Six Questions on the Construction of Ontologies in Biomedicine

Anand Kumar[a], Anita Burgun[b], Werner Ceusters[c], James J. Cimino[d], James Davis[e], Peter Elkin[f], Ira Kalet[g], Alan Rector[h], Jim Rice[i], Jeremy Rogers[h], Stephan Schulz[j], Kent Spackman[l], Davide Zaccagini[m], Pierre Zweigenbaum[n], Barry Smith[a,k]

[a]*IFOMIS, University of Saarland, Saarbruecken, Germany*
[b]*LIM, University of Rennes, France*
[c]*ECOR, University of Saarland, Saarbruecken, Germany*
[d]*Department of Biomedical Informatics, Columbia Univ. College of Physicians & Surgeons, NY, USA*
[e]*Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, USA*
[f]*Mayo Clinic College of Medicine, Rochester, MN, USA*
[g]*Radiation Oncology Department, University of Washington, WA, USA*
[h]*Department of Computer Science at University of Manchester, UK*
[i]*Muskegon Surgical Associates, MI, USA*
[j]*Department of Medical Informatics, Freiburg University Hospital, Germany*
[k]*Department of Philosophy, SUNY at Buffalo, NY, USA*
[l]*Oregon Health & Science University at Portland, OR, USA*
[m]*Department of Health S&T, Harvard-MIT, MA, USA*
[n] *Assistance Publique - Paris Hospitals, STIM; INSERM, U729; INALCO, TIM, Paris, France*

Abstract

Best practices in ontology building for biomedicine have been frequently discussed in recent years. However there is a range of seemingly disparate views represented by experts in the field. These views not only reflect the different uses to which ontologies are put, but also the experiences and disciplinary background of these experts themselves. We asked six questions related to biomedical ontologies to what we believe is a representative sample of ontologists in the biomedical field and came to a number conclusions which we believe can help provide an insight into the practical problems which ontology builders face today.

Keywords: Biomedical ontology, Formal ontology, Knowledge representation

Introduction

This communication has been prepared on the basis of an email and face-to-face discussion among AMIA's Knowledge Representation Special Interest Group (KR-SIG) members regarding the best practices for ontology building in biomedicine. Much has been written on this topic and conceptions of what are best practices for ontologies seem to vary between different groups, depending on field of application, experience and expertise. [1,2,3] The objective of this paper is to provide some answers to practical problems which builders of ontologies, terminologies and vocabularies typically face in the biomedical domain. Our goal is to provide a clear answer to the question whether different fundamental approaches to ontology-building lead to substantial differences in ontologies or rather affect only the representation of a small number of especially problematic classes. In polling the views of experts in biomedical ontologies we found that their views are not so far apart as they sometimes seem to be. Among the authors, there is a consensus that the term 'ontology' should be used in a broad sense which does not exclude any of those artifacts commonly referred to as 'biomedical vocabularies' in the medical informatics community.

We addressed 6 questions to the KR-SIG members:
P1. Are ontologies about concepts or entities in reality?
P2. & P3. Should we impose the rule of single inheritance in ontologies? Should we create ontologies embodying multiple partitions?
P4. Should we support the negative classes such as: *invertebrate*, *anosmia*, or even *unlocalized* of the type explicity represented within some ontologies?
P5. Is it always possible to construct ontologies in such a way as to involve only pairwise disjoint siblings?
P6. Is it always possible to construct ontologies in such a way that they will represent jointly exhaustive siblings?

P1. Concepts vs. Reality

While many definitions for terms like 'concept', 'class', 'type', 'universal', etc. exist, we will consider the following definitions here, noting that 'entity' (representing the top, all-inclusive category) is for the purposes of this discussion treated as a primitive notion .

Class: In the set-theoretic sense, a class is the set of objects instantiating a given universal (or: instantiating a given universal at a given time); i.e. it is the extension of a universal. In the literature on ontologies, however, 'class' is sometimes used as a synonym for 'universal'.
Universal: a general entity an invariant in reality, which can be instantiated in a range of particular instances; thus the sort of entity to which a general term like 'cell' or 'gastrulation' (e.g. appearing in a scientific textbook) corresponds. 'Type' and 'kind' are synonyms of 'universal'.
Instance: a particular entity which instantiates a universal; thus it is the sort of entity to which a proper name or an indexical expression ('this bone here') corresponds. 'Token' is a synonym of 'instance'.

