New Desiderata for Biomedical Terminologies

Barry Smith

Department of Philosophy, University at Buffalo, Buffalo, NY 14260, USA
Institute for Formal Ontology and Medical Information Science, Saarland University, 66041 Saarbrücken, Germany

Abstract

Only by fixing on agreed meanings of terms in biomedical terminologies will we be in a position to achieve the sort of accumulation and integration of knowledge which is indispensable to progress at the frontiers of biomedicine. Yet when we examine more closely the most popular methodology for achieving this goal, then we find that its reliance on the so-called 'concept orientation' places severe obstacles in the way of its realization. We summarize some of the reasons why this is so, and then sketch a new 'realist' orientation in which not concepts but universals (kinds, types) and instances (individuals, tokens) in reality play a central role. This incorporation of instances and of the associated machinery for tracking instances yields a tighter connection between terminologies and the electronic health record, and this in turn allows a new type of diagnostic decision-support based on the statistical comparison of alternative assignments of terms to instances.

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

The Concept Orientation and Its Problems

While the language of medicine is in constant flux, experience tells us that human beings can cope quite well both with the creation of new terms and with changes in the meanings of terms already in use. When computers enter into the picture, however, then familiar problems arise – problems which eventually become so serious that there is born the longing that someone will step forward to shout 'Stop!' It is in this spirit, I believe, that we can best understand James Cimino's important paper [1], which advances a set of desiderata which must be satisfied by medical terminologies if they are to be able to support computer applications. I shall argue in what follows that many of Cimino's desiderata can be accepted by those involved in terminology work – but only when they have been subjected to radical reinterpretation.

It is Cimino's principal thesis 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.

As can be gauged by the number of influential terminologies that have been developed in its spirit, adoption of 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 and in the most recent versions of the SNOMED vocabulary – 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. But the concept orientation is beset with problems still too little addressed in the literature, to the degree that, as I hope will become clear in what follows, it should now be regarded as a regrettable detour from the path of terminological virtue.

The first problem is that the term ‘concept’ is used in so many different ways in work on terminologies that it is difficult to know precisely what it means. 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 modeling’, ‘knowledge-based systems’, and the like). And on the ontological view, concepts are *abstractions of kinds, attributes or properties* (i.e. of general invariant patterns) on the side of entities in the world.

Elements of all these views are to be found mixed together in different combinations in the literature [2] in ways which bear testimony to the fact that no single one of the four views yields an account of concepts that can sustain all of the associated expectations.

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. One desideratum for a well-constructed medical terminology is accordingly that of ‘concept permanence’: *the meaning of a concept, once created, is inviolate.* Three further desiderata 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 nodes (preferred terms) 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 [3]. Echoing the authors of WordNet, Cimino recognizes that synonymy is not an equivalence relation dividing up the domain of terms neatly into disjoint synsets. Word-forms which are synonyms relative to some types of context are often not synonyms relative to others. To resolve this problem he invokes the further desideratum of *context representation*, which requires a terminology to contain formal, explicit information about how concepts are used in different types of contexts. If, however, we are right in our view that concepts for Cimino are themselves synsets, then concepts are thereby already relativized to contexts. Thus in formulating the mentioned desideratum he ought more properly to speak not of *concepts* but rather of *terms* themselves being used in different types of context. If this is so, however, then his strategy for realizing the *concept* orientation requires that he first of all take seriously precisely that *term* orientation which (if we understand his views correctly) the *concept* orientation was designed to do away with.

Concepts understood as synsets ought, presumably (as in WordNet itself), to be seen as standing in different kinds of meaning-relations: *is narrower in meaning than, is wider in meaning than,* and so forth. Cimino, however, follows the usage now common in much work on biomedical terminologies and ontologies in speaking of concepts as being linked together also by *ontological* relations such as *caused by* or *site of* or *treated with*. As I am sure he would be the first to accept, sets of synonymous terms do not stand to each other in causal or locational or therapeutic relations, and in fact in allowing the latter it seems that Cimino is embracing elements of the *ontological* view of concepts according to which concepts would be abstractions from entities in reality.

The ontological view has advantages over the linguistic view of concepts 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 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 as to what it is in the world to which disorder and similar concepts would correspond.

Concepts are Ethereal

The problematic features of common uses of the term ‘concept’ are not peculiar to the world of biomedical terminology. Indeed they arise quite generally in the literature of knowledge representation in work on semantic networks and conceptual models, where again concepts (variously called ‘classes’, ‘entity types’, ‘object types’) are called upon to perform a plurality of conflicting roles. On the one hand they are delegated to represent inside the computer concrete entities and the classes of such entities existing in reality. Some abstract proxy – some ghostly *diabetes counterpart* – is required for this purpose, it is held, because *one cannot get the diabetes itself inside the computer*. On the other hand they are delegated also to play the role of representing the ‘knowledge’ in the minds of human experts. This knowledge is then itself characteristically (and again erroneously [4]) assumed to be identifiable with the meanings of the terms such experts use, and in this way the painful polysemy of ‘concept’ carries over also to the word ‘knowledge’ and its cognates.

