e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-1038

What's in a Class? Lessons Learnt from the ICD – SNOMED CT Harmonisation

Stefan SCHULZ, Jean-M. RODRIGUES, Alan RECTOR, Kent SPACKMAN,
James CAMPBELL, Bedirhan ÜSTÜN, Christopher G. CHUTE, Harold SOLBRIG,
Vincenzo DELLA MEA, Jane MILLAR, Kristina BRAND PERSSON
WHO – IHTSDO Joint Advisory Group (JAG)
http://apps.who.int/classifications/whoihtsdo

Abstract. The upcoming ICD-11 will be harmonized with SNOMED CT via a common ontological layer (CO). We provide evidence for our hypothesis that this cannot be appropriately done by simple ontology alignment, due to diverging ontological commitment between the two terminology systems. Whereas the common ontology describes clinical situations, ICD-11 linearization codes are best to be interpreted as diagnostic statements. For the binding between ICD codes and classes from the ontological layer, a query-based approach is favoured.

Keywords. ICD, SNOMED CT, Ontology, Terminology, Classification

Background

On a global scale, the International Classification of Diseases (ICD) is being used for morbidity and mortality statistics, as well as for billing in many jurisdictions. The WHO is currently preparing ICD's 11th release [1]. This process is different from past revisions in several aspects, regarding web-based authoring workflows [2], but also its overall architecture, which is more comprehensive and flexible:

ICD's *Foundation Component* (**FC**) is the central hub in which the terminological content needed for the creation of multiple, purpose-specific views or linear serializations called *Linearizations* (**LIN**s) are pooled, together with a rich set of additional data defined by a content model including text definitions, diagnostic criteria and classification rules. FC hierarchies are heterogeneous. They are neither mutually exclusive nor exhaustive. In contrast, LINs manifest the properties of strict mono-hierarchies with exclusions and residual classes (NEC, NOS), demanded for statistical classifications. LINs will support statistics and a smooth transition from ICD 10 to 11.

ICD 11 FC and, indirectly, the linearizations (LINs) will rely on a common model of meaning, the *Common Ontology* (**CO**). According to an agreement by a WHO – IHTSDO expert group, the Joint Advisory Group (JAG) [3] the CO will be built on ontological principles and shared with a subset of SNOMED CT [4].

That terminology systems do not restrict themselves to collections of terms but commit to ontological principles and use logical axioms dates back to the pioneering GALEN project [5]. In the meantime this tendency has been stimulated by the Semantic Web, especially the OWL language [6] and Description Logics [7]. The emergence of *Applied Ontology* as a new discipline [8] bridges between the theoretical achievements of analytic philosophy and the need to create theoretical foundations for semantic artefacts, both formally rigid and usable in real-world information systems. In this it follows the precedent of endeavours such as the Gene Ontology [9] in particular and the OBO Foundry [10] in general.

1. Ontological principles

What are the implications of this for the harmonization of ICD-11 and SNOMED CT? The following aspects are fundamental for ontology-based artefacts:

Logical framework: Description logics require a clear distinction between individual entities, classes of individuals, and relations. They require distinguishing the *subclass* relation between classes, the *membership* relation between an individual and a class, and *object properties*, which relate individuals.

Ontological commitment: With the view on the target domain, a precise and unambiguous understanding about the kinds of entities to be represented is required [11]. Top-level classes such a *Process* or *Material Entity* should be clearly characterized, as well as domain top-level classes, like *Procedure*, *Finding*, *Disorder* (such as in SNOMED CT). In this vein, the JAG and the SNOMED community have concluded that *Finding* and *Disorder* concepts in SNOMED CT should be interpreted as a special kind of *Process*, *viz. Clinical situations*, i.e. phases of a patient's life in which one or more well-distinguished conditions of clinical interest (e.g. a phenotype, a pathological body part, a risk, a disease process) are fully present [12].

Taxonomic order. There is a clear criterion for the arrangement of classes into taxonomies: *A* is a subclass of *B* if and only if all members of *A* are also members of *B*, i.e. if and only if a clinical situation fits the definition of *A* it also fits the definition of *B*.

2. The nature of statistical classifications

It is tempting to consider statistical classifications like ICD as just another case of terminologies, to be underpinned by an ontology conforming to the above principles. Our initial postulate that all is-a links in ICD should coincide with direct or inferred is-a links in SNOMED CT cannot be sustained, due to the complexity of the SNOMED CT architecture and the idiosyncrasies of ICD. To serve their intended purposes, ICD linearizations must conform to different principles. Our analysis has shown:

- For statistical reporting, LINs maintain hierarchies of disjoint and mutually exhaustive classes at each level. To achieve this, even when the logical meaning of the classes is not disjoint and the list of classes cannot be guaranteed exhaustive, exclusions are required that cannot be interpreted in strict logical terms. For example, "*Hypertensive heart disease*" is in the cardiovascular chapter, but "*Gestational hypertension*" is under *Pregnancy, childbirth and puerperium*. Using OWL, the ICD-specific meaning of "Hypertensive heart disease" could be harmonized with the SCT concept "Hypertensive heart disease" by the axiom:

  icd:*HypertensiveHeartDisease*" equivalentTo
  sct:*HypertensiveHeartDisease*" and not sct:*GestationalHypertension*

  This would, then allow to infer – from the ICD code "*Hypertensive heart disease*"

– that the patient is not pregnant. This contradicts the current pragmatics of ICD coding as much as guidelines for its proper use.