The responses to our questions can be divided in light of two cross-cutting oppositions: between formal and informal representations, and between congnitivist/conceptualist and realist representations. The arguments pertaining to the former concern the added value which is to be gained from providing rigorous logical structure (defended e.g. by GALEN[1]) vs. the more intuitive structures targeted to human users preferred for example by the representatives of the Gene Ontology[2] (GO). and by the proponents of the kind of approach typified by the Read codes[3].
The most prominent conceptualist ontologies include the Unified Medical Language System[4] (UMLS), the National Cancer Institute Thesaurus[5] (NCIT), and GALEN; realist ontologies are, for instance, the Foundational Model of Anatomy[6] (FMA), SNOMED CT[7] and (increasingly) GO. As GALEN makes clear, not all conceptualist approaches are informal; as GO makes clear, not all realist ontologies are formal.
Conceptualists start with *concept* (or some equivalent) as their root class, and realists with *entity*.[8] Conceptualists see no difference between the representation of existing and non-existing entities, thus *unicorn* and *uremia* are treated in the same way. Realists, on the other hand, exclude classes like unicorns from their representations. For obvious reasons the opposition in question has led to arguments pertaining to the existence criteria for classes/concepts.

[1] http://www.cs.man.ac.uk/mig/galen/index.html
[2] www.geneontology.org
[3] http://www.equip.ac.uk/readCodes/docs/index.html
[4] http://www.nlm.nih.gov/research/umls/
[5] http://nciterms.nci.nih.gov/NCIBrowser/Startup.do
[6] http:// sig.biostr.washington.edu/projects/fm/AboutFM.html
[7] http://www.snomed.org/
[8] HL7 Reference Information Model (RIM) has act and entity as their root class. They do not consider act to a subclass of entity; rather they represent act and entity as mutually exclusive classes.

Another apparent difference between the two methodologies is characterized by the example,

thumb part_of hand

In a realist ontology *hand* and *thumb* are classes of individual things. Only individuals can have parts. Therefore this relation must be understood e.g. as a disguised assertion about individuals, to the effect that every instance of the universal *hand* has some instance of the universal *thumb* as instance-level (mereological) part. Such *all-some* assertions are on the other hand problematic, since hands without thumbs do after all exist.
A conceptualist, in contrast, would look not at individual hands, but rather at the mental representation associated with the natural language expression 'hand'. In this mental representation a hand has a thumb, and the above relation is thereby justified.
Alternative ways of solving the apparent invalidity within a realist ontology include:
a. introduce more specific classes, e.g. *hand without thumb*, whose instances do not have a thumb as its part.

hand without thumb is_a hand

$\forall xy(\mathbf{inst}(x, \textit{hand without thumb}) \ \& \ \mathbf{inst}(y, \textit{thumb}) \rightarrow \text{not}(\mathbf{part\_of}\ (y, x)))$

hand with thumb is_a hand

$\forall x(\mathbf{inst}(x, \textit{hand with thumb}) \rightarrow \exists y(\mathbf{inst}(y, \textit{thumb}) \ \& \ (\mathbf{part\_of}\ (y, x)))$

where **inst** is the relation of instantiation between an instance and a universal, **part_of** is the parthood relation on the level of instances, and '$\forall x$' signifies: for all values of x)
b. Define the parthood and instantiation relationships in such a way as to take account of time [1]:

$A\ \textit{part\_of}\ B =_{\text{def}} \forall xt(\text{inst}\ (x, A, t) \rightarrow \exists y(\text{inst}\ (y, B, t) \ \& \ \mathbf{part\_of}(x, y, t)$

where '$\exists x$' signifies: for some value of x. Thumbs are always parts of hands at the initial moment of their existence.
c. Another way to deal with this issue is to provide cardinalities to the relations. For example, to represent *hand* could have either one or no *thumb* as part:

hand has_part(0,1) *thumb*

On the other hand, a thumb is always a part of a hand.

thumb part_of(1,1) *hand*

Beyond this point, formally correct ontologies, whether conceptual or real are similar.

