A Unified Framework for Biomedical Terminologies and Ontologies

Werner Ceusters, Barry Smith

Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, Buffalo NY

Abstract

The goal of the OBO (Open Biomedical Ontologies) Foundry initiative is to create and maintain an evolving collection of non-overlapping interoperable ontologies that will offer unambiguous representations of the types of entities in biological and biomedical reality. These ontologies are designed to serve non-redundant annotation of data and scientific text. To achieve these ends, the Foundry imposes strict requirements upon the ontologies eligible for inclusion. While these requirements are not met by most existing biomedical terminologies, the latter may nonetheless support the Foundry's goal of consistent and non-redundant annotation if appropriate mappings of data annotated with their aid can be achieved. To construct such mappings in reliable fashion, however, it is necessary to analyze terminological resources from an ontologically realistic perspective in such a way as to identify the exact import of the 'concepts' and associated terms which they contain. We propose a framework for such analysis that is designed to maximize the degree to which legacy terminologies and the data coded with their aid can be successfully used for information-driven clinical and translational research.

Keywords:

Ontology; terminology; mapping.

Introduction

Familiarly, biomedical information is published using multiple different sorts of terminologies, classifications and coding systems. This diversity produces silo effects, which reduce the value of annotations created on the basis of such systems by making data both difficult to access and resistant to integration. Ontologies such as the Gene Ontology, in contrast, seek to overcome these problems by providing corridors of semantic interoperability between distinct information resources [1]. The idea is that, if multiple bodies of relevant information can be annotated using common, non-redundant sets of categories with definitions formulated in some common logical language, then the information they contain will thereby be more easily accessible and more readily capable of being integrated together computationally. This strategy is now increasingly being applied also in the field of human health. [2] Unfortunately, many of the ontologies being employed in specific life science disciplines and in associated clinical specialisms are still built by groups working independently or with no resort to common ontological standards.

Increasingly, one or other version of description logic such as OWL 2.0 is being used in their development. However, the use of a logical representation language alone is clearly not enough to ensure the high quality of an information resource [3], and even ontologies employing the same formal language are often not combinable into a single resource because of incompatibilities between the ways this language is used by different groups. [4]

The goal of the OBO (Open Biomedical Ontologies) Foundry is to counter such tendencies by promoting the creation of a single, expanding family of ontologies designed to be interoperable and logically well-formed and to incorporate accurate representations of biological reality. Ontologies are admitted into the Foundry, and to its on-going process of review, only if their developers commit to an evolving set of common principles [2], of which the most important for our purposes are:

(1) that terms and definitions should be built up compositionally out of component representations taken either from the same ontology or from more basic feeder ontologies;

(2) that for each domain there should be convergence upon exactly one Foundry ontology; [5]

(3) that ontologies should use upper-level categories drawn from Basic Formal Ontology (BFO) [6] together with relations unambiguously defined according to the pattern set forth in the OBO Relation Ontology (RO) [7].

The concept orientation

Concept-based terminologies such as SNOMED CT consist of groups of terms, each such group being linked to a 'concept' that is said to define the meaning of the corresponding terms. We have argued that the inconsistent interpretations of the word 'concept' embraced by the creators and users of such terminologies have given rise to multiple distinct modeling practices, which in turn have given rise to inconsistent representations. [8-9]

Our identification of these problems – which are now acknowledged also by other experts in the field [10-11] – does not, however, imply that we dismiss traditional terminology resources as being without value. On the contrary, it is clear that the majority of these systems will continue to play an important role in the information-driven clinical and translational science of the future, and this for at least two reasons.

First, huge quantities of clinical and research data have already been annotated (and in some cases compiled *ab initio*) in their terms, and it cannot be expected that these data will be annotated a second time using OBO Foundry ontologies created *de novo*.

Second, where Foundry ontologies seek to represent the entities on the side of reality, traditional terminology systems are designed to reflect the ways language is used by clinicians and others in reporting (for example) patient encounters. This closeness to the needs of clinicians and healthcare institutions suggests that concept-based systems may still be in common use in the future.

