Journal of Biomedical Informatics 2006; 39(3): 288-298.

From Concepts to Clinical Reality: An Essay on the Benchmarking of Biomedical Terminologies

Barry Smith

Department of Philosophy and National Center for Biomedical Ontology, University at Buffalo, Buffalo, NY 14260, USA

Institute for Formal Ontology and Medical Information Science, Saarland University, 66041 Saarbrücken, Germany

Abstract

It is only by fixing on agreed meanings of terms in biomedical terminologies that we will be in a position to achieve that accumulation and integration of knowledge that is indispensable to progress at the frontiers of biomedicine. Standardly, the goal of fixing meanings is seen as being realized through the alignment of terms on what are called 'concepts'.

Part I addresses three versions of the concept-based approach – by Cimino, by Wüster, and by Campbell and associates – and surveys some of the problems to which they give rise, all of which have to do with a failure to anchor the terms in terminologies to corresponding referents in reality.

Part II outlines a new, realist solution to this anchorage problem, which sees terminology construction as being motivated by the goal of alignment not on concepts but on the universals (kinds, types) in reality and thereby also on the corresponding instances (individuals, tokens). We outline the realist approach, and show how on its basis we can provide a benchmark of correctness for terminologies which will at the same time allow a new type of integration of terminologies and electronic health records. We conclude by outlining ways in which the framework thus defined might be exploited for purposes of diagnostic decision-support.

Keywords

terminology, ontology, referent tracking, concept orientation, instances, realism, electronic health record, diagnostic decision-support

Part I: The Concept Orientation and Its Problems

1 Introduction

The language of medicine is in constant flux. And while human beings can cope quite well with changing patterns of use and meaning of biomedical terms, when computers enter the scene then familiar problems arise. The orthodox approach to solving these problems, which is illustrated by almost all the terminologies integrated together in the Metathesaurus of the Unified Medical Language System (UMLS) [1], rests on the view that the fixation of meanings is best brought about through the alignment of terminologies on what are called 'concepts'.

As can be gauged by the number of influential terminologies developed in its wake, this 'concept orientation' was in some respects an important step forward in terminology development, in that it resolved many unfortunate features of the treatment of terms in the vocabularies of an earlier era. Most recently – as in the GALEN project [2] and in the SNOMED Clinical Terms vocabulary [3] – it has facilitated the application of tools such as Description Logic to the task of formalizing concept definitions in ways that can be used for automatic processing and quality assurance of terminologies.

On the other side, however, the concept orientation is beset with a number of fundamental difficulties, the most important of which is that the term 'concept' is used in so many different, sometimes highly counterintuitive, ways that it is difficult to know precisely what is meant by this term even when it is used by the same author and in the same paper.

Four loose families of views can be distinguished, which we can refer to as the *linguistic*, the *psychological*, the *epistemological*, and the *ontological*, respectively. On the linguistic view, concepts are *general terms* whose meanings have been somehow regimented (or, as on some variants of the view, they are these meanings themselves). On the psychological view, concepts are *mental entities*, analogous to ideas or beliefs. On the epistemological view, concepts are *units of knowledge* (as the latter term is used in phrases such as 'knowledge representation', 'knowledge modelling', 'knowledge-based systems', and the like). And on the ontological view, concepts are *universals*, *kinds*, *attributes or properties* (i.e. they are something like general invariant patterns) on the side of entities in the world.

Each one of these views might, in and of itself, be in a position to sustain a coherent methodology for the fixation of meanings in terminologies. As we shall see, however, elements of all four views are to be found mixed together in different combinations in the standard literature, in ways which provide strong evidence for the thesis that no single reading of the term 'concept' can sustain all of the expectations which have become associated with its use [4].

2 Cimino's Desiderata

James Cimino's important paper [5] advances a set of desiderata which must be satisfied by medical terminologies if they are to be able to support computer applications, based on the central idea that those involved in terminology work should focus their attentions, not on terms or expressions or on associated meanings, but rather on what are called ‘concepts’. As Cimino puts it, it is concepts that should serve as the ‘unit of symbolic processing’ in the construction of terminologies.

Cimino himself comes close to embracing a linguistic view of concepts. A concept, he says, is ‘an embodiment of a particular meaning,’ which means that it is something like a term that has been extricated from the flow of language change. This reflects the desideratum of ‘concept permanence’ to the effect that *the meaning of a concept, once created, should be inviolate*.

Three further desiderata distinguished by Cimino are:

– *the concepts which form the nodes of the terminology must correspond to at least one meaning* (‘non-vagueness’),

– *they must correspond to no more than one meaning* (‘non-ambiguity’),

– *these meanings must themselves correspond to no more than one concept* (‘non-redundancy’).

The concepts which form the nodes of a well-constructed terminology will, if these requirements are met, be mapped in one-one fashion to corresponding meanings.

Concepts as conceived by Cimino are thus in some respects analogous to WordNet’s *synsets*, which are collections of word-forms substitutable for each other without change of truth-value in given types of sentential contexts [6]. Concepts thus understood stand in different kinds of meaning-relations: *is narrower in meaning than*, *is broader in meaning than*, and so forth. Cimino, however, follows a usage now common in work on biomedical terminologies in speaking of concepts as being linked together also by *ontological* relations such as *caused by* or *site of* or *treated with*. He thereby embraces – simultaneously with the linguistic view – elements of the *ontological* view, according to which concepts would be abstractions from entities in reality (at one point in the text he refers even to protons as ‘low-level concepts’ [7, text to note 28]).

