Applied Ontology 10 (2015) 71–77
DOI 10.3233/AO-150144
IOS Press

Editorial

What's a "disease"? Questions for applied ontologies of diseases

Alan Rector *
School of Computer Science, University of Manchester, Manchester, UK

1. Introduction

This issue of *Applied Ontology* includes an article by Rovetto and Mizoguchi (2015) that sets out an alternative ontological formulation of the notion of "diseases" as "causal chains" or "processes", which they call the "River Flow Model" (RFM). They contrast their account with that in the Ontology of General Medical Science (OGMS) as described by Scheuermann et al. (2009) that formalises diseases as "dispositions" to undergo "pathological processes" as the result of "disorders" that are kinds of "physical components". There have also been a series of articles published recently by Schulz and his colleagues arguing that the entities in the large coding system for diseases, SNOMED CT (IHTSDO, 2015), are best interpreted as "situations" (Schulz et al., 2011) by which they define as "phases of the life of a patient, during which he/she is bearer of a clinical condition". This interpretation has been largely accepted by both the SNOMED organisation and the group harmonising the latest revision of the World Health Organisation's International Classification of Diseases (ICD) with SNOMED CT (Schulz et al., 2012).

In addition, entities entitled "disease" occur in at least 45 of the ontologies in BioPortal (NCBO, 2015), as well as in other standards such as the *de facto* standard for exchange of laboratory test data, LOINC (LOINC, 2015), the National Library of Medicine's Unified Medical Language System's Semantic Network (UMLS, 2003), and various more general resources such as WordNet (WordNet, 2015). Some of these are not "ontologies" in the sense usually understood in this Journal. The word "ontology" is now used for such a variety of information artefacts that, without qualification, it is almost meaningless. We take it here in a relatively narrow sense familiar to most readers of this Journal as a formal representation of the definitions and necessary characteristics of the entities in the world, or more broadly, of the entities that appear in some information system.

Given this situation, how should we discuss and evaluate these various ontological formulations of "diagnosis" for purposes of applications? We take it as given that the purpose of the ontological. Although Rovetto and Mizoguchi discuss their application briefly, as with most papers in this area the focus is primarily on the internal structure consistency of the ontology. So, how should a developer decide which

*Address for correspondence: Alan Rector, 15 Ashland Road, Sheffield, S7 1RH, UK. Tels.: +44 114 255 3660, +44 771 511 7126; E-mail: rector@cs.man.ac.uk.

disease ontology to use for a given application? This editorial seeks to articulate some questions that such a developer might wish to ask and to initiate a broader discussion of the use of disease ontologies in applications.

2. What's it for? Is it fit for purpose?

How do we know if an ontology of diseases is fit for purpose – i.e. for the role to be played by an "ontology" in the application? In our experience there are three families of such roles:

- "Terminology" and "codes" for information capture, information retrieval, and/or information sharing (interoperability) in which we include marking up or annotating other information with metadata;
- Information presentation and natural language processing;
- Inference and Knowledge Representation (e.g., for clinical decision support and for work in systems biology).

It is the first of these – "terminology" and "coding" – that is currently the most prominent in actual practice. Ontologies are used for terminologies in electronic health records, in annotating databases of molecular biological or generic information, or as part of broader knowledge-representation systems. (The phrase "coding" has a long history in Health informatics. It is used for most uses of controlled vocabularies, terminologies and ontologies to record information in electronic health records, lab and other clinical systems, and many billing systems. The origin of the phrase is that terminologies are almost always represented ultimately as "codes" or "coded expressions" using arbitrary identifiers to avoid confusion of spelling, language, and other linguistic artifacts.)

Most of these applications emphasise "interoperability" or "data sharing". Stripped of jargon, "interoperability" and "sharing" come down to communication – between machines, between machines and people, and between people via machines, synchronously or across time. An implicit assumption is that the communication will be by means of a symbolic – usually "coded" – representations rather than natural language. One key question of ontologies of disease is, therefore: Do they work as required to communicate information? Between different information systems? Between humans and the information system? Between humans using the information system as an intermediary?