The problem must be addressed, however, that the data resulting from such annotation efforts, precisely because they stay so close to the language used in specific disciplinary communities, and because they are affected by the multiple modeling paradigms associated with the orientation around 'concepts', are marked by the detrimental effects of silo formation.

The widespread adoption of SNOMED CT would diminish such effects. But as long as SNOMED CT itself does not use a consistent ontological approach [12], we believe that the data expressed with its aid, too, will involve too high a degree of redundancy and of inconsistent coding [13].

SNOMED's structure does not as yet provide a consistently accessible and reliable representation of the reality on the side of the patient as this reality changes through time. Moreover, SNOMED in its current form will not be able to do justice in consistent fashion to the changes in our knowledge of this reality which will be brought by advances in translational science [14]. To address these problems we need a strategy to map legacy terminologies such as SNOMED CT to OBO Foundry ontologies in such a way as to ensure that both can contribute to the creation of the non-redundant common framework for data integration and exploitation that will be needed in the future.

Objectives

The underlying idea is that both terminology artifacts and ontologies contain representational units (such as single words) and combinations thereof (such as compound word phrases and whole sentences) – together called 'representations' in what follows. The goal is to subject such representations to careful inspection of a sort which can allow terminological representations organized around 'concepts' to be mapped to appropriate ontological counterparts. To this end, we must provide a framework for ontological analysis of terms in legacy terminologies that will support adequate mappings especially for those terms that, because they are declared as 'synonyms', are associated with single 'concepts' under the terminological view. Such terms must be mapped separately wherever they refer – on face value – to entities of different types.

Methods

Our framework rests on three principal distinctions: (1) between *generic* and *specific portions of reality (**POR**s)*, (2) between the various purposes that can be served by *definitions*, and (3) between three distinct levels of reality.

Generic versus specific portions of reality

The first distinction separates *generals* from *particulars*, or in other words it separates *generic* (***GPR***) from *specific portions of reality* (***SPR***). While this distinction, like the remaining proposals outlined in this section, can be applied to both *continuants* (such as cells and organisms) and *occurrents* (such as lives and deaths), we shall concentrate here exclusively on the case of continuants.

Amongst the generic portions of reality are *universals* (***UNV***) and what we shall call *generic configurations* (***GCO***).

Universals are denoted by general terms such as 'human being', 'president', 'nation', 'population'. Universals are *instantiated* by particulars such as President Obama, the USA, the inhabitants of Buffalo. [15]

Generic configurations are configurations formed by *generic portions of reality* (***GPR***) that stand in some relation to each other that can be represented by some statement. An example is the portion of reality represented by the statement '*cell membrane part_of cell*'. Here '*part_of*' represents the generic *part_of* relation as described in the Relation Ontology. [7] Another example is the portion of reality represented by the sentence '*clinicians are human beings*'. Here the word '*are*' denotes what we shall call the *subgroup* relation, which holds between *clinicians* and *human beings*.

Amongst the *specific portions of reality* (***SPR***) are, analogously, *particulars* (***PAR***) and *specific configurations* (***SCO***).

PARs are entities that exist only once and are confined in space and time. Examples are: Mary, Buffalo, and the World Health Organization (WHO). Some ***PAR***s are what linguists would describe as 'named entities', but the majority – a liver cell in Mary, the fracture in her leg, and so forth – are not.

Both specific and generic configurations are represented by statements. Each ***SCO*** involves at least one ***PAR*** that stands in some relation to something else, for example to another ***PAR***, as in the specific configuration represented by the statement '*Mary's left leg part_of Mary*'. If Mary's left leg is amputated, then the two ***PAR***s involved in this ***SCO*** may survive the amputation, but the ***SCO*** itself will cease to exist.