The ontological view has advantages over the linguistic view above all when it comes to understanding the ways the expressions in medical terminologies are in fact used by clinicians in making diagnoses. Cimino himself provides only one small hint in this connection, when he refers to the concept *diabetes mellitus* becoming ‘associated with a diabetic patient’. Presumably, this association does not come about because the physician has the patient on his left, and the concept (term, meaning) on his right, and decides that the two are fitted together to stand in some not further specified ‘association’ relation. Rather, there is something *about* the patient, something in reality, which the clinician apprehends and which makes it true that this given concept can be applied to this given case. Fatefully, however, like other proponents of the concept orientation, Cimino does not address this *what it is on the side of the patient* which would warrant ontologically the assertion that an association of the given sort obtains – he does not, in other words, address the issue of what it is in the world to which biomedical concepts such as *diabetes mellitus* would correspond.

3 International Standard Bad Philosophy

Cimino's reluctance is understandable. When concepts are pressed into service to perform incompatible roles – as standing both in relations like *narrower in meaning than* and in relations like *treats* or *causes* – then it becomes difficult to determine what, exactly, concepts might be and thus also difficult to specify in a coherent way how they might relate to actual clinical cases.

We can derive some illumination as to how this pass was reached, if we look back to the origins of the concept orientation in terminology work in the 1930s, when the Austrian engineer and businessman Eugen Wüster began to develop the astonishingly influential theory of terms and concepts which later became entrenched as the terminology standard promulgated by ISO, the International Organization for Standardization [8].

Wüster himself defended a *psychological* view of concepts, here echoing the views on the relations between thoughts (=Wüster's concepts), words and things articulated by Ogden and Richards in the form of the so-called 'Semantic Triangle' [9]. Thus Wüster held that words or symbols have direct reference not to things in reality but rather always only to concepts, which means to certain elements of thought existing in the minds of human subjects. Indeed Wüster sometimes writes as if, in order to apprehend concepts, we would need to gain access to the interiors of each other's brains:

> If a speaker wishes to draw the attention of an interlocutor to a particular individual object, which is visible to both parties or which he carries with him, he only has to point to it, or, respectively, show it. If the object, however, is in another place, it is normally impossible to produce it for the purpose of showing it. In this case the only thing available is the individual concept of the object, provided that it is readily accessible in the heads of both persons. [10]

An individual concept is (Wüster says) a memory of an individual object which can serve as its mental surrogate. His examples are: '"Napoleon" or the concept of my fountain pen'. A general concept, analogously, for example 'rabbit' or 'fruit', is a mental surrogate of a plurality of objects [8]. General concepts reflect similarities between objects which human beings are able to apprehend through perceptual experience and store in memory. They are human creations, arising as the cumulative reflection of what Wüster repeatedly insists are arbitrary choices made by humans in grouping objects together.

The perceived similarities which serve as starting points for such groupings are reified by Wüster under the heading of what he calls 'characteristics', a term which, like the term 'concept' has been embraced by the terminology community (and has thereby also fallen prey to a variety of conflicting interpretations). In some passages Wüster himself identifies characteristics with *properties* on the side of the objects in reality. In others, he identifies them as further *concepts*, so that they too would exist in the heads of human beings. [8]

The same uncomfortable straddling of the realm of mind (ideas, thoughts, meanings) and world (objects and their properties) shows up in Wüster's treatment of the *extension* of a concept, which he defines alternatively as 'the totality of all individual objects which fall under a given concept' and as 'the totality of all subordinated concepts'. Thus on the one hand the extension of the concept *pneumonia* would be the totality of cases (or, in philosopher-speak, of *instances*) of pneumonia; on the other hand it would be a collection of more specific concepts (*bacterial pneumonia*, *viral pneumonia*, *mycoplasma pneumonia*, *interstitial pneumonia*, *horse pneumonia*, and so on).

The unclarity of Wüster's thinking is reflected also in his definition of 'object' as 'anything to which human thought is or can be directed,' a definition which has unfortunately been given normative standing through its incorporation in different versions in many ISO standards. Thus ISO 1087-1:2000 defines an object as 'anything perceivable or conceivable', providing therewith the following *Note*: 'Objects may be material (e.g. an engine, a sheet of paper, a diamond), immaterial (e.g. conversion ratio, a project plan) or imagined (e.g. a unicorn).' [11] Similarly, Wüster's definition of 'object' would seem to imply that the extension of the concept *pneumonia* should be allowed to include not only your and my pneumonia but also, for example, cases of unicorn pneumonia or of pneumonia in Russian fiction.

With this, I believe, ISO undercuts any view of the relation between concepts and corresponding objects in reality that might be compatible with the needs of empirical science (where it is important to recognize that an imagined mammal is not a special kind of mammal). It thereby also cuts us off from any coherent understanding of that *what it is on the side of reality* to which the concepts used in biomedicine or other scientific disciplines would correspond.

4 Castles in the Air

A further illustration of the problems associated with the concept orientation is provided by [12], in which Keith Campbell, Diane Oliver, Kent Spackman and Edward Shortliffe present their account of the status of the UMLS in current terminology work.

The UMLS gathers into a single compendium terms from different vocabularies with the goal of creating 'unified meaning' across terminologies. The problem is that it does this even where the terms derived from separate source terminologies clearly have different extensions in this, the actual world, as for example when it assigns (in early versions of the UMLS) the same concept unique identifier (CUI) both to 'aspirin' and to a proprietary form of chewable aspirin called 'Aspergum'.

The thesis of [12] is that this is allowed because the UMLS is a 'Possible World' (the authors cite in this connection the work of Leibniz), in which 'aspirin' and 'Aspergum' do indeed refer to one and the same thing – and this in spite of the fact that 'many clinicians would not regard different formulations of aspirin ... as interchangeable concepts in the prescriptions they write'. But in what sense *is* the world thus defined *possible*, given that it would have to be governed by laws of nature different from those in operation here on earth? The answer is that it is possible, at best, as an *artifact* inhabiting that same strange Wüsterian realm in which aspirin 'may be an abstract concept'. In [12] the UMLS is accordingly referred to as an 'artificial world', as contrasted with 'our corporeal world' of flesh and blood entities.