Technical tests of communication between systems are relatively straightforward, for example as parts of various "connectathons" (e.g., see http://www.ihe.net/connectathon/). Tests of human–machine communication are rarer. Proxies such as inter-rater reliability in coding are sometimes performed (Andrews et al., 2007; Rothschild et al., 2005), but these do not necessarily capture whether the information entered is really what the consensus of coders intended. To test either human–system or human–system–human communication requires end-to-end testing and assessment of the resulting actions – something difficult to do and, to the best of the author's knowledge, not yet done at scale for any major ontology or coding system.

A second approach to testing the reliability of communication is to look at the accuracy of retrieval – when a question is asked does the system using the "ontology" find the expected answers? If not, can the problem be attributed to the "ontology"? Most queries will involve all entities in some branch – or in some Boolean combination of branches – of the hierarchy. So a simple test is to ask if all and only the entities that a clinician or scientist would expect to be find under a given abstraction actually do actually fall under it. Spot checks may reveal egregious errors (Rector et al., 2011), and systematic reporting

of anomalies found during use can reveal others. For a more comprehensive approach, a comparison between independently developed terminologies is required, as is occurring in the harmonization of the ICD and SNOMED CT. However, when discrepancies are found, how should it be determined which is correct? Ignoring simple blunders, this is likely to turn on the interpretation of how "diseases" are to be represented, and indeed a survey of such issues motivated the decision by ICD and SNOMED to opt for the "situation" interpretation for the entities in their respective systems (Schulz et al., 2012).

A third question is the complexity of the queries required for retrieval and whether they return the expected results. Does one formulation of disease lead to ontologies that make such queries simpler? More accurate? Lead to shorter expressions? How accurately do implementers of varying degrees of experience make formulate the queries? What errors are commonly made?

3. Which differences are substantive

Given two formulations, how do we know that they are different? The words used to label ontologies are frequently problematic – after all much of the issue is that the words are admitted to be ambiguous. For example might Rovetto and Mizoguchi's notion of "disease" in the RFM correspond more closely to Scheuermann notion of "disorder" in the OGMS? Or perhaps "disease course"? Even if the match is not perfect, might one of these be the best place to start? Rovetto and Mizoguchi make a start on related questions in the paper in this issue, but most issues are left as further research. Equally, do the entities in SNOMED CT correspond to entities in either the RFM or OGMS, or do they represent they something else entirely? For example, in SNOMED, syndromes and other conditions where the underlying process is ill understood are classified according to their manifestations – as in the example of "Tetralogy of Fallot". How would the River Flow Model cope with such cases where the causal chain is not fully known? Where classification has preceded understanding of the causal chain. Is there an information-preserving transformation from one formulation to another? Complete? Partial?

What is the expressivity of each formulation? Are there important notions that cannot be expressed? If so, are there work-arounds?

These questions are not necessarily easy to answer, but one can at least look at the hierarchies that result. Will, for example, a given notion such as "pulmonary stenosis" be classified under the same entities in both systems? If not, what are the discrepancies and why?

4. Formal questions

Most ontologies are represented using formal languages, most frequently some subset of first-order logic, typically a description logic (often a profile of OWL). Inference of subsumption is intended to be computable in most ontology languages, and some of these languages support other kinds of inference. Since first-order logic is computationally intractable, it is important to know the subset that is used and its computational properties, including any heuristic approaches to dealing with computationally intractable problems. The practical performance of inference in the formalism is particularly important in medical applications if they are to be performed at run time in clinical systems (e.g., for "post-coordination" of specialisations of existing entities; Green et al., 2006), as these are often time critical.

Given that most representations are limited to a subset of first order logic, how do they deal with notions that cannot be directly represented in that subset? At least two occur frequently. The first is the representation of cyclical processes and feedback, which falls outside of the DL framework, although

extensions to DLs have been proposed for analogous problems in other domains (Hastings et al., 2010). Closely related is the notion of "same" or, even more important, "same kind as" (e.g., to define "autocatalysis" or indeed reproduction – that organisms reproduce organisms of the same kind). The notion of "same kind" is intrinsically higher order, and difficult to represent in logic because our traditional notions of diseases do not label sorts of classes analogously to biology's "species", "genera", and so on. Lacking such established sorting, one has to capture the intended meaning of "same kind as" while avoiding notions such as "immediate parent", which is dependent on the granularity of the hierarchy on the one hand, and, on the other, the fact that everything is a kind of the top entity.