Particulars can be divided into *atomic particulars* (***APA***) and *groups* (***GRP***). An atomic particular is a ***PAR*** that constitutes a unity in the sense that it has a complete, spatially connected external boundary. Examples, again, are: Mary and Mary's left leg. '*Atomic*' is here not to be understood as implying that the entity in question is not further decomposable. If Mary's left leg is amputated, then it may still exist, though not any more as part of Mary. Nor is it to be understood that anatomic particulars cannot themselves contain parts which are atomic (for example Mary herself contains parts which are her cells).

GRPs are entities denoted by generic terms such as 'limb of vertebrate', 'limb of human being', and even 'limb of Mary'. Although the latter example will likely not be found in a terminology or ontology, terms of the same sort do occur, examples being 'citizen of the United States', 'Nobel Prize winner', 'veteran of the Second World War'. Terms denoting ***GRP***s are typically formed via combination of smaller terms which themselves denote universals, particulars, or other ***GRP***s.

If Mary is a healthy human being, the entity denoted by the noun phrase 'Mary's limbs' is an example of a group (***GRP***). Each of healthy Mary's limbs is at the same time a *part* of Mary and a *member* of the corresponding ***GRP***. All members of a ***GRP*** at any given time are such as to exist at that time.

Among ***GRP***s, we distinguish further between, *bona fide groups* (***BGR***), *fiat groups* (***FGR***) and extensions (***EXT***) [16]. While these distinctions are by no means trivial, their correct understanding is important if we are to find coherent ways to manage the large families of terms (for example in SNOMED CT the family consisting of terms such as 'absent leg', 'amputated leg', 'withered limb', 'absent bone in leg', 'limb amputee', 'amputation of lower limb', 'amputation of limb'), whose meanings are otherwise difficult to capture in a coherent way.

A bona fide group (***BGR***) is a group whose members are homogeneous, are causally linked together, and which is maximal in the sense that all causally linked entities of the relevant sort are members of the group. Examples are: Mary's limbs, Mary's cells, Mary's molecules.

A fiat group (***FGR***) is a group that is demarcated by fiat, such as: left lungs of people currently in Buffalo, the left lungs of all the people now participating in clinical trial #77639.

At any time at which the ***BGR*** constituted by healthy Mary's 4 limbs exists, a cognitive being may explicitly recognize the simultaneous existence of any combination of two or more of her limbs. Some of these combinations, for instance any group of 3 of her limbs, are distinct ***FGR***s, since they fall short of being maximal. The groups formed by her two arms and by her two legs, in contrast, are ***BGR***s. The relation between fiat subgroups of the bona fide group that is formed by Mary's limbs is analogous to the relation between some proper part of Mary that is demarcated by fiat and Mary as a whole. There is a fiat boundary between healthy Mary's left arm and the rest of Mary's body in the region of her left shoulder.

To each continuant universal corresponds a group, called its *extension* (***EXT***), formed by all and only those particulars that are instances of that universal at any given time.

The purposes of definitions

Our second distinction recognizes three purposes which a *definition* of a representational unit may serve:

P1: to specify the conditions that must be satisfied for a term to be an acceptable designator for a given entity in some given community. An example would be:

chronic pain =def. *a pain that has been present for more than 3 months*

P2: to specify what is characteristic of particulars that instantiate a certain universal, for instance:

disorder =def. *a part of an organism which serves as the bearer of a disposition to pathological processes* [17]

P3: to demarcate groups and classes by specifying characteristics that their members or elements must exhibit..

P1 definitions are essentially a matter of terminological decisions. The definition given as example excludes the use of the term 'chronic pain' for pains lasting less than 3 months. This does not mean, however, that a pain in a specific patient that has already lasted for 90 days *becomes* a chronic pain one day later. It was, in fact, a chronic pain already from the very beginning, even though this fact was unknown to any observer.

P2 and P3 definitions help in determining whether a given particular is to be classified in a given way. P2 does this at the level of universals, while P3 does it for ***GRP***s and as further explained, classes.