Because concepts are pressed into service to perform so many different roles, they acquire certain remarkably ethereal qualities. Concepts, in fact, are triply ethereal – existing in a different sort of denatured guise in, respectively, the machine, the human mind, and among the meanings stored in language. Their ethereal nature implies that concepts are not the sort of thing that can be examined or inspected. Thus we know what it means to raise and answer questions about, say, a case of diabetes, or about the disease *diabetes* itself – we can turn towards both of these things by directing our attentions to corresponding entities in the world; we can make the *what it is on the side of the patient* the target of our mental acts; we can concern ourselves with its traits, compare and weigh the separate views advanced by different observers. But it seems that we can do none of these things in relation to entities in the realm of concepts. What would it be like to *examine* a concept and compare the way it appears now from this side, now from that? If, as the pertinent literature in philosophy and psychology would suggest [5], concepts are most properly understood not as *targets* of our cognitive acts but rather as their *contents*, then our puzzlement in the face of such questions is understandable. The concept orientation, however, rests however precisely on the assumption that concepts would serve as targets – indeed as the primary targets of work on terminologies.

International Standard Bad Philosophy

The concept orientation in terminology work goes back at least as far as the 1930s, when 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. [6]

Wüster himself defended not a linguistic but a *psychological* view of concepts. That is, he saw concepts as mental entities, sometimes writing 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. [7]

A concept, for Wüster, is an *element of thought*, existing entirely in the minds of human subjects. An individual concept is (he says) a mental surrogate of an individual object; a general concept is a mental surrogate of a plurality of objects. [6] General concepts reflect similarities between objects which human beings are able to apprehend through perceptual experience. They are human creations, arising as the cumulative reflection of what Wüster also 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 views). In some passages Wüster himself seems happy to identify characteristics with *properties* on the side of the objects themselves. In others, however, he identifies them as further *concepts*, so that they too (again incoherently) would exist in the heads of human beings. [6]

The same uncomfortable straddling of the realm of mind (ideas and 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, *instances* of pneumonia; but 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).

Another characteristic unclarity of Wüster's thinking is reflected in his definition of 'object' as 'anything to which human thought is or can be directed.' This definition has been given normative standing through its adoption in the relevant ISO standards, which similarly define an object as 'anything perceived or conceived', stressing thereby that 'object' can embrace, in Wüsterian spirit, not only the material but also the immaterial, not only the real but also the 'purely imagined, for example, a unicorn, a philosopher's stone or a literary character' [8]. With this, we 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. Indeed its 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.

The same problems still present themselves in more recent ISO documents, where an effort has been made to increase clarity by embracing elements of the ontological reading of the term ‘concept’. Unfortunately, however, this is done in such a way that remnants of the other views are still allowed to remain. Thus in ISO 1087-1:2000 ‘concept’ is defined variously as a ‘unit of thought constituted through abstraction on the basis of properties common to a set of objects’, or as a ‘unit of knowledge created by a unique combination of characteristics,’ where ‘characteristic’ is defined as an ‘abstraction of a property of an object or of a set of objects’. Since ‘object’ is still defined as ‘anything perceivable or conceivable’, however (a unicorn still being listed as a specific example of the latter), the clarificatory effects of this move are once again rendered nugatory.

Focusing terminology development on the study of concepts is, for Wüster and for ISO terminologists, a way of shouting ‘Stop!’ in the attempt to stem the tide not of language but of human cognition – effectively by inventing a new realm – the realm of concepts – in which the normal ebb and flow of thought would have been put out of action. It thereby, as Temmerman argues [9], stands in conflict with many of the insights gained through research in cognitive science in recent years. Wüster’s account of concept learning especially, and his insistence on the arbitrariness of concept-formation, rest on ideas that have long since been called into question by cognitive scientists. Even very small children manifest in surprisingly uniform ways an ability to apprehend objects in their surroundings as instances of natural kinds in ways which go far beyond what they apprehend in perceptual experience. Thus there is now much evidence (documented for example in [10]) to the effect that our ability to cognize objects and processes in a domain like that of biology rests on a shared innate capacity to apprehend our surrounding world in terms of (invisible) underlying structures or powers (whose workings we may subsequently learn to comprehend, for example through inquiries in genetics).

Castles in the Air

A further illustration of the problems associated with the concept orientation is provided by [11], in which Keith Campbell, Diane Oliver, Kent Spackman and Edward Shortliffe present their account of the status of the Unified Medical Language System (UMLS) in current terminology work.

The UMLS, as is well known, 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 the same concept unique identifier (CUI) to both ‘aspirin’ and ‘Aspergum’.

