An Ontology-Based Methodology for the Migration of Biomedical Terminologies to Electronic Health Records

Barry Smith, PhD[a,b], Werner Ceusters, MD[b]

[a]Department of Philosophy, University at Buffalo, Buffalo, NY 14260, USA

[b]Institute for Formal Ontology and Medical Information Science and European Centre for Ontological Research, Saarland University, 66041 Saarbrücken, Germany

<Please do your best to make this smoother and to incorporate responses to reviewers comments, especially regarding computational problems of RT.>

Biomedical terminologies are focused on what is general, *Electronic Health Records (EHRs) on what is* particular, *and it is commonly assumed that the step from the one to the other is unproblematic. We argue that this is not so, and that, if the EHR of the future is to fulfill its promise, then the foundations of both EHR architectures and biomedical terminologies need to be reconceived. We accordingly describe a new framework for the treatment of both generals and particulars in biomedical informatics that is designed: 1. to provide new opportunities for the sharing and management of data within and between healthcare institutions, 2. to facilitate interoperability among different terminology and record systems, and thereby 3. to allow new kinds of reasoning with biomedical data.*

The General and the Particular

Much effort has been invested in recent years in the development of structured vocabularies of medical and biological terms. The resultant terminologies, such as SNOMED-CT[i], GALEN[ii], ICD-10[iii], or GO[iv], consist overwhelmingly of general terms ('cell', 'tumor' or 'postoophorectomy with pathological fracture') linked by relational assertions for example of the forms '*A is_a B*' or '*A part_of B*'.

The Electronic Health Record (EHR), in contrast, is a record of particular entities belonging to a wide variety of different general categories. It is a record of particular histories taken, of the particular

attributes, events and processes described in these histories, of the particular symptoms, disorders and associated pathological and non-pathological anatomical structures on the part of each patient, and so on. For all this variety, however, most existing EHR architectures allow direct reference to just two kinds of particulars in reality:

(i) to *human beings* (patients, care-providers, family members) via proper names or via alphanumeric IDs, and

(ii) to the *times* at which actions are performed or observations made.

For particulars in all other categories they provide merely general codes. This limited repertoire of labels allowing direct reference to what is particular means that current EHRs have no adequate means to keep track of one and the same particular (for example a tumor, a breast implant, a shadow revealed on a succession of radiological images) over extended periods of time. When interpreting EHR data it is correspondingly difficult to distinguish clearly between multiple references to the same particular and multiple particulars of the same general kind.[vi]

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 at present be created anew, via some combination of general terms with designators for persons and times, for example in expressions like: *the fever of Patient McX first noted by Patient at* t_1 *and observed by Physician O'Z at* t_2. Unfortunately, the need to use such composites creates logical obstacles to cross-identification of the corresponding entities as they occur in different contexts, and thus also to reasoning about these entities in software systems. This is so especially when we are concerned to keep track of how such entities develop over time, the facility for which is thus for poorly developed in the domains of bio- and medical informatics.

To resolve these problems, we have proposed a new type of EHR regime, in which software would ensure that explicit alphanumerical IDs are automatically assigned (and from that point on, i.e. after being assigned to aparticular, are called IUIs - *Instance Unique Identifier*) in the course of data entry to each individual real-world entity at the point where it first becomes relevant to the treatment and care of a given patient.[vi] Such IUIs would be assigned not just to each particular tumor but also to the gland or duct in which the tumor is located, to each biopsy taken, to each associated radiological image, and indeed (in principle) to instances of all the diagnostically salient categories recorded in the EHR.

Part of this idea has been implemented already within the DICOM standard (for Digital Imaging and

Communications in Medicine), which assigns unique IDs to radiological images, with a range of associated attributes such as *image type*, *patient orientation*, *acquisition time*, and so forth.[v] On the approach here envisaged, however, not only would each separate image receive its own IUI, but so also would each feature appearing on the image at the point where it is identified as medically salient by the practitioners involved.