First-order entities versus representations

The third distinction concerns the *level of reality* at which the referent of some representation exists. Of importance here is the distinction between

1. *first-order entities* such as patients, disorders, families,
2. *beliefs* in people's minds (including beliefs putatively about objects such as unicorns which do not in fact exist), and
3. *representations* in some publicly accessible medium, for instance a term in an ontology.

Applying the framework

When a terminology has been selected as one that needs to be mapped to OBO Foundry ontologies, each of its representational units should be inspected to identify, in terms of corresponding representations in Foundry ontologies, what sorts of ***POR***s it is able to denote. A problem is that terms from concept-based terminologies often denote multiple distinct sorts of ***POR***s, for example because of asserted subtype relationships, as in SNOMED CT, whose concept 'Finger structure' subsumes the concepts 'entire finger' (a ***UNV*** under a realist framework) and 'all fingers' (a ***GRP***) (though SNOMED does not specify whether the latter means: 'all fingers in the world', 'all fingers of a given patient', 'all fingers on a given hand').

To address this problem, we introduce an intermediary layer made up of *classes* (***CLA***), understood as arbitrary totalities of elements which are either (i) defined through some descriptor referring to ***POR***s of any of the sorts described thus far (for example: 'the disorders in all the patients treated by Dr. McX'), or (ii) totalities whose elements are themselves so defined, or (iii) combinations of (i) and (ii).

Classes under (i) thus carve out ***POR***s in ways which go far beyond ***GRP***s as defined in the foregoing. Classes under (ii) and (iii) allow simultaneous reference to entities associated together in ways which have no counterpart ***POR***, for example when we wish to assert heritability relations between Mary and certain of her ancestors who died many years before she was born.

Defined classes

Where groups have *members*, classes have *elements.* A *Defined Class* (***DCL***) is a class all of whose elements are specified by some class description. In the simplest case, this will be of the form '*ξ which stands in R to λ*', where '*ξ*' names some universal, for example 'person born in Belgium', which defines what we shall call a *Specifically Defined Class* (***SDC***), or 'patient who has tuberculosis', which defines a *Generically Defined Class* (***GDC***), each of whose elements enjoys the same relation (*exemplifies*) with instances of the single universal: tuberculosis. In more complex cases the definition will be of a logically more complex form, such as '*ξ* **has duration** *which stands in R to λ*', for example in the ***GDC*** *chronic pain*, where *ξ* is the universal: pain, *R* is the relation *longer_than* and *λ* is the temporal interval: 90 days. Many of the terminological definitions distinguished under P1 above will define terms which refer to ***GDC***s in the outlined sense.

For each ***GDC*** and for each ***SDC*** there is some universal from whose extension all its elements are drawn. An *Ad Hoc Class* (***AHC***), in contrast, is a ***CLA*** formed through combinations of ***GDC***s and ***SDC***s which is such that there is no such overarching universal. An example is, again, the SNOMED CT concept '*finger structure*', since among the entities that can be denoted by this term are both ***GRP***s and ***APA***s

Among ***AHC***s, too, we can distinguish both *Generic* (***GAC***) and *Specific Ad Hoc Classes* (***SAC***). An example of a ***SAC*** is the class whose elements are the clinical signs exhibited by some specific patient with tuberculosis [17]. An equivalent ***GAC*** would be the class whose elements are the clinical signs exhibited by all tuberculosis patients assigned to the control group of a given clinical trial.