The thesis of [11] is that this is allowed because there is a Possible World (the authors cite in this connection the work of Leibniz) in which ‘aspirin’ and ‘Aspergum’ refer to

one and the same thing. Of course, as its authors admit,

> many clinicians would not regard different formulations of aspirin ... as interchangeable concepts in the prescriptions they write. Although aspirin may be an abstract concept, Ecotrin and Aspergum have specific formulations (extensions) in our corporeal world, and use of those particular formulations is subject to different indications, mechanisms of therapy, and risks to the patient. Clearly then, in at least a pharmacy order-entry system, any extensional relationship that was used to determine allowable substitution of pharmacologic formulations would need to have different relationships (representing a different Possible World), than the one currently embodied within the UMLS. However, for a system primarily concerned with the active ingredients of a drug, such as an allergy or drug interaction application, the Possible World embodied in the UMLS may be optimal.

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 the same high-plasticity conceptual realm that is postulated by Wüster and his colleagues, a realm in which aspirin 'may be an abstract concept'. In [11] the UMLS is itself correspondingly referred to as an 'artificial world', as contrasted with 'our corporeal world' of flesh and blood entities. And 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.'

Three worlds have hereby been distinguished:

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

How can world 1. link worlds 2. and 3. together? The answer, surely, must involve some appeal to the *extensions* of the concepts in the UMLS, understood as collections of the individual objects (actual patients, actual pains in actual heads, actual pieces of Aspergum chewed) in the corporeal world. The authors themselves suggest a reading along these lines when they point out, in regard to the terms existing in the UMLS source terminologies, that

> [o]n the one hand there are the physical objects to which [an expression like 'aspirin'] 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).

When it comes to the UMLS itself, however, they abandon this traditional view 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:

> [t]hese 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.

In thus interpreting 'extension' in Wüsterian fashion, however, which means in conceiving extensions in abstraction from the corresponding instances in reality, our authors deny 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.

They face further problems, too, in regard to their account of how the UMLS was constructed:

> 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.

Because the creators of the UMLS often could not grasp what the developers of its source terminologies had in mind just by looking at the constituent terms, and because they could not – à la Wüster – climb into the brains of those developers, they attempted to tease out the relevant information by examining certain clues left behind in the course of terminology creation.

In ICD-10, for example, we have "C75.0: parathyroid gland," a term which on the face of it belongs to the domain of anatomy. However, the associated term C75 reads: "malignant neoplasm of other endocrine glands and related structures." Thus what the developers of ICD-10 really had in mind with C75.0 is "malignant neoplasm of parathyroid gland." In building the UMLS, accordingly, care was taken to ensure that C75.0 would be linked to the UMLS CUI that stands for "malignant neoplasm of parathyroid gland" and not to the UMLS CUI for "parathyroid gland."

The application of this methodology is seen as bringing the advantage that the users of the UMLS will face a reduced likelihood of confusion brought on as a consequence of the bad term-formation principles used in some of its source vocabularies. Unfortunately, however, [11] provides no indication as to how those who applied the methodology were able to tell the difference between those clues which reveal what the developers of source terminologies 'had in mind' and those other clues which reveal merely aspects of their thinking clouded over by bad term-formation principles.

The answer from the side of the realist conception is clear: in constructing the UMLS its authors were able to separate out the 'good' clues because they were implicitly taking into account their own knowledge of the corporeal world to which the terms in question

were, however, inchoately pointing. The universals in reality with which all physicians are at least tacitly familiar from both their training and their everyday experience thereby played the role of anchors. No such answer is available, however, to those who are attempting to work within the constraints of the concept orientation. The text of [11] accordingly merely describes in neutral terms a variety of clues different types, telling us that they

> take 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 developers 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.

It is tempting to caricature this proposal as consisting in the thesis that we learn the meanings of the terms in the UMLS *by examining old typesetting tapes*. While this is not quite fair, it does bring out in vivid fashion the degree to which our authors have failed to provide an adequate account of how, on their own terms, the UMLS could have been constructed.

In hindsight we can see that, with their talk of the UMLS as building a bridge between computers and corporeal reality, Campbell, Oliver, Spackman and Shortliffe have projected on to the UMLS a goal more ambitious than that which it was really intended to serve. Its actual goal was that of finding 'unified meaning' across terminologies. This weaker goal has proved unrealizable, for the same reason that the concept orientation in general is unrealizable (though there may be some practical value in its imperfect realization, for example in expanding the number of synonyms that can be used to find a target term in a specific terminology). We are still free, however, to readdress the more ambitious goal of building a bridge between computers and corporeal reality, a goal which, with the ineluctable expansion in the use of computers in clinical care (and especially in evidenced-based medicine), becomes ever more urgent.

The Role of Logic

One obstacle to its realization is the poor logical resources bequeathed by the early advocates of the concept orientation to their terminologist successors.

The logical tools needed to deal in a formal way with real-world instances, and to relate such instances to the general terms used in terminologies, have been available since Frege's invention, in 1879 [12], of the logic of quantifiers, which has constituted the standard in formal logic ever since.

Unfortunately Wüster, in laying down the central principles of the ISO standard for terminologies, adopted not the Fregean logic, but rather an older, and weaker, 'concept logic' propagated *inter alia* in the work of Kant, in which real-world objects play no essential role.