We have outlined elsewhere the benefits to diagnosis and treatment within a single healthcare institution which may derive from an EHR architecture conceived along the lines described and also the practical problems (for example faced by the practitioner at the point of data entry) involved in its realization.[vi] Here we focus on certain formal aspects of the referent tracking approach, especially in light of its relation to the terminologies, ontologies and coding systems which will need to be constructed if the full promise of biomedical informatics is to be harvested in the future.

Our idea is that each IUI (and thus each corresponding particular entity) stored in the IUI-repository of a referent tracking system will be associated with a vector comprehending both relevant general-term-coding-assignments and also cross-references to the IUIs assigned to each of the particulars (above all to the patients) with which the entity under scrutiny is associated. Each coordinate in the vector would in addition be indexed by time of entry, source, and estimated evidence. (Such coordinates might comprehend also the measured values of medically salient attributes such as temperature, weight, etc., as well as gene expression and other bio-assay data.)

Because the vectors themselves would cross-refer, through the IUIs of other particulars included among their coordinates (for example family members, earlier events in the patient's life history), they would together form a complex graph representing the associations between such particulars as they exist in reality. This graph would in addition contain large amounts of redundancy, of a sort which can be used for cross-checking and thus for error-prevention in relation to the data entered into the associated EHRs.

Because disorders and associated anatomical structures would receive IUIs of their own, independently of any identifying reference to the corresponding patients, our proposal would allow automatic compilation of pseudonymized data pertaining for example to specific kinds of disorders or to the multiple disorders of specific kinds of patients. The ever growing pool of vectors could further be managed in such a way that different kinds of associations between IUIs could be subject to different levels of encryption, thereby allowing new

types of research collaboration based on the automatic exchange and tracking of different kinds of instance data.

In the ideal case, uniqueness of IUIs would be guaranteed by means of the same sorts of mechanisms as are currently used for maintaining the uniqueness of patient IDs or webpage addresses. The vectors pertaining to particular identifiers might be stored in a single (ideally, internationally administered) pool. Or they could be stored locally, but made internationally available for example through some suitably ambitious extension of the methods used in the Life Sciences Identifier (LSID) framework,[vii] or another similar paradigm. In either case, our approach would facilitate the gathering of more adequate statistics on patient care and outcomes, and also allow the application of new types of instance-based data-mining techniques.[viii]

Importantly, the software associated with the pool of vectors would need to have the facility to unravel those ID assignments already made which have been discovered to be erroneous. Where multiple IDs are assigned to what proves to be a single particular it may be possible simply to merge the vectors associated with each separate IUI. In the dual case, where a single ID has been assigned to what turns out to be multiple particulars, software tools would need to be provided which would help the physician or coding specialist to decompose the corresponding vector in such a way that relevant segments come to be assigned to their associated particulars. (In the worst case, the entire existing vector might continue to be associated with each of the new IDs, but now flagged as having a low degree of evidence.) In the same way, the system would need to be able to accommodate the sorts of corrections to the codes contained in specific vectors which become necessary when terminologies themselves change because of scientific advance.

Interestingly, the availability of such tools for the handling and correction of mis-assignments would give coders the possibility to experiment with alternative ID assignments on an experimental basis, 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. Statistical methods of pattern matching applied to these alternatives could then be used as a basis for diagnostic decision support.

By keeping track of the ways in which prior ID assignments have been corrected with the gathering of new information, the system could in principle also learn to associate specific recurring patterns of mis-assignment and correction with corresponding kinds of disorders. Thus it might learn to recognize the characteristic patterns of correction which arise in

the early phases of diagnosis of degenerative diseases such as multiple sclerosis.

The Ontological Problem

Even with an adequate system for tracking referents of the sort described, however, there is an obstacle to the effective migration of biomedical terminologies to the EHR environment, which turns on the currently predominant treatment of the terms and relations in such terminologies.