Solving the semantic proximity problem

In its January 2009 version SNOMED CT associates the concept '*Fractured nasal bones (disorder)*' with the following synonyms: 'Fractured nasal bones' (S1), 'Broken nose' (S2), 'Fractured nose' (S3), 'Fracture of nose' (S4), 'Fracture of nasal complex' (S5), and 'Fracture of nasal bones' (S6). One consequence of the multiple interpretations that are given to the term 'concept' both inside [12] and outside [8] of SNOMED CT is that it is difficult to understand precisely how this 'association' is to be understood. In practice, what it means is that SNOMED is here acknowledging the different ways language users capture nasal bone fracture-related information when entering patient data into a record, and providing an aid to translating the corresponding bodies of data into SNOMED form. As realist ontology (and common sense) would suggest, however, it can be assumed that when a study nurse enters the term 'fractured nasal bones' into a patient record, then what he means thereby is not a *nose of a certain (fractured) sort* but rather a certain *group of bones.* If, accordingly, we are to devise a strategy for translating the resultant SNOMED data into the OBO Foundry framework, then our mapping will need to take account of the mentioned 'associations' in a more careful way than is possible when all the mentioned synonyms are treated *en bloc*. It is for this reason that we introduce the machinery of ***CLA***s and ***GRP***s in the above. This machinery is designed to make apparent the unarticulated complexity of SNOMED's synonymy relation by allowing each synonym to be treated separately in a way which at the same time allows formulation of the needed mappings to the corresponding OBO Foundry terms.

Human bones and noses are represented in the FMA Anatomy Ontology [18] by means of representational units denoting the universals *bone* and *nose* respectively. Fractures, in contrast, would be included in an ontology of disorders [17]. To realize our proposed strategy, now, scholars developing a mapping from SNOMED CT to OBO Foundry ontologies would have to decide, in collaboration with the SNOMED authors, what precisely the synonymous terms (S1–6) mentioned in our list above should properly be understood as denoting. In the framework here proposed, for example, S2 and S3 would both denote a ***GDC*** that is a subgroup of the extension of the universal *nose*. S1 would denote, according to further context, either a ***GRP*** which has *nasal bones* as members or a ***GDC*** denoted by the plural term 'bones of the nose'.

Another advantage of our strategy is that it helps us to understand the structure of the *is a* hierarchy in SNOMED CT. 44 concepts in SNOMED CT are described as being *is a* parents of *Fractured nasal bones (disorder)*. Where all of the synonyms referred to above denote first-order entities on the side of the patient, this is not the case for all 44 of the parent concepts listed. '*Disorder by body site (disorder)*', for example, reveals itself upon inspection to denote not a disorder at all but rather the way the representational units about disorders are further organized.

Another problematic case is '*Finding by site (finding*)': fractured nasal bones cannot, in our terms, be a (type) of finding, since something can only be found – and hence give rise to a finding – if it pre-exists, and is thus independent of, the corresponding act of observing. On our strategy, in fact, finding data would be mapped, not to bones directly, but rather to the corresponding datable observations.

Conclusion

It has been stated that '*Terminologies should not be developed by reference to a system of preferred terms, rather they should be developed in such a way that their individual nodes and* [the] *relations amongst these nodes are modeled on an underlying formal ontology, where the linguistic content of these nodes will be filled in based on a system of terms and synonyms (from many different languages) that is associated with each node based on the intended ontological interpretation of that node'*. [19] Few, if any, existing biomedical terminologies exhibit these characteristics. The framework we propose is designed to promote progress in this respect, with the goal, not of developing an underlying formal ontology for these terminologies themselves, but rather of achieving appropriate mappings to OBO Foundry ontologies. The approach provides a tool for terminologists to detect ambiguities and conflations in the conceptual structures they have designed and to determine the correct handling of terms proposed as synonyms; it also forces developers of realism-based ontologies to be more precise about what exactly the representational units in their artifacts denote. Certainly there is a long way to go. We acknowledge that the proposed approach is not easy to apply because of the subtle distinctions it requires, distinctions which are perhaps not easy to understand especially for adepts of the concept-based approach. We believe, however, that the approach promises significant benefits, both practical and theoretical, in the long run.

Acknowledgements

The work described was funded in part by the John R. Oishei Foundation and also by grant R21LM009824 from the National Library of Medicine and by the NIH Roadmap grant 1 U 54 HG004028. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

References

[1] Bodenreider O. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. Yearbook of Medical Informatics: access to health information. Stuttgart: J. Schattauer; 2008.