The job of this artificial world is asserted to be that of providing 'a link between the realm in which we live and the symbolic world in which computer programs operate.' To achieve this end, accordingly, we need to distinguish three worlds:

1. the possible ('artificial') world which is the UMLS,
2. the 'symbolic world' of computers,
3. the 'corporeal world' in which we live.

Given that so much hangs on fulfillment of the task of knowledge integration in biomedicine, linking worlds 2. and 3. together would be a valuable achievement indeed. But how is this linkage to be effected? By appeal, surely, to the *extensions* of the concepts in the UMLS, understood as collections of particular entities (actual patients, actual pains in actual heads, actual pieces of Aspergum chewed) in the corporeal world. The authors themselves suggest that they accept a view along these lines when they embrace the standard Fregean interpretation of 'extension', according to which (as they interpret [13]):

> the "meaning" of expressions can be divided into two components: On the one hand there are the physical objects to which the expression refers (the expression's *extensional* component) and on the other there are the characteristic features of the physical object used to identify it (the expression's *intensional* component). Understanding the interrelationship between intensional and extensional meaning is essential to understanding the "senses" of meaning represented within the UMLS. Only with this understanding can we know when symbols (such as 'morning star' and 'evening star') can be substituted for one another without loss of truth. [12, single quotation marks added]

When it comes to the UMLS itself, however, they abandon this Fregean reading in favor of a view according to which (if we have understood their formulations correctly) the extensions of the concepts in the UMLS would be *sets of concepts drawn from source terminologies*:

> the developers [of the UMLS] collected the language that others had codified into terminologic systems, provided a framework where the intension (connotation) of terms of those systems could be preserved, and unified those systems [into one *unified* system] by providing a representation of extensional meaning by collecting abstract concepts into sets that can be interpreted to represent their extension.

They then assert that:

> These extensional sets are codified by the *Concept Unique Identifier* (CUI) in the UMLS. We argue that the "meaning" of this identifier is only understandable extensionally, by examining the characteristics shared by all abstract concepts linked by a CUI.

If the extension of a UMLS concept is not a set of instances in reality but rather a set of concepts, then the term 'extension' has come to mean what is ordinarily called 'intension'. With this reinterpretation, however, our authors have denied themselves the possibility of a conception of the UMLS as providing the desired link between the symbolic dimension of computer programs and the domain of real-world entities. For by abstracting the extensions of UMLS concepts away from corresponding instances, they have also left themselves no means of giving an account of how these concepts would relate to the *what it is on the side of reality* that is addressed by clinicians and biomedical researchers when they use the corresponding terms.

5 The Birth of the UMLS

It is a subsidiary goal of [12] to provide a theoretically illuminating account of how the UMLS came to be constructed. Here we must bear in mind that the state of its source terminologies was often not such that the creators of the UMLS could ascertain what characteristics had been associated with the concepts in these terminologies just by looking at the corresponding terms. The proposal of [12] is that the creators of the UMLS were able to tease out the relevant information by examining certain clues left behind in the course of terminology construction:

> When developers of source terminologies developed their systems, they had very specific thoughts about what the individual terms 'meant' (in the intensional sense) with respect to the terminology they were developing and

the human beings who would interact with those systems. Although we cannot directly know what was in the minds of the developers of the source terminologies, the UMLS developers have used clues embodied within the sources to try to infer what those thoughts were and to try to codify those thoughts within the UMLS.

ICD-10, for example, includes "C75.0: parathyroid gland," a term which on the face of it belongs to the domain of anatomy. However, the associated parent term C75, which reads: "malignant neoplasm of other endocrine glands and related structures," provides a better clue to what the developers of ICD-10 really had in mind with C75.0. In building the UMLS, accordingly, care was taken to ensure that the code C75.0 would be linked to the UMLS CUI for "malignant neoplasm of parathyroid gland" and not to that for "parathyroid gland." (More recently, the string or term corresponding in ICD-10 to C75.0 has additionally been marked in the UMLS as 'lacking face validity'.)

[12] tells us that the clues which served as basis for making such decisions took several forms:

the term used by a source to describe the thought; the synonyms used by a source to describe other statements that its developers considered equivalent to the thought; and any formal or informal relationships used by the developers to relate terms within the terminologic system to one another. Some of the informal relationships had to be inferred [by the creators of the UMLS] from processing the typesetting tapes for a particular source, using constructs such as how many tabs appeared before the word, whether the word was in bold or italics, and what page of the printed book the word occurred on.

The methodology described by Campbell *et al*. in [12] thus presupposes a distinction between two sorts of clues left behind by the developers of UMLS source terminologies: those which do indeed reveal what the latter 'had in mind', and those which reveal merely aspects of their thinking clouded over by the bad term-formation principles which the UMLS needed to correct for. The two sorts of clues do not, of course, come ready labelled as such. If the proposed methodology is truly to have been applicable in coherent fashion, therefore, then the creators of the UMLS must have had some means of separating out good clues from bad. Note that the concept orientation cannot itself provide us with any help in understanding how they were able to carry out this task. For it is precisely concepts that are supposed to have served as both input and output when the methodology in question was being applied, and this means there is no way in which concepts themselves could have served also as benchmark of correctness.

How, then, were the creators of the UMLS able to find their way successfully through the mass of clues left behind by their predecessors? One intuitively appealing answer is this: that they were able to recognize the good clues because they were implicitly taking into account that corporeal world to which the corresponding terms were, however inchoately, pointing. That it was, in other words, precisely the real world of real biomedical phenomena, with which physicians are familiar from their training and everyday experience, which played the role of measure of correctness for their coding choices.