Any formal representation involves primitives that have to be explained outside the formalism. How clear are these explanations. In particular how clear are notions such as "realises" in OGMS. For example, what are the criteria for a particular case to "realise" diabetes, even if only some of the manifestations of diabetes are present. In the River Flow Model, *Causal Structure* (*Chain*) "inherits from" *Dependent Causal Structure*. How are the semantics of "inherits" different from "subsumed by"? What are their semantics? In both cases, are there operational criteria to determine whether or not the relations hold?

More generally, the issues of specialisation and abstraction are less obvious for diseases than might be imagined. If to be an instance of a disease class, an entity must share all of the necessary and sufficient conditions of that class; how does the system cope with the fact that most diseases are variable in their presentation? In the River Flow Model, when is one causal chain an instance of a class of chains? In OGMS, when does one disposition realise another? When is a disease an instance a disease class? How does this relate to "Disease course"? If the abstraction relation is not logical subsumption, how is it defined? What are the criteria?

Finally, it is difficult to account for medical or scientific discourse without dealing with hypotheticals – things that may or may not be true. Medical discourse traditionally proceeds by means of *differential diagnoses* – a list of hypotheses about a patient's diagnosis. Scientific discussion is founded on hypothesis testing. Ontologies that derive from Smith's Basic Formal Ontology (Smith, 1998), which only allow the representation of entities that are known to exist (e.g., OGMS and Schulz's BioTop; Beisswanger et al., 2008), must make special provisions for such cases. For these or any other ontology, how are hypotheticals dealt with? If they require formal "tricks" such as the unintuitive use of universal rather than existential quantification, do the tricks work? Do they have undesirable side effects?

Beyond hypotheticals, what role, if any, do notions of uncertainty, fuzziness and probability play? Few processes in biomedicine are invariant. Most involve probabilities that give rise to uncertainties. Many of our notions are fuzzy (e.g., when a patient becomes "elderly" or the point at which elevated blood pressure is accepted as hypertension). How does the formulation of disease account for these issues? Or can it be scoped to exclude them? If so, are the criteria for scoping well defined? Acceptable?

5. Human factors, language, pragmatics and standards

Scheuermann et al. (2009) rightly point out that the use of words in the OGMS may seem unintuitive to clinicians, biomedical scientists, and other potential users. The same is true, to a greater or lesser extent of almost all formal ontologies. How is this situation dealt with? Is this equally true of the River Flow Model? Are its notions more or less intuitive to clinicians and researchers? In whichever system, are there systems to aid users to make unfamiliar distinctions correctly? What are the implications in terms of user training and documentation? What are the risks of misinterpretation resulting in clinical or scientific errors?

This is a special aspect of the more general problem that clinical usage often conflicts with logical rigour. Grice (1957) pointed out many years ago that standard usage often depends on emphasizing exceptions and leaving the usual case to be understood implicitly. In medicine, where one kind of a class of diseases predominates, common usage is often to use the generic class name for the common case and to use a qualification for the unusual case. For example, unqualified, "subdural hematoma" (a blood clot under the covering of the brain and spinal cord) will be assumed to be inside the skull. The rare cases occurring in the spine will be normally qualified as "spinal subdural hematoma". The diagnosis is common and an error in classification life threatening. There is a strong argument for clarity for labelling their common parent something like: "subdural hematoma intracranial and/or spinal" rather than just "subdural hematoma" (Rector et al., 2011).[1] Although the latter is technically correct, the likelihood of miscommunication is high. Such potential errors need to be taken into account in any program of quality assurance or safety testing of the ontology as it is intended to be used.

More generally, human beings normally communicate in natural language. What support is given beyond a single name for language processing? In the primary language (usually English)? In other languages? For technical users? For non-technical users? What support for or links to linguistic resources are there, either for language interpretation or generation? For example, an essential but unusual feature of GALEN, an early medical ontology, was its emphasis on multilingual resources for both language generation and understanding, which played a key role in both user interfaces and quality assurance (Baud et al., 1997).

What other opportunities are there for systematic quality assurance, internal and external? If ontologies are to become part of biomedical systems, some will be safety-critical. What sorts of internal quality assurance are possible? Have been implemented? How does it lend itself to external validation? Are there forms easily understood by domain experts? Comparison to related resources? Have external validations been performed? Have any comparative evaluations been performed?