If Health Level 7[9] (HL7)'s Reference Information Model (RIM), has

document is_a act

diagnosticImage is_a act

[9] http://www.hl7.org/

then, it does not matter if 'act' here designates a concept or an entity: it is still incorrect. Rather, we should have on the left-hand side of these two assertions something like: *act of taking documentation* and *image-based diagnostic act* respectively.
When the UMLS Semantic Network asserts:

bacteria causes experimental model of disease

then the problem here has to do not with the realism vs. conceptualism debate, in the sense that representatives from both sides would find the given assertion puzzling.
When the National Cancer Institute Thesaurus asserts:

tissue is_a other anatomical concept

then the problem is not that tissue has been classified as a concept, but that the term 'other anatomical concept' has no coherent meaning from the ontological point of view. Thus even if we find it correct to classify tissue as a concept, it should properly be classified is as a child of say *anatomical entity concept*, which would then in turn be classified as a child of *concept*.

There are situations where we need to represent the relation of prevention, even though entities in the corresponding prevented classes then do not, by definition, exist. So, for example, a contraceptive drug prevents conception and SNOMED CT correspondingly has a class *prevented pregnancy*. We need to distinguish such classes from cases such as *unicorn* or *goblin.*
Of course, the following relation is incorrect

prevented pregnancy is_a pregnancy

But it is true that,

pregnancy prevention event part_of contraception event

Builders of vocabularies and ontologies have often failed to take such differences between ontologies into consideration, and they have thus tended to create representations portions of which are realistic and others conceptualistic. This practice should be avoided.

P2. Single vs. Multiple Inheritance & P3. Partitions in ontologies

Binomial ontologies has been used as a justification for making all taxonomies into trees, however some argue that single inheritance is not a sufficient answer in many classifications. [4,5,6,7] Another opinion is that the structure of ontology should be chosen to accurately represent the real world and that the decision to use single or multiple inheritance is domain-specific. Yet another opinion favors the use of single inheritance only.

An ontology is **partitioned** when only one criterion is used for its classification. There are many partitions which are present within multiple inheritance and within each such partition, there are further subclassifications. For example, within the pathology partition, there are subclassifications of the sort:

by infection: (e.g., Granulomatous pancreatitis)
by type of malignancy: (e.g., Adenocarcinoma pancreas)
by metabolism: (e.g., Diabetes Mellitus)

These disparate views can be reconciled on the basis of following practices:
a. A clear distinction should be made between primitive and defined classes.
b. At any level the primitive siblings should be disjoint and any primitive should have only one primitive parent.
c. All multiple inheritance should be achieved using formal classification based on explicit definitions.

The rationale behind this proposal is:
i. **Representational clarity**. Ontology split into partitions brings clarity of structure. [2]
ii. **Modular interfaces**. The cross-module definitions and restrictions (semantic links, conditions, and constraints) provide the interface between modules.
iii. **Explicitness**. When some universal is a child of two parents, the reason stating why the dual subsumption is present provides explicitness and helps in the drawing of inferences and in avoiding coding errors.
iv. **Ease of maintenance**. Modular ontologies are easier to maintain and error-check e.g. with Description Logic-based tools
v. **User-friendliness**. The users of ontologies find it easier to navigate a modular ontology as they can choose which tree they want to traverse.
vi. **Usage of user-defined classes**. A terminology restricted only to those classes users want to see is often too complicated for them to find what they want to see, and then one is left with the problem of how to filter and otherwise process the content in order to deliver what the user needs. In that case, having more detail than any user actually needs can provide the key leverage needed to give specific users what they actually want and such detailed representations are easier in a modular ontology. [2]

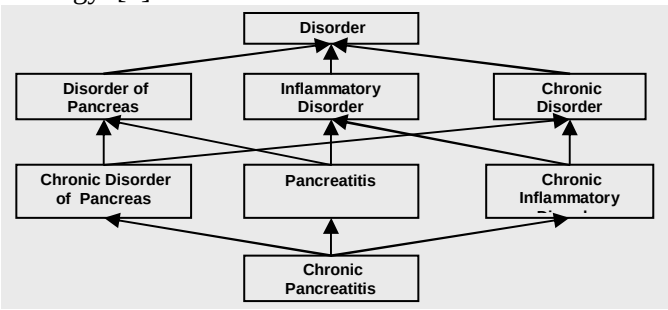

Fig. 1. Example of multiple inheritance

The *is_a* relations represented by the arrows in figure 1 mean, for example:

every instance of *chronic pancreatitis* is thereby also an instance of: *chronic disorder of pancreas*, *pancreatitis*, and *chronic inflammatory disorder*