[2] Smith B, Ashburner M, Ceusters W, Goldberg L, Mungall C, Shah N, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;25:1251-5.

[3] Schorlemmer M, Kalfoglou Y. Institutionalising Ontology-Based Semantic Integration. Journal of Applied Ontology. 2008;3(3):131-50.

[4] Brinkley JF, Suciu D, Detwiler LT, Gennari JH, Rosse C. A framework for using reference ontologies as a foundation for the semantic web. Proceedings of the AMIA Fall Symposium 2006. p. 95-100.

[5] Smith B. Ontology (Science). In: Eschenbach C, Grüninger M, eds. Formal Ontology in Information Systems - Proceedings of the Fifth International Conference (FOIS 2008). Amsterdam: IOS Press; 2008. p. 21-35.

[6] IFOMIS. Basic Formal Ontology. 2009; Available from: http://www.ifomis.uni-saarland.de/bfo/.

[7] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biology. 2005;6(5):R46.

[8] Smith B, Ceusters W, Temmerman R. Wüsteria. In: Engelbrecht R, Geissbuhler A, Lovis C, Mihalas G, editors. Connecting Medical Informatics and Bio-Informatics Medical Informatics Europe 2005. Amsterdam: IOS Press; 2005. p. 647-52.

[9] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: Where do they come from and how can they be detected? In: Pisanelli DM, editor. Ontologies in Medicine Studies in Health Technology and Informatics. Amsterdam, The Netherlands: IOS Press; 2004. p. 145-64.

[10] Cimino JJ. In Defense of the desiderata. Journal of Biomedical Informatics. 2006;39(3):299-306.

[11] Solbrig HR, Chute CG. Concepts, Modeling and Confusion. In: Smith B, editor. ICBO: International Conference on Biomedical Ontology. Buffalo NY: National Center for Ontological Research; 2009. p. 123-6.

[12] Schulz S, Cornet R. SNOMED CT's Ontological Commitment. In: Smith B, editor. ICBO: International Conference on Biomedical Ontology. Buffalo NY: National Center for Ontological Research; 2009. p. 55-8.

[13] Hogan WR. What's in a 'is a' Link? In: Smith B, editor. ICBO: International Conference On Biomedical Ontology. Buffalo NY: National Center for Ontological Research; 2009. p. 174.

[14] Ceusters W, Spackman KA, Smith B, editors. Would SNOMED CT benefit from Realism-Based Ontology Evolution? American Medical Informatics Association 2007 Annual Symposium Proceedings, Biomedical and Health Informatics: From Foundations to Applications to Policy; 2007 November 10-14; Chicago IL: American Medical Informatics Association.

[15] Smith B, Kusnierczyk W, Schober D, Ceusters W. Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. KR-MED 2006, Biomedical Ontology in Action. Baltimore MD, USA 2006.

[16] Smith B. Fiat Objects. Topoi. 2001;20(2):131-48.

[17] Scheuermann RH, Ceusters W, Smith B. Toward an Ontological Treatment of Disease and Diagnosis. Proceedings of the 2009 AMIA Summit on Translational Bioinformatics, San Francisco, California, March 15-17, 2009: American Medical Informatics Association; 2009. p. 116-20.

[18] Rosse C, Jr MJ. The Foundational Model of Anatomy Ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy Ontologies for Bioinformatics: Principles and Practice. London: Springer; 2007. p. 59-117.

[19] Baud R, Ceusters W, Ruch P, Rassinoux A-M, Lovis C, Geissbühler A. Reconciliation of Ontology and Terminology to cope with Linguistics. In: Kuhn K, Warren J, Leong T, editors. Proceedings of MEDINFO 2007, Brisbane, Australia, August 2007. Amsterdam: Ios Press; 2007. p. 796-801.

Address for correspondence

Werner Ceusters
Center for Excellence in Bioinformatics & Life Sciences
University at Buffalo
701 Ellicott street
Buffalo NY, 14203, USA