Knowledge changes. Can the ontology be easily maintained as knowledge evolves? Are there mechanisms for doing so? For repeating the quality assurance? To log errors detected and avoid their recurrence?

Finally, there is the issue of how well the proposed formulation fits with widely used standards or how well it can be made to interact with them. Issues of standards and use by collaborators are likely to be critical criteria for acceptance.

6. Practical issues specific to diseases

There are some common practical issues related to diseases that any formulation ought to address. When should a disease class be split into two? When should two disease classes be combined? For example, historically, there was a long debate about whether the peculiar abdominal symptoms in what we now recognise as "abdominal migraines" should be included as a kind of migraine – which to many meant "headache" by definition. Splitting of diseases is more common and becoming ever more so as our biological knowledge improves, witness the number of different subtypes of different cancers, or even simpler cases such as type 1 and type 2 diabetes mellitus. Looking at the history of such names, explicating the causal mechanisms has often been the deciding criteria. This sits well into the River Flow Model's notion of causal chain. How might be handled in the OGMS? Can the OGMS' notion

[1] This issue is being addressed in recent revisions of SNOMED CT.

of "maximally connected" "causally relatively isolated" collection of "abnormal physical components" accommodate such cases? What about Schulz et al.'s (2011) notion of "situation"?

A related issue, when might a patient have two intercurrent diseases and when might she have one complex disease? As a simple example, both congestive heart failure and pneumonia are typified by lung congestion. The treatment implications of each are very different. Determining whether one or the other or both are present is, therefore, critical. This is a notoriously difficult and important problem for systems for clinical decision support.

How does the ontology handle "syndromes" in either of the two senses of the word: (a) poorly understood collections of manifestations that commonly occur together or (b) well understood processes that result in multiple, seemingly independent manifestations – the classic example cited by Schulz et al. (2011) in their argument to interpret as a "situation" Tetralogy of Fallot – a combination of congenital abnormalities of the heart now understood to occur because of a specific failure in the developmental process?

Finally, how does the ontology cope with inherent ambiguity of the notion of disease and the heterogeneity of the collection of entities we call "diseases"? Some disease entities seem focused on structure (e.g., tumours and fractures), whereas others seem focussed on processes (e.g., infections and hormonal imbalances) and some processes may become self-sustaining once initiated, with little or no obvious structural foundation. This ambiguity thwarts the dichotomy in many ontologies between *continuant* and *occurrent*. The causal chains of the River Flow Model appear to focus process (although causal chains are themselves classified as continuants). In OGMS, *dispositions* are related to processes, but the *disorders* that give rise to dispositions are defined as structures. In the situation interpretation, *Situations* are defined as portions of a patient life during which either kind of abnormality may be present. Can these formulations be reconciled? How does each formulation cope with cases that do not seem intuitively to fit its perspective?

7. Conclusions and questions for disease ontologies

Clearly many of these issues go far beyond the formulation of "disease" itself to the entire artefacts that are based on these formulations, and even to the organisations that build and maintain them. However, for the potential user wishing to apply a disease ontology, the choice of ontology needs to be based a wide range of issues. Where, as is inevitable, there are limitations in how the ontology can be used, the limitations need to be acknowledged and dealt with. As a summary, therefore, I list ten questions that developers intending to apply disease ontologies might usefully ask and advocates of disease ontologies be prepared to address. I invite others to contribute to the list or related debate.

(1) What's it for? For what *purpose* was the ontology developed? Does that purpose match the proposed application?
(2) What can it do? What are its *capabilities*? Does it meet the requirement of the application? Are there tests available (e.g., a set of plausible queries to be answered)? If it requires supplementation, are such supplements available? If not, can the effort to develop them be estimated?
(3) Are its formulations *appropriate* to the application? How does it cope with limitations in the formalism? With special issues such as hypotheticals, syndromes and multiple intercurrent diseases?