In Frege's logic we can quantify over instances. Thus for example we can formulate assertions like

every instance of *A* is part of some instance of *B* (commonly abbreviated: *A part_of B*)

or

every instance of *B* has some instance of *A* as part (abbreviated: *B has_part A*)

by using universal (∀) and existential (∃) quantifiers, variables (*x*, *y*, ...) ranging over instances, and the primitive instance level relation **part_of**, as follows:

∀*x* (if *Ax* then ∃*y* (*By* & *x* **part_of** *y*))

∀*y* (if *By* then ∃*x* (*Ax* & *x* **part_of** *y*)).

That these assertions are logically distinct is seen, for example, if we consider that for *A* = *cell nucleus* and *B* = *cell* the first is true, but the second, false [13].

Concept logic, in contrast, because it deals with ('associative') relations between concepts without any concern for underlying instances, has proved itself unable to capture this logical distinction. It is the same concept logic approach which underlies the various 'semantic networks' constructed in the AI field in the 1970s, which were associated initially with considerable optimism concerning their potential to support what is still called 'knowledge representation and reasoning'. The dawning awareness that this optimism was misplaced coincided with the initial experiments in the direction of what would later come to be called Description Logics [14]. The latter fall squarely within the Frege tradition – they are effectively a family of computable fragments of Fregean first-order logic – and thus they, too, have some of the resources needed to deal with reasoning about instances. Unfortunately, however, while instances do indeed play a role in the DL world, the instances at issue are not of this world; thus they are not instances of the sorts encountered, for example, in clinical practice. Rather, work within the DL community has been focused on certain mathematical proxies for real-world instances which exist inside artificial models created *ad hoc*, not for the ontologist's practical purposes of relating a terminology to instances in reality, but rather for the logician's technical purposes of testing consistency and other properties of their systems. Thus also the DL community has not paid attention to the treatment of instances in different ontological categories, for example to the differences between instances of attribute kinds (your temperature, your blood pressure) and instances of event kinds (your breathing, the beating of your heart). Applications of DL-based formalisms in medical terminologies such as GALEN, SNOMED CT and the National Cancer Institute Thesaurus have similarly not exploited its resources for reasoning about instances, in spite of the fact that one central purpose for which such terminologies were devised is to support the coding of clinical records (which relate, as one might suppose, precisely to instances in reality).

How Terms are Introduced into the Language of Medicine

To find our way back to the domain of such real-world instances, consider what happens 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 becomes 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. This kind is 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 exemplifiable in an arbitrarily 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 diagnosis possible, as they allow uniform treatments and associated clinical guidelines to be applied to pluralities of diagnosed disorders encountered in different times and places.

The problem, of course, is that it is in many cases difficult to grasp what universal given particulars are instances of. When a disease universal first begins to make itself manifest in a family of clinical cases 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.

There arises then the need for a new term to refer to the kind newly visible in a given family of cases. Eventually 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 WordNet-style synset, a handy shorthand for a grouping together of terms in different natural languages synonymous with 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. Responding, however, to the edicts of ISO and of healthcare messaging and W3C-style standards bodies, terminologists see themselves as exploring from the start, not the realm of instances and universals in reality, but rather the special realm of concepts. Concepts and definitions are thereby de-anchored from the world of universals and their instances and they begin to acquire a life of their own. Terminologies come to be regimented by taking concepts as benchmark, rather than by reference to the *what it is on the side of the patient* out there in corporeal reality.

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 which underlie many uses of the term 'concept' in terminology circles, proponents of the concept orientation may find it difficult to understand why we are insisting so forcefully on the distinction between concepts and universals. One reason is that there are very many terms in medical terminologies which may be associated (somehow) with concepts but not with any universal in reality. There are no universals corresponding, for example, to terms from ICD-9-CM such as

probable suicide

possible tubo-ovarian abscess

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. Rather, they have the status of disguised sentences pertaining to our ways of gaining knowledge of such entities [15]. Cimino, too, seems to acknowledge part of what we are getting at here with his desideratum rejecting *not elsewhere classified* terms. The latter are not intrinsically problematic if they are called upon to correspond just to concepts (and if they are coded syntactically in such a way as to avoid the problems associated with changes in successive versions of a terminology). They are problematic only when evaluated in light of their potential correspondence to universals in reality.

Problems with Definitions

Concepts, according to the original ISO-Wüster paradigm, are formed when we identify the properties/characteristics given objects share by perceiving and grouping similarities between them. We know what concepts are, on this view, by picking out certain characteristics shared in common by similar objects and joining these together in what are called *intensional definitions*. This account works well enough in the area of woodworking equipment, where Wüster worked out his ideas on concepts and definitions (though they would work less well if the domain of objects were conceived, ISO-style, as including also *immaterial* woodworking equipment). It works well, too, in a domain like chemistry, for molecular structures can indeed be precisely and unproblematically defined in terms of exactly repeatable patterns. Unfortunately however it confronts an obvious problem in the domain of medicine. For there we often have to deal with instances of a single universal which manifest few characteristics identifiable in encounters of similars (consider, for example, an embryo in the successive phases of its development, or the successive phases in the development of even simple diseases). In fact, of course, the provision of intensional definition for the terms in a medical terminology reflects the ideal case only. So many of the concepts in systems like

SNOMED-CT remain undefined [16] because in this, the real world of real instances and of real and pervasive clinical ignorance, it has often been found impossible to reach agreement on how to define them.