Where such relations are asserted in an ontology, a computer can derive that *chronic pancreatitis* has three parents. But the computer needs also to derive the reasons why the corresponding *is_a* relations obtain. There are at least two ways to achieve this information.

a. Via the relational assertions we can make concerning the class *chronic pancreatitis*, namely:

has_pathology inflammation

has_onset chronic

has_location pancreas

For each of the three corresponding parent classes one of these relations is not assertible. Thus, for *chronic inflammatory disorder* not *has_location pancreas*; for *pancreatitis* not *has_onset chronic*, and for *chronic disorder of pancreas* not *has_pathology inflammation*.

Then, however, one can assert the non-*is_a* relations only, and one would not explicitly need to represent the multiple inheritance. This is what the OIL plug-in in Protégé allows.[10]

b. The non-*is_a* relations are represented as in case a. However, the *is_a* relations are further specified in terms of the reason for subsumption. Thus we could assert of *chronic pancreatitis* that it:

is_pathology_specification_of chronic disorder of pancreas

is_onset_specification_of pancreatitis

is_location_specification_of chronic inflammatory disorder

In order to understand the reason for a given subsumption the computer then does not need to look into the various relations of the class.

Thus, every multiple inheritance can be decomposed via the sort of normalization described in [2]. Such partitions are used in processors like GRAIL for GALEN or SMK to drive PEN&PAD. According to one's perspective they can be seen either as promoting multiple inheritance or as helping to eliminate it. But clear, these alternatives reflect no genuine underlying difference.

P4. Use of negative classes

Should ontologies support negative classes, for example, *invertebrate*, *anosmia*, *unlocalized*? All three are currently used in existing ontologies. Some apparently negative classes are perfectly acceptable to the realist, as for example those which figure in oppositions like: *insulin-dependent* vs. *non-insulin dependent diabetes mellitus*, or *gram-positive* vs. *gram-negative bacteria*, because they are not genuinely negative classes, and the same view can be held of *anosmia*. Similarly, the distinction between *VHL+ Kidney Carcinoma* and *VHL- Kidney Carcinoma* is real (referring to the *Von Hippel Lindau (VHL) mutation*, an important parameter in renal cancer. [8]

There are other situations where seemingly negative classes point in fact to entities acceptable to the realist. For example, *blindness* is defined as: lacking or deficient in sight; especially: having less than 1/10 of normal vision in the more efficient eye when refractive defects are fully corrected by lenses. This definition tells us that *blindness* signifies not something negative but rather the impairment of a function.

P5. Pairwise disjunction

A classification C_1, ... , C_n is pairwise disjoint if and only if no pair of distinct classes share a common instance.

Pairwise disjointness can be achieved within ontologies, but only when the ontology is modularized/partitioned in such a way as to eliminate multiple inheritance along the lines suggested under P2-P3 above. In a multiple inheritance hierarchy this will not be possible, but it can be achieved within each of the partitions . For example, *acute hepatitis*, *subacute hepatitis*, *chronic hepatitis* and *viral hepatitis* are not pairwise disjoint. However, within a partition in which hepatitis is classified on the basis of time of onset, *acute*, *subacute* and *chronic hepatitis* will be pairwise disjoint.

Most common classes of endurants (continuants) are unproblematically pairwise disjoint. No cell is an organ, No head is a foot, and oxygen is not hydrogen. When representing anatomy, mereological disjointness is also commonly encountered. .

Another approach to this representation is to distinguish between primitive and defined classes and allow negative classes only within the defined classes built over the primitives. The problem is that then one would need to draw a distinction between which classes are primitive and which are defined; a boundary hard to agree upon.