(4) Is it well specified? What is the *formal basis*? Can it be implemented repeatably? If based on logic, what subset? If based on a formalism that is not logic-based, is the specification rigorous and clear?
(5) Is it *computationally tractable*? What computational resources are required? Will it scale? What performance can be expected?
(6) Can it be understood? Are the *human factors* implications appropriate to the application? Have there been tests of reliability in use? Experiences of the training and other resources required during deployment.
(7) What *language resources* are included? Is there more than a single set of names? Are there multilingual versions? If not, are there any hooks to make it easy to do so? Has it been used in National Language Processing?
(8) Can it be *maintained*? Does the formulation help in ontology evolution? Technically? In terms of human factors and resources? Is it maintained? Will it evolve?
(9) What *quality assurance* is possible? Has it been performed? Internally? Externally? Against what criteria? Is it included in the maintenance process?
(10) Who uses it? What is the *community of use*? Is it a standard? *De facto*? *De jure*?

References

Andrews, J.E., Richesson, R.L. & Krischer, J. (2007). Variation of SNOMED CT coding of clinical research concepts among coding experts. *Journal of the American Medical Informatics Association*, *14*(4), 497–506.

Baud, R.H., Rodrigues, J.-M., Wagner, J.C., Rassinoux, A.-M., Lovis, C., Rush, P., et al. (1997). Validation of concept representation using natural language generation. *Journal of the American Medical Informatics Association*, Fall Symposium Supplement, 841.

Beisswanger, E., Schulz, S., Stenzhorn, H. & Hahn, U. (2008). BioTop: An upper domain ontology for the life sciences: A description of its current structure, contents and interfaces to OBO ontologies. *Applied Ontology*, *3*(4), 205–212.

Green, J.M., Wilcke, J.R., Abbott, J. & Rees, L.P. (2006). Development and evaluation of methods for structured recording of heart murmur findings using SNOMED-CT® post-coordination. *Journal of the American Medical Informatics Association*, *13*(3), 321–333.

Grice, H.P. (1957). Meaning. *Philosophical Review*, *66*, 377–388.

Hastings, J., Dumontier, M., Hull, D., Horridge, M., Steinbeck, C., Sattler, U., Stevens, R., Hörne, T. & Britz, K. (2010). Representing Chemicals using OWL, Description Graphs and Rules. In *OWLED*. Available at: http://researchspace.csir.co.za/dspace/bitstream/10204/4919/1/Britz1_2010.pdf.

IHTSDO (2015). SNOMED CT home page. Available at: http://www.ihtsdo.org/snomed-ct/ [retrieved 2 April 2015].

LOINC (2015). Laboratory Observations Identifiers Names and Codes (LOINC) home page. Available at: https://loinc.org [access date 13 July 2015].

NCBO (2015). BioPortal. Available at: http://bioportal.bioontology.org/ [retrieved 3 April 2015].

Rector, A., Brandt, S. & Schneider, T. (2011). Getting the foot out of the pelvis: Modelling problems affecting use of SNOMED CT hierarchies in practical applications. *Journal of the American Medical Informatics Association*, *18*, 432–440.

Rothschild, A.S., Lehmann, H.P. & Hripcsak, G. (2005). Inter-rater agreement in physician-coded problem lists. In *AMIA Annual Symposium Proceedings* (Vol. 2005, pp. 644–648). American Medical Informatics Association.

Rovetto, R.J. & Mizoguchi, R. (2015). Causality and the ontology of disease. *Applied Ontology*, *10*(2), 79–105.

Scheuermann, R.H., Ceusters, W. & Smith, B. (2009). Towards an ontological treatment of disease and diagnosis. *Summit on Translational Bioinformatics*, *2009*, 116–120.

Schulz, S., Cornet, R. & Spackman, K. (2011). Consolidating SNOMED CT's ontological commitment. *Applied Ontology*, *6*(1), 1–11.

Schulz, S., Rector, A., Rodrigues, J.M., Chute, C.G., Üstün, B. & Spackman, K. (2012). Ontology-based convergence of medical terminologies: SNOMED CT and ICD 11. In *Proceedings from eHealth 2012 – Health Informatics Meets eHealth*, Vienna, Austria.

Smith, B. (1998). The basic tools of formal ontology. In *Proceedings from Formal Ontology in Information Systems (FOIS)*. Amsterdam.

UMLS (2003). The UMLS semantic network. Available at: http://semanticnetwork.nlm.nih.gov/ [retrieved April 2015].

WordNet (2015). WordNet home page. Available at: https://wordnet.princeton.edu/ [access date 13 July 2015].