The definition of a term *p* we can identify as a statement of the necessary and sufficient conditions for the correct application of *p*. For many medical terms, only some small number of necessary conditions have thus far been identified. Yet even in those cases a definition is sometimes created from out of the air, often simply by permuting the constituent words of the term to be defined (as for example when, in the National Cancer Institute Thesaurus, *cancer death rates* is defined as 'mortality due to cancer'). Merely partial definitions are sometimes taken as descriptions of what the corresponding entities in flesh-and-blood reality *ought* to be (of what they are agreed upon to be) – rather than as expressions of (partial and still amendable) views of what they actually *are* and thus of the real-world conditions under which it is appropriate to use the corresponding term. It is as if SARS is *defined* as: *severe acute respiratory syndrome*, when in fact SARS is much rather *this particular severe acute respiratory syndrome first identified in Guangdong, China in February 2002 and caused by instances of this particular coronavirus whose genome was first sequenced in Canada in April 2003*.

Where intensional definitions are erroneously allowed to take the place of such case-based specifications, the relation to real-world instances is obscured and the user of a concept-based system is encouraged, again, to be content with access only to the abstract nodes in the ethereal realm of concepts. This development is then reinforced still further by the fact that concepts and definitions are seen as something that can be stored electronically and as being ideally suited to serve as units for symbolic processing (a fact which goes hand in hand with what often turn out to be erroneous expectations as concerns the powers of software to detect and eliminate errors in terminologies [17]).

The Concept Orientation Makes Terminology Work Seem Suspiciously Easy

According to the concept orientation, the meanings of preferred terms in a terminology would belong to a realm whose denizens exist as products of agreement; according to the realist orientation here defended 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 they mean the same thing (for example because differentiating clinical manifestations were initially misinterpreted). Nothing in the realm of concepts (or of the ideas or beliefs of language users) would inform us of this coincidence; rather it takes arduous inquiry. We then rely at every stage on the instances in the world to which we have gained access to tell us what the meanings of our general terms are (which is to say: what universals they correspond to), in an empirical process of discovery that is never brought to an end [4,18].

It must be admitted that this realist approach will not work for those terms in our terminologies (*prevented pregnancy, replacement of unspecified heart valve, other life circumstance problem, other European in New Zealand*) which are not associated with any claim to correspond to universals in reality. How such terms are to be dealth with from the realist point of view is a difficult issue, which would here lead us too far from our goal. In what follows, therefore, we focus exclusively on those terms which are clearly associated with such a claim.

As we have seen already above, many such terms are introduced and used before their users have any clear understanding of what they mean and thus before any definition can properly be formulated. 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 is difficult to convey this information to others. 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*.

The terminologist seeks to plug this gap by aligning on concepts. Where a universal has not yet been identified, a concept is postulated in its stead. Each general term is identified either as a preferred term for this concept or as a synonym of some preferred term already recognized. The resulting synsets can then be called upon to ensure that every general term *p* has its own precisely tailored referent, called '*the concept p*' – a referent which is guaranteed to exist (in the 'realm of concepts') even when the term in question has no application to either universals or instances in reality. This, however, amounts to the abandonment of the goal of coming to grips with reality (with this corporeal world) and of substituting instead the goal of reaching consensus on use of words.

The conceptualist approach thereby makes the task of building a terminology suspiciously easy to complete: if a terminology captures our consensus concept of, say, HIV-AIDS, then, so long as this concept does not change, the terminology remains correct even if the corresponding universal has in fact evolved (as it were behind the scenes) in such a way as to be replaced by a plurality of distinct universals. This throws new light on one Cimino desideratum, that of concept permanence: *the meaning of a concept, once created, is inviolate*. Cimino himself embraced this desideratum for reasons having to do with the need always to be able to understand records from the past incorporating codes for terms now rendered obsolete. (Perhaps it is possible to preserve such understandability by means of syntactic devices along the lines sketched in [19] – devices of a type which were not used, as we have seen, by the developers of many of the source terminologies of the UMLS.) When we conceive the desideratum of concept permanence from a more forward-looking perspective, however, then it becomes exposed as practically unrealizable, since it would involve matching up terms not just with the putative concepts which human beings happen to have agreed upon at given stages in the advance of science, but rather with the way the world itself is. It is however precisely this

goal that is characteristic of the scientific method – and it is precisely this latter goal that is tacitly abandoned by the advocates of the concept orientation.