6 How Terms are Introduced into the Language of Biomedicine

The realist orientation in biomedical terminology development is the result, now, of generalizing this simple proposal. It consists in the view that terms in terminologies (treated straightforwardly as linguistic items, as strings of certain special sorts) are to be aligned not on concepts but rather on entities in reality. This realist orientation has a small but growing band of defenders, which includes the authors of the Foundational Model of Anatomy [14] together with members of the Gene Ontology [15] and Open Biomedical Ontologies [16] consortia collaborating together in the National Center for Biomedical Ontology [17]. The proposal on which it rests now needs to be carefully unpacked if it is to fulfill its promise.

Consider, to fix our ideas, how new biomedical terms are formed, for example when a new disorder first begins to make itself manifest. Slowly, through the official and unofficial cooperation of physicians and other involved parties, a view begins to become established to the effect that a certain family of cases, manifesting a certain newly apparent constellation of symptoms, represents instances of a hitherto unrecognized kind. We are confronted, therefore, with a phenomenon involving both *real world instances* and also the *biomedical kinds* which these instantiate.

A kind is, for the realist, a part of reality: it corresponds to what philosophers call a 'universal', or in other words to an invariant pattern in reality which is multiply exemplified in an indefinitely extendable range of different instances. It is such universals which, by allowing us to describe multiple particulars using one and the same general term, make science possible. Such universals also make science-based clinical care possible, as they allow uniform treatments and associated clinical guidelines to be applied to pluralities of disorders diagnosed to be of like kind. And it is universals, too, which make terminologies possible.

The problem, of course, is that it is in many cases difficult to establish what universals or kinds given particulars are instances of. Again, when a disease universal first begins to make itself manifest it will still hardly be understood. Something similar applies when a new virus or gene is first detected, or a new kind of biochemical reaction in the cell. While, in regard to each individual case, users of the term may know precisely what they are referring to – they can point to it in the lab or clinic – it may be difficult to convey this information to others. This is because the user has a clear understanding of what the term designates in reality, but only at the level of instances and not yet at the level of universals.

With increasing understanding of what universal given instances have in common, however, the new term becomes entrenched as a means to refer to the corresponding family of cases, which itself begins to be apprehended as the *extension* of the corresponding universal (sometimes also called the 'class' of its instances).

Those involved come to an agreement to use from here on

(1) *this term*

for: (2) *these instances*

of: (3) *this kind*.

Against the background of the concept orientation, however, there is postulated also:

(4) a new *concept*,

together with

(5) a *definition*.

If (4) is simply a chosen 'privileged term', or a WordNet-style synset, or some handy shorthand for a grouping together of terms in different natural languages which are allowed to be used as translations or synonyms of a given selected term, and if definitions under (5) are promulgated only after the point where the corresponding universal is properly understood in terms of necessary and sufficient conditions, then little harm is done by the postulation of (4) and (5). Responding, however, to the edicts of ISO and of healthcare messaging and W3C-style standards bodies, and to the needs of computer programmers, terminologists have come to see (4) and (5) as offering a new and special realm for exploration, the ‘realm of concepts’. Concepts and definitions come thereby to be de-anchored from the world of universals and their instances and they begin to acquire a life of their own.

One advantage of this move for terminologists is that it can be ensured that each and every general term *p* has its own precisely tailored referent – called ‘*the concept p*’ – a referent that is guaranteed to exist even when the term in question (‘unicorn’, ‘phlogiston’) has no application to either universals or instances in reality. One disadvantage, however, is that the move to concepts hampers the goal of coming to grips with the universals in reality (in this our corporeal world), because it substitutes instead the much weaker goal of reaching consensus on the use of words. It then postulates entities called ‘concepts’ wherever such consensus has been reached.

Matters are complicated still further by the fact that, on the ontological reading of ‘concept’, concepts themselves come close to being identified with universals in reality. And because there are traces of this ontological reading underlying many uses of the term ‘concept’ in the standard literature, proponents of the concept orientation may find it difficult to understand why it is necessary to insist so forcefully on the distinction between concepts and universals.

One reason is precisely the very many terms in biomedical terminologies which have been associated too readily with concepts even though they correspond to no universal in reality. There are no universals corresponding, for example, to UMLS terms such as

probable suicide

possible tubo-ovarian abscess

other European in New Zealand

gallbladder calculus without mention of cholecystitis

atypical squamous cells of uncertain significance, probably benign.

Such terms represent not entities in reality as they exist in advance of and independently of our testing and measuring and inquiring activity. Indeed they represent nothing in reality at all. Rather, they are disguised assertions about our ways of gaining knowledge of or referring to entities of other types in specific kinds of contexts [18].

More important, however, is the fact that where, according to the concept orientation, the meanings of terms in a terminology would belong to a realm whose denizens exist as products of agreement, according to the realist orientation they belong to a realm which exists prior to and independently of any agreements we are able to make. According to the concept orientation, if practitioners have agreed that two terms have different meanings, then they do indeed have different meanings. According to the realist orientation, it is possible that we discover that two terms (‘Bilharzia’ / ‘Schistosomiasis’; ‘morning star’ / ‘evening star’) mean the same thing, for example because differentiating (clinical; astronomical) manifestations were initially misinterpreted. Nothing in the realm of concepts, or of the ideas or beliefs of language users, can inform us of this coincidence; rather it takes arduous inquiry, directed at reality itself. We must thus rely at every stage on the instances in the world and on the patterns of similarities between them to tell us what the meanings of our general terms are, in an empirical process of discovery that is never brought to an end [19, 20]. This implies that, where the concept orientation views terminology development as effectively a linguistic exercise (it is, again, a matter of coming to agreement on words), the realist orientation views terminology development as associated much more closely with the advance of science on the basis of reality itself as benchmark.