A biomedical terminology is part of a wider reality, which includes users, records, diseases, patients, acts of observation, acts of coding, and so forth. The proper understanding of this wider reality is, we hold, an indispensable presupposition of an adequate computer representation of the link between the terms in a terminology and the particular instances or cases to be documented with its aid. In the development of many existent terminologies, unfortunately, too little consideration was given to this wider framework and to the need for a clear link, or bridge, between terms in terminologies and instances in reality. <We can add that this bridge is missing, too, from other influential projects, such as the Semantic Web, <OTHERS?> I would not do this. The paper needs to be shortened to fit in the 5p limit>. Rather, relations between terms were (and often still are) conceived in the ways in which dictionary-makers might conceive them, which is to say: as reflecting corresponding relations between *meanings of words*, independent of any reference to instances in reality.

The Semantic Network (SN) of the Unified Medical Language System (UMLS)[ix] provides an illustration of the sort of problem we have in mind in its definition of the *is_a* relation:

Definition. If one item "isa" another item then the first item is more specific in meaning than the second item.

That to which the terms of the SN correspond – called, variously, 'concepts' or 'items' or 'entities' or 'Semantic Types' – would seem in light of this definition to be precisely *the corresponding meanings.* Yet some SN relations seem on the other hand to demand that the relevant relata be *particular instances*. *Part_of*, for example, is defined by SN as: *composes, with one or more other physical units, some larger whole*. *Location_of* is defined as: *the position, site, or region of an entity or the site of a process*. *Co-occurs_with* is defined as: *occurs at the same time as, together with, or jointly*.

SN itself gives no answer to the question what the nodes (the 'concepts' or 'Semantic Types') to which its constituent terms would correspond of a sort which would make its own relational assertions come out true simultaneously. One consequence is that no coherent reading is available as to how the terms of

the SN might link to instances in reality.[x]

The SN is of course not itself designed to be used as a terminology in healthcare records. But the problem of polysemy of the term 'concept' applies to almost all the terminologies included in the UMLS Metathesaurus, where the use of relations like *synonymous_with* or *narrower_in_meaning_than* or *associated_with* or *conceptually_related_to* is at best problematic when terms in terminologies need to be related to corresponding instances.

The absence of a coherent reading of the term 'concept' in standard terminologies has helped to ensure also that the relations which structure these terminologies were introduced in informal and inconsistent ways, so that the logical interconnections between the corresponding assertions were left unclear. We believe that a large part of the errors in coding and documentation can be explained by this [xi], and that it has also blocked those kinds of logical reasoning within and between terminologies and EHRs which would be possible given clear and consistent definitions[xii].

A New Regime of Definitions

How, then, are we to reconceive biomedical terminologies in such a way that they will (1) support the provision of clear definitions of relations such as *is_a* and *part_of* and at the same time (2) allow cascading inferences across such relations and (3) facilitate application to corresponding instances in reality? The answer we propose in [xiii] consists in an ontology of general-purpose relations, applicable across all biomedical domains each of which is such as to allow the provision of simple formal definitions of a type which would be both intelligible to human users and also able to support logic-based tools for automatic reasoning. This Relation Ontology,[xiv] which has now been incorporated into the ontology library maintained by the Open Biomedical Ontologies Consortium, is at this stage a starting-point only, but efforts are underway to expand it to include standardized definitions of all important relations currently employed in biomedical ontology development.

Foundational relations: *is_a, part_of*
Spatial relations (connecting one entity to another in terms of relations between the spatial regions they occupy): *located_in, contained_in, adjacent_to*
Temporal relations (connecting entities existing at different times): *transformation_of, derives_from, preceded_by*
Participation relations (connecting processes to their bearers): *has_participant, has_agent*

Table 1: The OBO Relation Ontology (Version 1.0)

The goal in formulating such definitions, is to provide the optimal combination of intelligibility (in order to ensure maximally reliable curation of the ontologies in which the corresponding relations would be applied) and formal rigor (designed to support logic-based reasoning tools and thus to guarantee maximal leverage in integrating knowledge derived from different domains).