Tracking Referents

We can summarize our arguments thus far as amounting to the thesis that currently predominant approaches to terminology development have in common that they are top-down (word-to-world) approaches, in which it is concepts (terms, meanings) which wear the trousers, while the real-world phenomena to which such concepts relate receive correspondingly short shrift. We owe the predominance of this top-down approach to a peculiar constellation of forces, which has too long dictated the rules of play in the terminology domain, above all:

> (i) the needs and expectations of computer programmers and of the builders of 'semantic networks' and 'conceptual graphs',
>
> (ii) Wüsterianism and the ISO terminology standards,
>
> (iii) a variety of other hold-overs from an earlier era when an 'electronic dictionary' just meant a dictionary whose pages are readable on a computer screen.

Clinicians, on the other side, who one might imagine to be the natural supporters of a bottom-up approach which begins with actual clinical cases, have made their voice too little heard in counteracting these forces. They of course have full and direct access to actual cases, and so they have not noticed the degree to which general proxies for such cases have usurped their role in computer applications.

To bring about the needed rectification, we must turn the concept-based approach on its head. We need to begin, in other words, not with terms, concepts, meanings, ideas, and typesetting tapes, but rather with what the physician is confronted by at the point of care – with disorders, lesions, disease processes, therapies – and to build up from there to terminologies. We need to focus, in other words, on particular entities of a range of different types, including both continuant objects such as cells or fractures or inflammations and occurrent processes such as disease histories or rises in temperature or the clottings of particular portions of blood.

We can then distinguish two interrelated sorts of biomedical knowledge: the *general* knowledge recorded in the textbooks of biomedical science, and the *particular* knowledge recorded in clinical records. More precisely (since of course some of what is recorded in textbooks and in clinical records will at any given stage reflect not *knowledge* but rather *false beliefs*), we have two collections of assertions deriving respectively from scientific texts and from clinical records. Assertions in the first group pertain to universals; those in the second to particulars or instances. If a particular tumor in John's

lung is an instance of the universal *tumor* then it is in relevant respects (to be more precisely specified for each type of universal [20]) similar to other tumors.

While universals are invariants in reality which exist independently of our human cognitive activities, terminologies are 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. (This is distinct from a representation of concepts, not least because there will at any given stage in each domain of science be many concepts in relation to which the consensus view is that they do *not* correspond to universals in reality.)

Each terminology comprehends within its domain 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), biological processes (disease pathways, processes of development and growth), and more. And for each of the latter there is a corresponding family of particulars which instantiate the universal in question.

Most existing systems for keeping track of clinical phenomena, on the other hand, allow direct reference to just two sorts of particulars in reality: (i) to the *human beings* (patients, care-providers, family members) via proper names or via alphanumeric patient IDs, and (ii) to the *times* at which actions are performed or observations made. [21]

This impoverished repertoire of types of direct reference to particulars 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 correspondingly 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 inferences from the existence of different instances of the same clinical universal in different patients, for example for public health purposes [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 thus be created anew, via some combination of general terms (or associated codes) with designators for persons and times, for example in expressions like: *the fever of patient* #1001 *observed by physician* #4001 *at time* #9001. Unfortunately, such composites, even where they are formulated by the same physicians using the same general terms deriving from the same coding system, constitute barriers to reasoning about the corresponding entities in software systems. (Imagine a regime for reasoning about human beings as they change and develop 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*.)

To make the corresponding instances directly visible to reasoning systems (which means: visible without need for prior processing), and thus to give extensions in corporeal reality their due in the creation and use of biomedical terminologies, we need to create a regime in which explicit alphanumerical IDs – analogues of proper names – are 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 [21]. 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, too, we become able to do justice to the ways in which different views of one and the same instance of a given disorder may become incorporated into the record, for example when physician A writes 'tumor' and physician B writes 'CAAA12'. The use of IUIs allows us to *map* the corresponding particulars in our computer representations in a way which will (as far as possible automatically) makes it clear when different physicians are referring to one and the same particular – and the cumulative result of such use can indeed be understood as a map of the domain in question, with some of the properties of the maps we use to represent instances in the geospatial domain [23].

Terms Should be Aligned not with Concepts but with Universals in Reality

To find an appropriate way of treating *universals* in a domain like that of biomedicine will require, now, that we take account of the fact that not only terminologies but also (and in consequence) the clinical records coded with their aid are subject at every stage to the possibility of revision. Each terminology, like each clinical record, is a work-in-progress, reflecting the more or less secure beliefs held at the pertinent stage in the development of biomedicine about how particular entities in reality are to be classified as instances of universals.

Note that this recognition of the fact of constant revision on both the terminology (universal) and EHR (instance) levels is perfectly compatible with the realist view (which also underlies the methodology of modern science) according to which both the vast and settled majority of the beliefs expressed in biomedical texts, and the vast majority of assertions captured in clinical records, are both true and uncontroversial. 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, a broad accumulation of knowledge.