Part II. Grounding Biomedical Terminologies in Clinical Reality

7 Tracking Referents

While universals and instances exist independently of our human cognitive activities, terminologies and clinical records are, like scientific theories, human creations. Each terminology should represent the universals about which the consensus of researchers in its domain believes itself to have gained knowledge at the stage when the terminology is created. Each clinical record should represent what its compilers believe about the corresponding cases in light of (and in some cases of course also in spite of) the terminologies they are using at the stage when the clinical record is created.

Terminologies and clinical records are connected together through single cases (instances), and it is the totality of such cases in reality which serves as benchmark of correctness for both. Our task is to find a means of understanding this fact theoretically, but in such a way that our understanding can be exploited for the practical purposes of healthcare and biomedical research.

The idea, simply put, is to devise an approach which will allow terminologies to be built up from what the physician is confronted by at the point of care. To this end we need to engineer a shift of focus in terminology construction to particular medically salient entities of a range of different types, including both *objects*, such as cells or fractures or inflammations, and *processes*, such as disease histories, rises in temperature, or the clottings of particular portions of blood.

Major terminologies such as ICD-10 or SNOMED CT already comprehend a wide variety of different kinds or categories of universals in the realms of disorders, symptoms, pathological and non-pathological anatomical structures, acts of human beings (for example anesthetizings, observings, interpretings of symptoms), biological processes (for example processes of digestion, movement, development, growth, aging), and many more. And while, for each of the latter, there is a family of particulars which instantiate the corresponding universals, these instances themselves are, under existing EHR regimes, not directly entered in a clinical record.

This is because existing systems for keeping track of clinical phenomena allow direct reference to just a small number of types of particulars, normally just to (i) *human beings* (patients, care-providers, family members, via proper names or via alphanumeric patient IDs), (ii) *times* and (iii) *places* at which actions are performed or observations made [21, 22].

This impoverished repertoire of types of direct reference means that no adequate means is available to keep track of one and the same particular (for example a specific wound or tumor) over an extended period of time. When interpreting health record data, it is accordingly difficult to distinguish clearly between multiple examples of the same particular and multiple particulars of the same general kind [21]. The same limitation also places obstacles in the way of drawing reliable inferences, for example for public health purposes, from the existence of different instances of the same clinical universal in different patients [22].

Under present EHR regimes, when the need arises to refer in different contexts to some single particular as it exists at different points in time, each such reference must be created anew, via some combination of general terms (or associated codes) with designators for persons, times and places, for example in expressions like: *the fever of patient* #1001 *observed by physician* #4001 *at time* #9001 *in hospital ward* #7001. Unfortunately, such composites, even where they are formulated by the same physicians using the same general terms deriving from the same coding systems, constitute barriers to reasoning about the corresponding particular entities in software systems. (Imagine a regime for reasoning about human beings as they change and develop and move about over time in which people could be referred to only by means of expressions like: *patient in third bed from left*, or *person discharged after appendectomy*, or *relative of probable smoker*.)

In [21, 22] we have proposed a method by means of which the corresponding instances would be made directly visible to reasoning systems without need for prior processing. This involves the creation of a new sort of EHR regime in which explicit alphanumerical identifiers – analogues of proper names – would be as far as possible automatically assigned in the course of data entry to individual real-world entities at the point where they first become relevant to the treatment of the corresponding patients. Such instance unique identifiers (IUIs) would be assigned to instances of universals in all the diagnostically salient categories recorded in a clinical record as a means of doing justice to the *what it is on the side of the patient* in all its richness and complexity. In this way, they can serve to tie together different views of one and the same instance of a given disorder which may become incorporated into the record, for example when physician A writes ‘tumor’ and physician B writes ‘CAAA12’. They can thereby also, as we shall see, help to solve the anchorage problem – the problem of explaining how terms in terminologies can be anchored to corresponding referents in reality. For the use of IUIs would allow us to identify the corresponding particulars both in written records and in computer representations in a way which would make it clear when different physicians or biomedical researchers in different disciplines are referring to one and the same particular. The cumulative result of such use would then amount to a giant, growing, map-like representation of the particulars in the healthcare domain and of the interrelations between them.

8 Understanding Terminology Revision

As in the past, so also at every foreseeable stage in the future, terminologies and clinical records will be subject to the need for correction of errors. Note, however, that this recognition of the need for constant revision on both the terminology (universal) and EHR (instance) levels goes hand in hand with another component of the realist orientation, according to which both the vast and settled majority of the beliefs expressed or presupposed in biomedical texts, and the vast majority of assertions captured in clinical records, are both true and uncontroversial. (It is not controversial, for example, that the terms 'parathyroid gland' and 'malignant neoplasm of the parathyroid gland' refer to two different entities in reality.) It is also compatible with another (surely also correct) view according to which the sum total of true beliefs of both kinds is constantly increasing, so that there is, in biomedicine as in other fields of science, a broad accumulation of knowledge.

The phenomenon of constant revision tells us, however, that mixed in with the knowledge that is captured in terminologies and in clinical records there is at every stage a small and ever-changing admixture of *false belief*. The part of this admixture which most concerns us here takes the form of terms in a terminology which are associated with a claim to refer to some corresponding universal but where this claim is not fulfilled. This can be either because there is no universal at all which can serve as referent of the term in question, or because the term refers ambiguously to what is in fact a plurality of universals. This means that the realist counterparts of the three central Cimino desiderata:

- each term in a terminology must correspond to at least one universal ('non-vagueness')
- each term must correspond to no more than one universal ('non-ambiguity')
- each universal must itself correspond to no more than one term ('non-redundancy').

should be accepted only as long-term goals, to the ever closer but never quite complete realization of which terminologists are condemned. In moving towards the realization of these goals, they must follow always in the coat-tails of those engaged in empirical research in attempts to expand our knowledge of biomedical universals and their instantiations.