P6. Jointly exhaustive classes

A classification C_1, … , C_n is jointly exhaustive if and only if every individual in the relevant domain instantiates some class in the classification.

One should not try to achieve joint exhaustiveness except in cases where the subdivision of a class is done within a single partition. But one should not try to force

[10] http://www.ontoknowledge.org/oil/

this artificially. If one partitions all the infectious pneumonias according to their etiologies, we would have to either create siblings for every possible etiology (including ones that have never been observed nor are likely to be) or have some 'other' sibling, as in ICD-9/ICD-10[11] to provide a place to add new concepts when they are observed. Sometimes, negative classes can be of use to solve the problem. For example, for a *viral hepatitis* caused by a virus with is neither A nor B nor C, instead of using *other forms of viral hepatitis*, one could use *non-A, non-B, non-C viral hepatitis*. Although these classes emphasize absent features (i.e., the features present in the complement class), we argue that they essentially correspond to valid, genuine classes for which no specific positive name or names have as yet been crafted. [8]

As for negative classes, for both pairwise disjunction and jointly exhaustiveness, we need to make distinctions that these are present for primitive classes and defined classes might not always have these properties.

Conclusion

The main conclusions reached are:

- There is no difference when one uses conceptual or realist ontologies, as long as the concepts created directly correspond to the object in the reality.
- Creation of formal ontologies should be encouraged over informal ones.
- In building an ontology one should be consistent in using either exclusively realist principles or exclusively conceptualist principles.
- It is useful to begin with single inheritance trees while building ontologies on the basis of a single partition that is a single cause for subsumption within one tree. This helps in ontology creation and maintenance, and also in the drawing of inferences and in navigation and information retrieval.
- When *is_a* relations are specified in terms of the module in which it is present, only one relationship is needed to understand the reason behind a subsumption. On the other hand, we need to recompute the relations of the child and parent class when multiple inheritance is present, in order to understand those reasons. Such reasons for subsumption are useful for navigation across the tree and for drawing inferences.
- Ontologies should support classes which represent minimization or lack of certain functions or attributes. However negative classes such as *unlocalized*, *disease not otherwise specified* should be avoided.
- Jointly exhaustive and pairwise disjoint classes should be represented when the ontology where the subdivision of class is carried out within a single partition. In those situations where this is not achievable, ontologies should use the open world assumption that missing classes within the ontology will be filled up at a later date and thus avoid making artificial classes to cover those cases.

[11] http://www.cdc.gov/nchs/icd9.htm

References

[1] Smith B, Ceusters W, Klagges B, Koehler J, Kumar A, Lomax J, Mungall C, Neuhaus F, Rector A, Rosse C. Relations in biomedical ontologies. *Genome Biology*:; 2005 (in press)

[2] Rector A. Modularisation of domain ontologies implemented in Description Logics and related formalisms including OWL. *Knowledge Capture* 2003, (Sanibel Island, FL, 2003), ACM, 121-128.

[3] Bodenreider O, Smith B, Kumar A, Burgun A: Investigating subsumption in DL-based terminologies: A case study in SNOMED CT . *KR-MED 2004*: 12-20

[4] Russell S. and Norvig P. *Artificial Intelligence: A modern approach*. Chs. 10.6 and 10.7.

[5] Brachman R. and Levesque H. (Eds.) *Knowledge representation and reasoning* (2004), Ch. 9.

[6] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. *J Biomed Inform*. 2004 Apr;37(2):77-85.

[7] Smith B, Koehler J, Kumar A. On the application of formal principles to life science data: A case study in the Gene Ontology. DILS 2004;:79-94 .

[8] Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: A case study in medical terminology. *Proc. FOIS*, 4-6 November 2004.

Acknowledgement

We are thankful to the AMIA staff and KR-SIG chair for providing the infrastructure for this discussion. IFOMIS members are supported by the Wolfgang Paul Program of the Alexander von Humboldt Foundation, the European Union Network of Excellence entitled Semantic Interoperability and DataMining in Biomedicine (SemanticMining) and the Volkswagen Foundation project "Forms of Life."

Contact Information

Anand Kumar, IFOMIS, University of Saarland, Postfach 151150, D-66041 Saarbruecken, Germany. Email: akumar@ifomis.uni-saarland.de