To illustrate the results of this methodology, we here provide examples of relations involving continuant entities, which is to say particulars (such as lungs, diseases, tumors) that endure through time while undergoing changes of various sorts. (For particulars in other categories see [xiii,xv].) Such particulars instantiate *universals*, which are those invariants in reality in virtue of which we are able to describe multiple particulars using one and the same general term. (It is such invariants which make possible, *inter alia*, the use of general terms in scientific inquiry.) The definitions are formulated using variables *c*, *d*, ..., *C*, *D*, ... ranging over continuant particulars and continuant universals, respectively. Because continuant particulars can instantiate different universals and include different instance-level parts at different times (consider, for example, a carcinoma in its successive stages of development), our definitions require also variables *t*, t_1, ... to range over instants of time (assumed to be linearly ordered by a relation **earlier_than**). They require also certain primitive (which is to say, undefined) relations involving entities on the instance level:

c **instance_of** *C* **at** *t* (particular *c* instantiates universal *C* at time *t*)

c **part_of** *d* **at** *t* (particular *c* is an instance-level part of particular *d* at time *t*)

c **located_in** *d* **at** *t* (the spatial region occupied by *c* is an instance-level part of the spatial region occupied by *d* at time *t*).[xvi]

The needed formal definitions of relations between biomedical universals can then be formulated as follows:

C is_a D =def. for all *c*, *t*, if *c* **instance_of** *C* **at** *t* then *c* **instance_of** *D* **at** *t*.

C part_of D =def. for all *c*, *t*, if *c* **instance of** *C* **at** *t* then there is some *d* such that: *d* **instance_of** *D* **at** *t* and *c* **part_of** *d* **at** *t*.

C located_in D =def. for all *c*, *t*, if *c* **instance_of** *C* **at** *t* then there is some *d* such that: *d* **instance_of** *D* **at** *t* and *c* **located_in** *d* **at** *t*.

C transformation_of D =def. for all *c*, *t*, if *c* **instance_of** *C* **at** *t*, then there is some t_1 such that: *c* **instance_of** *D* **at** t_1 and t_1 **earlier than** *t*.

The relation *transformation_of*, for example, serves

the representation of the phenomena of growth, development and pathological change. Where the transformation relation obtains on the instance level, then we have some single continuant particular which instantiates distinct universals at different times in virtue of phenotypic changes.

Note how the definitions listed ensure that the corresponding relational assertions on the level of universals have embedded within them an automatic reference to the corresponding instances and times. This is achieved characteristically via an *all–some* structure,[xvii,xviii] as for example in the definition of parthood, where we have: universal *C part_of* universal *D* when ***all*** instances of *C* have ***some*** instance of *D* as part. The failure to recognize this *all–some* structure is at the root of many characteristic families of coding errors in existing biomedical terminologies[xvii,xviii] (and there is, to our knowledge, only one major terminology which is in its present form not problematic in this respect, and that is the Foundational Model of Anatomy[xix]).

Our analysis of relational assertions appeals only to the simplest resources of quantificational logic, and it has thus long been familiar to those working for example on Description Logics (DLs) such as are used in the GALEN and SNOMED terminologies. Its significance, however, and most especially its application to the problem of binding terminologies to health records, has seldom been appreciated in the relevant user communities.

A New Regime for Clinical Coding

Suppose that we wish to use the resources provided by a terminology such as SNOMED in order to code information about a specific patient suffering from recurring breast cancer. History-taking involves finding ways of referring to instances such as: this present incidence of breast cancer, earlier incidences, present tumors (at successive points in time), earlier (distinct) tumors (at earlier points in time), processes of mastectomy, types of tumor, etc.

The project described in [xx] holds that we can provide some of what we need to achieve this end, not via the systematic referencing of the mentioned particulars but rather *through inferences from statements at a general level*. The idea is that if, for example, a SNOMED term *A* is used at t_1 to describe 'something a physician observed' and at t_2 the same general term is used again by the same physician, and if t_2 is close in time to t_1, then it can be inferred that the physician referred to the same 'something' on two successive occasions.