Mixed in with this knowledge, however, there is at every stage a small and ever-changing admixture of false beliefs. The part of this admixture which 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. Our realist counterparts of the three central Cimino desiderata:

– each preferred 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')

are now not realizable by any terminological adjustments motivated merely by considerations of meaning and language. Rather, they need to be accepted as long-term goals, to the ever closer but never quite complete realization of which terminologists are forever condemned. In moving towards their realization they must rather always follow 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 these artifacts, and to this end ways must be found to keep track of time in two different ways, so that *changes in reality* and *changes in our scientific beliefs about reality* can be monitored independently. This idea involves nothing that is essentially new: we already track events of many sorts (for example observations) by indexing with times; and we track changes in terminologies (which here go proxy for our 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.

Terminologies Consist of Terms

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 *preferred term*, S_p a set of *synonyms*, and *d* an (optional) *definition*,

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 or ordered pairs $\langle p, q \rangle$ of those preferred terms for which '*p r q*' represents a consensus assertion of biomedical science about the corresponding universals at the time when the given terminology is prepared,

and

v is a version number, which encodes this time.

The variables *p, q, d, r, v, …*, on the account here defended, range over syntactic entities (strings of characters in some regimented language). The concept orientation yields an account of terminologies in which these variables can be seen also as ranging over concepts in virtue of the one-one correspondence between preferred terms and concepts. On our realist account, in contrast, no such supplementary reading is possible. This is because some values of *p, q* ... will correspond to no universal in reality (and thus, like 'unicorn' or 'phlogiston' they will be *empty names*, which correspond to nothing in reality at all). Others will correspond to too much (i.e. they will refer ambiguously to a plurality of universals). We need to take both these alternatives into account when we consider the whole terminology $T = \langle N, L, v \rangle$ in light of its status as a map of an analogous structure of universals in the corporeal world.

While in the ideal state of terminological virtue we could indeed associate *N* in one-one fashion with some corresponding set *U* of the universals designated by its constituent nodes, 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. This means (roughly, and for our present purposes) that at any given stage its preferred terms will be divided into three groups:

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

where N_1 consists of those preferred terms in *N* which correspond to exactly one universal, $N_>$ of those preferred terms in *N* which correspond to more than one universal (in various combinations), and $N_<$ of those preferred terms in *N* which correspond to less than one universal (again in various combinations, and in the simplest case 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*, though this assumption must be qualified in reflection of the fact that *N* is itself changing, as is the totality of biomedically salient universals in the reality beyond.

Our knowledge of the successes of medical science gives us strong reason to believe that N_1 constitutes a large portion of *N* (*N*, remember, is a collection of terms already in use which are associated with the claim to represent a biomedical universal, including the very many uncontroversial terms which we are normally inclined to overlook). 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 just 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* consists of the low value $N_>$- and $N_<$-type terms. The reason for this is clear: if we did

know where these 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 different position.

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. Thus the *whole* of N is, as far as the developers and users of a given terminology are concerned, such as to consist of *names of universals.*

But if we do not know how the terms are presently distributed between the three groups, then does this mean that the distinction between N_1, $N_>$, and $N_<$ is of purely theoretical interest – a matter of abstract (philosophical) housekeeping that is of no concrete significance for the day-to-day work of terminology development and application? Not all all. For recall that we will typically have at our disposal not just one version of a terminology, but a whole, developing series. In uncovering errors immanent to a terminology, we thereby uncover terms which must be excluded from future versions because they do not correspond to universals. Given the resources of our realist approach, however, we do not need to wait for the actual discovery of error. For we can carry out experiments with terminologies themselves, which means that we can explore through simulations the consequences of different kinds of mismatch between our terms and reality.

A Formal Framework for Terminology Experimentation

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 unknown – and the challenge is then to *solve* for this unknown. (Compare 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.) We can now see that there are three different kinds of solution which can present themselves: the cases in the pool – not patients, remember, but cases of 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 cross-references to the identifiers assigned to those other particulars (including the relevant patients) with which the entity under scrutinity is related, for example in the ways catalogued in [21]. Coordinates in the vector would include also the measured values of medically salient attributes such as temperature,

blood pressure, etc., as well as gene expression and other bio-assay data (for instance combining molecular imaging with clinical radiology), each coordinate being indexed by time of entry, source, and estimated level of evidence.

In more formal terms, we can define an *instance vector* as an ordered triple

$$<i, p, t>$$

consisting of a IUI *i* [22], a preferred term *p* in a terminology, and the designation of a time at which the particular designated by *i* is asserted to be an instance of the universal (if any) designated by *p* (the variables *i*, *p*, *t*, ... range once again over strings in a regimented language).

The IUIs in our repository will typically already have been associated at the point of entry into the EHR with preferred terms from one or more terminologies. For example the EHR will contain the assertion that tumor instance #5001 is 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 period), we can now define a *t-instantation* $\boldsymbol{I}_t(T, D)$ 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 *p* in the particular domain of reality selected for by *D* at the time *t*.