The proper understanding of terminologies and EHR systems must accordingly take account of the dynamic nature of both types of artifacts. Moreover, they must do this in such a way as to recognize two levels of dynamism, reflecting *changes in reality* and *changes in our scientific beliefs about reality*, which implies the need to keep track of time in two different ways. This idea involves nothing that is essentially new: our EHRs already track events of many sorts by indexing with times; and we track changes in terminologies (which here go proxy for corresponding beliefs) by means of version numbers. The relative independence of these two temporal dimensions is seen in the fact that we can in principle direct a version of a terminology created today to the task of classifying or reclassifying instances existing a week or a decade or 5,000 years ago.

9 Terminologies: A Formal Treatment

We can now define a *terminology*, more technically, as a graph-theoretic object consisting of nodes joined together by links, the whole indexed by version number. More precisely, a terminology is an ordered triple:

$$T = \langle N, L, v \rangle$$

where:

> N itself is a set of triples $\langle p, S_p, d\rangle$, called *nodes*, with p a unique *label* (which may be either some alphanumeric identifier or what is sometimes called a 'preferred term'), S_p a set $\{s, s', s'', ...\}$ of *synonyms* (including alphanumeric codes in systems like SNOMED CT), and d an (optional) *definition* (the precise format and standing of which we can here leave out of account),
>
> L is a set of ordered pairs $\langle r, L_r\rangle$, called links, consisting of a relation designation r ('*is_a*', '*part_of*', etc.), together with a set L_r of ordered pairs $\langle s, s'\rangle$ of those terms for which '$s\, r\, s'$' represents a consensus assertion of biomedical science about corresponding universals (if any) at the time when the given terminology is prepared,

and

> v is a version number, which encodes this time.

The variables p, s, d, r, v, …, range over syntactic entities (strings of characters in some regimented language). Importantly, some values of s, s', ..., will correspond to no universal in reality. (Like 'unicorn' or 'phlogiston' they will be *empty names*, which correspond to nothing in reality at all.) Others will correspond to too much on the side of reality (i.e. they will refer ambiguously to a plurality of universals). Every terminology will in this sense be marked by two kinds of defect, both which must be taken into account when we consider the whole terminology $T = \langle N, L, v\rangle$ in light of its status as a (partial) map of an analogous structure of universals in the corporeal world. Our approach thus differs radically from the standard approaches evinced by the majority of terminologies in the UMLS Metathesaurus, which skirt round both kinds of defect by postulating 'concepts' to serve as precisely tailored referents wherever needed. This means, however, that the curators of these terminologies are unable to come to grips with the ways in which clinical reality can serve as benchmark both for the correctness of terminologies in the large, and for the correctness of local applications of terminologies to particular cases in the small.

In what follows we shall be concerned almost exclusively with the set N as this exists in modified form in successive versions of a single terminology. While in the ideal state of terminological virtue we could indeed associate the nodes in N in one-one fashion with the universals in the corresponding domain of reality, really existing terminologies fall short of this ideal in the three ways identified in our realist counterparts of Cimino's criteria of non-vagueness, non-ambiguity, and non-redundancy above. This means (roughly, and for our present purposes) that at any given stage these nodes will in and of themselves be divided into three groups:

$$N = N_1 \cup N_> \cup N_<,$$

where N_1 consists of those nodes in N which correspond to exactly one universal, $N_>$ of those nodes which correspond to more than one universal (in various combinations), and $N_<$ of those nodes which correspond to less than one universal, which means in practice to no universal at all.

It is an assumption of scientific realism that, with the passage of time, $N_>$ and $N_<$ will become ever smaller, so that N_1 will approximate ever more closely to N. (This assumption must be qualified in light of the fact that N will change in reflection not only of changes in our knowledge, but also of changes in the totality of biomedically salient universals in the reality beyond. Changes of the latter sort are pervasive for example in the domain of therapies, reflecting advances in drug design, and in molecular biology and related domains, reflecting advances in scientific knowledge about proteins and other gene products. They will not, however, affect our argument here.)

Our knowledge of the successes of medical science gives us strong reason to believe that, at any given advanced stage in the development of a terminology, N_1 will constitute a large portion of N (N, remember, is a collection of terms already in use among domain experts, each of which is associated with the implicit or explicit claim to represent a biomedical universal; remember, too, that N will standardly include very many uncontroversial terms such as 'ear' or 'cough' or 'gland'). At the same time however our knowledge of the ways errors become locally manifest in specific terminologies gives us reason to believe that we have some way to go before $N_>$ and $N_<$ can be excised completely.

Moreover, we know *a priori* that at no stage (prior to that longed-for end to our labors that is forever out of reach) will we know precisely where the boundaries are to be drawn between N_1, $N_>$, and $N_<$ – that is, we will never know precisely which portions of N consist of the low value $N_>$- and $N_<$-type terms we are seeking to eliminate. The reason for this is clear: if we did know where these low value terms were to be found, then we would already have the resources needed to expand correspondingly the size of N_1 and thus to move its boundaries to a position closer to those of N.

This unavoidable lack of knowledge of the boundaries of N_1 is not a problem, however. For it is, after all, N, and not N_1, which is the focus of our practical labors. It is N which represents our (putative) consensus knowledge of the universals in the relevant domain of reality at any given stage.

Even though we do not know how the terms are presently distributed between the three groups, we shall see that this does not mean that the distinction between N_1, $N_>$, and $N_<$ terms is of purely theoretical interest – a matter of abstract (philosophical) housekeeping, of no concrete significance for the day-to-day work of terminology development and application. For we shall see that it is this very distinction which will provide us with the resources we need to exploit instance data as benchmark for terminology revision.