We find this idea implausible. We propose instead that the physician or other specialist entering data about a given particular in the EHR should first check an instance-tracking database to establish (with the aid of suitable software and the accessible vector

of SNOMED codes already entered) whether the particular in question has already been identified. If that is the case, he simple adds new information to the vector associated with its IUI, using further codes. The proposed framework then makes it easier to decide which codes to use at each successive stage, since the user is presented immediately with the codes already entered in previous descriptions. If, on the other hand, the instance has not yet been described, then a new ID is created which is henceforth subject to SNOMED coding in the usual way.

Yet caution is still needed if one wishes to use a system like SNOMED for more ambitious reasoning purposes. This is because, for the reasons just explained, one often cannot use the relational assertions incorporated within SNOMED on the general level to infer further information about particular instances. SNOMED currently offers no way to tell which relational assertions do support inferences of this sort, and so its relational organization is still best conceived as a convenient mechanism for browsing through the terminology in order to find better descriptors for given instances – not as a representation of how these instances are related together in reality. When the paradigm here advanced has been in use for some time, however, then the accumulated data could be exploited *post hoc* to correct SNOMED's treatment of relations in such a way that it would, by degrees, be in a position to support such inferences. In this way our methodology for the treatment of instances might also lead to improvements in terminologies. But it might also further the goal of interoperability between terminologies and other systems for recording biomedical data. For our paradigm would allow the simultaneous use of a variety of different coding systems within a single record. Indeed the use of multiple codes could yield in automatic fashion an ever-growing system of associations between the terms in the separate systems, reflecting their use in annotating common particulars, in a process which would eventually supplant current efforts to create mappings between such systems.

Conclusion

The instance-based methodology of formal definitions here described is being used by curators of the GO and FMA ontologies and also of the ChEBI chemical entities vocabulary in order to improve the reliability of their respective systems. It has also been applied in systematic evaluations of the National Cancer Institute Thesaurus[xxi] and as a basis for an ontology of colorectal cancer.[xxii] There is however still one serious obstacle to the use of formal definitions to support reasoning with instance-level data, which turns on the fact that such reasoning is

expensive in computational resources. But as is done in description logics, the right approach is to concentrate first on those problems that allow tractable reasoning, and focus on the hard cases later. We are currently testing a prototype reasoner which can help us to overcome this obstacle. Given assertions of specific relations between instances of given types, our prototype calculates, on the basis of definitions in the OBO Relation Ontology, an exhaustive list of all relations which can hold between instances of the types in question. It thereby becomes possible to transform *reasoning* with instance data into *search* across the corresponding relation space, which entails far fewer demands on computational resources. A framework is hereby being built which can, we believe, help in bringing together the distinct ways of treating data that have evolved in the worlds of clinical records and of medical terminology.

Acknowledgements

Work on this paper was carried out under the auspices of the Wolfgang Paul Program of the Humboldt Foundation and the Volkswagen Foundation Project "Forms of Life".

Dealing with the reviewers' comments

>It isn't clear that the boundaries between multiple examples of the same particular and multiple particulars of the same kind are always obvious to the clinician either. This will make this sort of entry tricky.

We changed the problematic part of the sentence from "... multiple **examples of** the same particular ..." to "....multiple **references to** the same particular ..."

>Lacking access to reference (6), it isn't obvious about how systems would go about assigning these identities in a rational, non-combinatorial fashion. Are the authors proposing some sort of RDF-like triple-store or something more elaborate?

That is explained in column 1 of 2nd page

>Typo: "Part of what he have in mind..."

Type corrected

>While a fascinating idea, the pragmatics of managing id pools and their referents seems a tad overwhelming - especially a single, internationally administered pool. The LSID solution is way too simple for this context - its primary function is to map unique identifiers to digital resources - unique bit strings. It is unlikely that LSID could be extended to do this.

Text adapted

>
>While I don't agree with all of ISO TC 37's methods, I think that the author is reading more into ISO's advice to avoid philosophical discussions about whether an object actual exists than is intended. The purpose is to avoid endless discussion about realism vs. phenomenology, constructivism, etc. and to focus on the terms (tokens) used in human communication. In any case, what is the connection between the ISO TC37 specifications and the UMLS Semantic network?