For each term *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 stands to other instances represented by IUIs in *D*. Our three alternative scenarios now once again present themselves, according to the status of each preferred term *p* in relation to the world of actual cases (the world which serves as standard 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.

Reasoning with Instance Identifiers: Three Applications

A system of vectors along the lines described could be used first of all for purposes of quality-control of terminologies (and thus for purposes of automatically generating improved versions of terminologies). 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 when measures are applied to instances for which we have similar amounts of data of similarly high evidence-value, 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.

Such revision of terminologies will in due course spawn, in the opposite direction, revisions 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 lead in turn, we believe, to better quality clinical record data, which may in turn spawn yet further revisions in our terminologies.

Such methods for reasoning with terminology and instance data might be used, secondly, for purposes of decision support in the process of diagnosis. One goal of an adequate terminology-based reasoning system, in a world of abundant instance data, would be to allow the clinician to experiment with alternative term-assignments to given collections of instance data in ways which would allow measurements, on the basis of statistical properties of the patterns of association between terms and instances which result, of the greater and lesser likelihood of given diagnoses. Thus we could imagine software which would allow experimentation with alternative IUI and term assignments, for example when it is unclear whether successive clusters of symptoms of a given patient should be counted as manifestations of single or of multiple disorders. The machinery of instantiations could then be used to test out alternative hypotheses regarding how to classify given particulars by giving us the facility to experiment with different scenarios as concerns the division between N_1, $N_<$, and $N_<$ in relation to given cases.

In the real world, of course, such methods cannot be applied successfully in every case. For example we may not have all the data needed to convince a computer armed

with a given stock of universal terms and associated instance data that a given case meets the requirements for any available diagnosis. The case is then however no different from that which is 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. He learns how to do this, first, from medical textbooks and education, then through experience and by following guidelines and protocols.

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 the patient nose-length in the way that we do now for, say, temperature or blood pressure? The answer is that we do so already. Many hundreds of thousands of patients 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 instance-level data in other domains.

Conclusion

In the ideal case, a biomedical terminology would provide not merely the resources for assigning preferred terms for universals to the corresponding 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 automatically partitioned at the point of data entry in ways reflecting the structure of the world of clinically relevant universals. This partitioning of instances is currently masked from view in the clinical record because the instance-level data that exists in separate EHRs is currently available 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 level of both instances and universals [13]. 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”.

References

1. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
2. 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.
3. Fellbaum C. (ed.) WordNet. An electronic lexical database. Cambridge, MA: MIT Press; 1998.
4. Putnam H. The meaning of ‘meaning’, in Putnam H. Mind, language and reality, Cambridge: Cambridge University Press; 1975, p. 215-271.
5 Margolis E, Laurence S (eds.). Concepts: Core Readings. Cambridge MA: MIT Press, 1999.
6. Smith B, Ceusters W, Temmerman R. Wüsteria. Medical Informatics Europe (MIE 2005), in press.
7. Wüster E. The wording of the world presented graphically and terminologically. Terminology 2003;9(2):269-297.
8. ISO/IEC JTC1 SC36 N0579 Information Technology for Learning, Education, and Training.
9. Temmerman R. Towards new ways of terminology description. Amsterdam: Benjamins; 2000.
10. Gelman SA, Wellman HM. Insides and essences: Early understandings of the non-obvious. Cognition 1991;38:213-244.
11. 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.
12. Gottlob F. Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens, Halle 1879. English translation by TW Bynum in: Conceptual notation and related articles, Oxford: Oxford University Press; 1972.
13. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector AL, Rosse C. Relations in Biomedical Ontologies. Genome Biol, 2005;6:R46.
14. Nardi D, Brachman RJ. An introduction to Description Logics. In F. Baader, et al (eds). The Description Logics Handbook, Cambridge: Cambridge University Press; 2003, p. 1-40.
15. 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.
16. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in DL-based terminologies: A case study in SNOMED CT. In: Hahn U, Schulz S, Cornet R, eds. Proc First Int Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004), 2004;:12-20.
17. Ceusters W, Smith B, Flanagan J: Ontology and medical terminology: Why Description Logics are not enough. TEPR 2003 Towards an Electronic Patient Record; San Antonio, 2003 (electronic publication).
18. Putnam H. Is water necessarily H_2O? in H. Putnam, Realism with a Human Face, Cambridge MA: Harvard University Press, 1990.
19 CEN/TC 251/WGII Terminology and knowledge bases. CEN Document WGII/N03-02 (2003).
20. Welty C, Guarino N. Supporting ontological analysis of taxonomic relationships, Data and Knowledge Engineering, 39 (1); 2001: 51-74.

21. Ceusters W, Smith B. Tracking referents in electronic health records. Proceedings of MIE 2005, in press.
22. Ceusters W. Strategies for referent tracking in electronic health records. WG6 Conference Presentation.
23. Smith B, Mejino JLV, Schulz S, Kumar A, Rosse C. Anatomical information science, in A. C. Cohn and D. Mark (eds.). Conference on Spatial Information Theory (COSIT 2005), in press.