10 A Framework for Terminology Refinement

Consider once more our scenario concerning the way in which a medical term is introduced into our language. While the *instances* in our initial pool of cases, and certain patterns of irregularities (deviations from the norm) which they exemplify, are well known to the physicians involved, the *universal* which they instantiate is as yet unknown. The challenge is then to *solve* for this unknown (in something like the way in which astronomers postulated an unknown heavenly body, later identified as Pluto, in order to explain irregularities in the orbits of Uranus and Neptune). And we can now see that three different kinds of solution can present themselves: the cases in the pool – not patients, remember, but the corresponding particular disorders – are (i) instances of exactly one universal, (ii) instances of no universal at all, (iii) instances of more than one universal.

To see how we might make practical use of this idea, we need to imagine, again, a future world of sophisticated electronic health records in which instances in all clinically salient categories are tracked by means of IUIs. Each IUI would be associated with a vector, comprehending both relevant assignments of preferred general terms in one or more terminologies and also, utilizing the relational (*L*-) component associated with *N*, cross-references to the IUIs assigned to those other particulars (including the relevant patients and medically salient attributes such as temperature, blood pressure, etc.) with which the entity under scrutiny is related, for example in the ways catalogued in [21, 22, 23].

We can then define an *instance vector* as an ordered triple

$$<i, p, t>$$

consisting of a IUI *i*, a node *p* in a terminology, and a string *t* designating a time at which the particular designated by *i* is asserted to be an instance of the universal (if any) designated by *p*. (Here and in what follows we refer to nodes in *N* via their corresponding labels which we have specified to be unique.) Such instance vectors reflect the fact that the IUIs in our repository will typically already have been associated at the point of entry into the EHR with preferred terms or associated codes from one or more terminologies. For example the EHR will contain the assertion that tumor instance #5001 is an instance of the universal associated with the SNOMED-CT code for *glomus tumor* (*morphologic abnormality*).

For a given set *D* of IUIs (gathered for example by a single healthcare institution in a given time period), we can now define a *t-instantiation* $\boldsymbol{I}_t(T, D)$ of a terminology $T = <N, L, v>$ as the set of all instance vectors $<i, p, t>$ for *i* in *D* and *p* in *N*. We can also define for each term *p* in *T* its *t-extension* $\boldsymbol{I}_t(T, D)(p)$ as the set of all IUIs *i* for which $<i, p, t>$ is included in $\boldsymbol{I}_t(T, D)$. The *t*-extension then goes proxy for (is a map of) the extension of the universal (if any) designated by the node *p* in that particular domain of reality which is selected for by *D* at time *t*. (In brief, it comprehends all the known instances of *p* in the relevant domain.)

For each node *p* we can now examine its *t*-extensions for different values of *D* and *t*, in order to determine statistical patterns of different sorts, taking into account also, for each *i*, the other instance vectors in which *i* is involved through the relations in which the corresponding instances stand to other instances represented by IUIs in *D*. Three alternative scenarios once again present themselves, according to the status of each node *p* in relation to the world of actual cases (the world which serves as benchmark for the truth and falsity of our assertions):

1. *p* is in N_1(there is a single universal designated by *p*): in this case the instances in $\boldsymbol{I}_t(T, D)(p)$ share in common a specific invariant pattern (which should be detectable through the application of appropriate statistically based tools),

2. *p* is in $N_>$ (*p* comprehends a plurality of universals, for example in a manner analogous to the term '*diabetes*'): in this case the instances in $\boldsymbol{I}_t(T, D)(p)$ manifest no common pattern, but they (or the bulk of them) can be partitioned into some small number of subsets in such a way that the instances in each subset do instantiate such a pattern,

3. *p* is in $N_<$ (*p* comprehends no universals): in this case the instances in $\boldsymbol{I}_t(T, D)(p)$ manifest no common pattern and there is no way of partitioning them (or the bulk of them) into a combination of one or a small number of subsets in such a way that all the instances in each subset instantiate such a pattern.

On the basis of inspection of the ways in which different terms fall under one or other of these three headings we can then subject terminologies, either manually or automatically, to evidence-based processes of refinement and correction, picking out questionable terms which need to be subjected to further testing and to subsequent deletion or subdivision as necessary.

For a given disorder term *p*, we gauge whether *p* is in N_1, $N_>$ or $N_<$ by applying statistical measures to the similarities between the vectors associated with each of the members of relevant instantiations. If, for example, the measure of similarity between such vectors is both roughly similar for all members of a given instantiation and also roughly constant across time, then this will constitute strong evidence for the thesis that *p* is in N_1. If, on the other hand, we find high similarity for some disorder term before a certain time *t*, but much lower degrees of similarity after some later time t^+, then we can hypothesize that the relevant disorder has itself undergone some form of mutation, and we can experiment with adding new terms and then repartitioning the available sets of IUIs in such a way as to reach once again those high levels of similarity which are associated with the N_1 case.

11 Applications of the Referent Tracking Methology

By allowing instances in reality to play the role of benchmark for the correctness of a terminology, the referent tracking methodology can be used also in other ways. Thus, on the (still highly speculative) scenario we are here outlining, the gradual revision and refinement of terminologies will in due course spawn, in the opposite direction, better patient records through revision of the information associated as vectors to each of the relevant IUIs, for example when we discover that a given single disorder term has thus far been applied incorrectly to what are in fact instances of a plurality of distinct disorders. This will thereby lead in turn to better quality clinical record data, which may in turn spawn yet further revisions in our terminologies.

Secondly our methods for cross-calibrating terminology and instance data might be used for purposes of decision support in the process of diagnosis. For where instances in reality are able to serve as global benchmark for the correctness of a terminology, then they can serve this role also locally, at the point where the clinician is confronted with the individual patient.