Ref to ISO removed.

>
>Perhaps some references or examples of "systematic errors in coding and documentation" and "blocked kinds of logical reasoning" would be useful? This is a strong assertion with otherwise.

References are given

>
>I'm pleased to see someone advocating a relation ontology - but is it really an "ontology" at this point? The list is quite shallow.

>

But the relations are precisely defined, which is not done in any other "ontology" in healthcare IT thus far.

>I'm also pleased to see a formalization of relation/inductance instance/instance relationships - something that has been woefully neglected in the current DL community - there is no way in OWL, for instance, to say that c has been known to cause b.
>

This comment does not require a change in the paper.

>The idea that we can reach the point that we can infer that the use of term A at two "close in times" involves the same referent seems far-fetched. Contrast diabetes (lifetime) with amour (months to years) to cough (days to years) to wound (hours to years).
>

Absolutely. For this reason, we do not endorse the vision of the author whom we refer to. That has been made more explicit.

>It seems like some of the types of reasoning proposed by the authors have been proved to be computationaly intractable - something that will be difficult to overcome with any reasoner, prototypical or elsewise.
>

approach added on last page

>

>This paper relates an excellent work and is well written.Not easy to read, such as every conceptual/theoritical document. No special comment.
>

>It is not clear how the present paper differs from the other work cited; either it doesn't, or insufficient detail is provided to distinguish it.>

details are given

References

i. http://www.snomed.org.
ii. Rogers J, Rector A. The GALEN ontology. *MIE* 1996;:174-178.
iii. http://www.who.int/classifications/icd/en.
iv. The Gene Ontology: http://www.geneontology.org.
v. http://medical.nema.org/dicom/2004/04_03pu.pdf.
vi. Ceusters W, Smith B. Tracking referents in electronic health records. *MIE* 2005. In press.
vii. Martin S, Smith D, Szekely B. LSID Foreign Authority Notification (FAN) Service Recommendation. http://lsid.sourceforge.net/reference/proposals/fan-rec.html.
viii Ceusters W, Smith B. Strategies for Referent Tracking in Electronic Health Records. Journal of Biomedical Informatics (submitted).
ix. http://semanticnetwork.nlm.nih.gov/
x. Smith B. Beyond concepts: Ontology as reality representation. *FOIS* 2004;:73-84.
xi Farhan J, Al-Jummaa S, Al-Rajhi A, Al-Rayes H, Al-Nasser A. Documentation and coding of medical records in a tertiary care center: a pilot study. Ann Saudi Med. 2005 Jan-Feb;25(1):46-9.
xii Rector AL. Clinical terminology: why is it so hard? Methods Inf Med. 1999 Dec;38(4-5):239-52.
xiii. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C. Relations in biomedical ontologies. *Genome Biology* 2005, 6:R46.
xiv. http://obo.sourceforge.net/relationship.
xv. Rosse C, Kumar A, Mejino JLV, Cook DL, Detwiler DL, Smith B. A strategy for improving and integrating biomedical ontologies. *AMIA* 2004 (in this volume).
xvi. Donnelly M: Layered mereotopology. *Proc IJCAI* 2003;:1269-1274.
xvii. Smith B, Rosse C. The role of foundational relations in the alignment of biomedical ontologies. *Proc. Medinfo* 2004; 444-448.
xviii. Donnelly M, Bittner T, Rosse C. A formal theory for spatial reasoning in biomedical ontologies. *Artificial Intelligence in Medicine*, in press.
xix. Rosse C, Mejino JLV. A reference ontology for bioinformatics: The Foundational Model of Anatomy. *J Biomed Inform* 2004.
xx. http://www.cs.man.ac.uk/mig/projects/current/clef.
xxi. Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus, Meth Inform Med, in press.
xxii. Kumar A, Yip YL, Smith B, Grenon P. Bridging the gap between medical and bioinformatics using ontological principles: A colon carcinoma case study, *Computers in Biology and Medicine*, in press.