One goal of an adequate terminology-based reasoning system in a world of abundant instance data would thus be to allow the application of statistical tools, for example tools for association rule mining [24], in the service of diagnostic decision support. We can imagine a scenario in which the clinician is able to experiment with alternative diagnoses (which is to say: alternative term-assignments to given collections of instance data) on the basis of measurements of the statistical likelihood of given patterns of association between terms and instances. We could imagine also software which would allow clinicians to experiment with alternative IUI assignments in those cases where it is unclear whether successive clusters of symptoms of a given patient should be counted as manifestations of a single or of multiple disorders. The machinery of instantiations could then be used by the clinician to test out alternative hypotheses regarding how to classify given particulars, by giving him the facility to experiment with different scenarios as concerns the division between N_1, $N_<$, and $N_<$ in relation to a given case, taking into account also oft-repeated patterns of error in diagnosis made by physicians confronted with analogous instance-data in the past. The goal is a software tool which would allow statistical *tuning* of the relevant local parts of a terminology to given instance-based EHR data.

In the real world, of course, such methods will not be able to be applied successfully in every case. For example we may not have all the data needed to convince a statistical reasoner armed with a given stock of universal terms and associated instance data that a given case meets the requirements for any available diagnosis. This scenario is however no different from that which is often faced already by the practicing physician, who must decide from case to case how much data to collect (for example how often to take the temperature of a given patient) in order to achieve a succession of better approximations to what then establishes itself as a good diagnosis on the basis of successful treatment.

The methodology can be used, finally, to support the making of scientific discoveries. Suppose, for example, that the length of a patient's nose is correlated with a certain specific disease, but that this fact is unknown to medical science. Why should anyone start to register patient nose-length in the way that we do now for, say, temperature or blood pressure? The answer is that such data has been collected already for the many hundreds of thousands of patients who have undergone plastic surgery for cosmetic nose corrections. In each such case, the length of the nose is measured as a matter of course. Many of these patients visited other physicians for totally different problems (before, at the same time, or later). If all the physicians involved had been exploiting the potential of referent tracking, then it would not be too difficult to correlate these data, just by using brute-force techniques such as cluster analysis, principle component analysis, or factor analysis, in order to tease out the correlation in question in just the way that scientific discoveries are sometimes made on the basis of statistical analyses of instance-level data in other domains.

12 Conclusion

In the ideal case, a biomedical terminology would provide not merely the resources for assigning terms for universals to the instances in reality, but also a perspicuous map of how these universals themselves are related together in reality. As we conceive the EHR system of the future, instance data will be to a large degree automatically partitioned at the point of data entry in ways reflecting the structure of the world of clinically relevant universals, with alternative partitions included as options in those cases where diagnosis is still uncertain. This partitioning of instances is currently masked from view in the clinical record because the instance-level data that exists in separate EHRs is accessible only via the detour of reference to the individual patient. A regime for the management of terminologies and clinical data along the lines described in the foregoing, however, would allow us to map directly the instances that are salient to medical care in such a way as to mirror how the latter are themselves related together in reality at the instance level. In this way it would make possible a new level of sophistication in reasoning about the *what it is on the side of the patient* that is the primary focus of medical care.

Acknowledgements

Work on this paper was carried out under the auspices of the Wolfgang Paul Program of the

Humboldt Foundation, the European Union FP6 Network Semantic Datamining, and the Volkswagen Foundation Project “Forms of Life”. Thanks are due also to Werner Ceusters, to James Cimino, and to an anonymous referee, for helpful comments.

References

1. The Unified Medical Language, System: www.nlm.nih.gov/research/umls; last accessed August 4, 2005.
2. Rector AL, Nowlan WA. The GALEN project. Comput Methods Programs Biomed. 1994 Oct; 45(1-2): 75-8.
3. SNOMED Clinical Terms: http://www.snomed.org/snomedct; last accessed August 4, 2005.
4. Smith B. Beyond concepts: Ontology as reality representation. Proc Third International Conference on Formal Ontology in Information Systems (FOIS 2004). Amsterdam: IOS Press; 2004, p. 73-84.
5. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
6. Fellbaum C. (ed.) WordNet. An electronic lexical database. Cambridge, MA: MIT Press; 1998.
7. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
8. Smith B, Ceusters W, Temmerman R. Wüsteria. Medical Informatics Europe (MIE 2005), p. 647-52.
9. Ogden CK, Richards IA The meaning of meaning. London: Kegan Paul, Trench, Trubner, 1923.
10. Wüster E. The wording of the world presented graphically and terminologically. Terminology 2003;9(2):269-297.
11. ISO 1087-1:2000. Terminology work. Vocabulary. Theory and application
12. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc 1998;5(5):421–431.
13. Frege G. On sense and reference. In: Geach PT, Black M (eds.). Translations from the Philosophical Writings of Gottlob Frege, Oxford: Blackwell, 1960.
14 Rosse C, Mejino JLV Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform, 2003 Dec;36(6):478-500.
15. http://www.geneontology.org
16. http://sourceforge.net/projects/obo
17. http://ncor.us
18. Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: A case study in medical terminology. Proc Third International Conference on Formal Ontology in Information Systems (FOIS 2004), Amsterdam: IOS Press; 2004, p. 185-195.
19. Johansson I. Biomedical informatics and biological reality. In this volume.
20. Putnam H. Is water necessarily H_2O? in H. Putnam, Realism with a Human Face, Cambridge MA: Harvard University Press, 1990.
21. Ceusters W, Smith B. Tracking referents in electronic health records. Proceedings of MIE 2005, p. 71-6.
22. Ceusters W, Smith B. Strategies for referent tracking in electronic health records. In this volume.
23. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector AL, Rosse C. Relations in biomedical ontologies. Genome Biol 2005, 6: R46.
24. Hipp J, Güntzer U, Nakhaeizadeh G. Data mining of association rules and the process of knowledge discovery in databases. Industrial Conference on Data Mining: Advances in Data Mining, Applications in E-Commerce, Medicine, and Knowledge Management (Lecture Notes In Computer Science 2394), 2002;:15-36.