- Numerous ICD classes seem to be motivated by convenience or epidemiological principles rather than by ontology or logic. Figure 1 shows a part of the malignancy section of ICD 11 *Joint Linearization for Mortality and Morbidity Statistics*. The plural labels in all levels but the lowermost one (with codes), suggests an interpretation as simple groupings of codes.
- Parent-child relations in ICD linearizations are sometimes only approximate, e.g. AB80.58 *Postthrombotic syndrome* as a child of AB80.5 *Chronic peripheral venous insufficiency*. A situation after an acute deep vein thrombosis, due to a coagulopathy, may not always imply chronic venous insufficiency. However, such cases have been judged sufficiently rare to be disregarded for statistical reporting.
- LIN Classes mix epistemic and ontological issues: the three sibling classes 1E10.1 *Suspected rabies*, 1E10.2 *Probable rabies*, 1E10.3 *Confirmed rabies* are under 1E10 *Rabies*, which is a child of *Infections due to Rabies virus*. Under ontological scrutiny, it is prohibitive that the situation of a patient, of whom rabies is merely suspected, is classified as a rabies situation. This so-called epistemic intrusion [13] is characteristic of terminology systems that blend the reference to a type of domain entity with the state of knowledge or belief of the terminology user.

The examples demonstrate a general problem, *viz.* to interpret classification "classes" in strictly logical terms such as by assuming description logics semantics, where the notion of "class" is different.

- *Neoplasms*
  - *Malignant neoplasms*
    - *Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissues*
      - *Malignant neoplasms of digestive organs*
        - *Malignant neoplasms of liver and intrahepatic bile ducts*
          - *2C90 Malignant neoplasm of liver*
          - *2C91 Malignant neoplasm of intra-hepatic bile ducts*
          - *2C9Y Other specified malignant neoplasms of liver and intrahepatic bile ducts*
          - *2C9Z Malignant neoplasms of liver and intrahepatic bile ducts without mention of type*

Fig. 1. Sample hierarchy from a ICD-11 linearization draft

One could criticize ICD as being ontologically improper. But then one would misjudge its function and risk rendering it less rather than more useful for its purposes. The following aspects should be taken into consideration:

- <u>Non-disruptive evolution</u>. Health statistics over time should be affected as minimally as possible by changes in the underlying coding vocabulary. ICD has evolved for more than 120 years, which explains most of its structure. , especially the single-hierarchy principle.
- <u>Honest limits on precision</u> ("unavoidable vagueness"). Medical statements are often fuzzy and diagnoses are approximate, especially where sophisticated diagnostic procedures are not available, which is the case in many low- and middle-income countries. Epidemiological data always bear a certain bias, which reduces the relevance of a certain "improperness" in the hierarchies

- <u>Use case and action orientation</u>. There are diseases that are difficult to diagnose, and where suspicion requires action, such as in the case of rabies. This explains why in some cases epistemic criteria are distinctive and required for ICD's proper use. Statistical reporting will fail to reflect the realities of medical practice if such cases cannot be recorded accurately.

ICD as used (ICD 11 LINs and previous tabular ICD versions) is not an ontology and should not be criticized for not conforming to all criteria for ontology or terminology well-formedness [10,14]. The structure of ICD LINs is the result of an evolutionary process, driven by users' needs. It would be more appropriate to consider the ICD LINs as classifications of diagnostic statements rather than of diseases. As such, there are analogies with information models, e.g. openEHR archetypes or HL7 clinical models. Under this assumption, the meaning of a hierarchical link between *A* and *B* changes radically. Whereas a domain ontology would reject a subclass relation between *A* and *B* as soon as there is one single member of *A* that is not member of *B*, this would be tolerable between nodes of information models, as they represent information entities and not the clinical reality. Revisiting the rabies example, the subclass link between 1E10.1 and 1E10 would be justified if the latter is rephrased as *Information about rabies*, while the former would have to be interpreted as *Suspected information about rabies.* "Suspected" modifies the information entity, not the disease.

A decision that ICD LINs commit to information entities (on clinical entities) has the promise of a clear solution, but several problems remain to be solved:

- The relation between LIN codes and the underlying CO classes must be formalised, and the attribution of unintentional meanings to codes has to be avoided. E.g., although hypertension excludes hypertension in pregnancy, coding patients as having hypertension should not imply that they are not pregnant. We suggest using queries such as that below to select just those CO codes that apply.

  SELECT *?code* WHERE (*?code* **is_subcode_of** *Hypertensive heart disease*)
  MINUS (*?code* **is_subcode_of** *Disorders of Pregnancy*)

- The problem of representing codes with epistemic content such as "suspected…" needs to be further addressed so as not to imply the existence of entities that may not exist. One approach is to use expressions of the form "*Information structure* that **is_about_situation** only *Rabies situation*" since such expressions do not imply that there actually is any rabies, only that it is not something else. However, there are technical difficulties in such models, as they lead to the possibility of statements that are not about anything at all. A second possibility is to allow hypothetical entities. More satisfactory but more complicated would be the use of a higher order logic so that the uncertainty can be correctly targeted on the statement or belief rather than on the underlying state of the world. However, this is likely to remain beyond the scope of easily used computational logics for the near future, so that approximations using description logic are required.
- It would have to be decided whether ICD codes that refer to classes of information are considered individuals (related by the transitive relation **is_subcode_of** [15]), whereas, e.g., the *SemanticHealthNet* project treats them as classes of information objects [16]. Related to this question is whether a distinction between a model of codes and an information model proper should be made.
- The fine details of the algorithms for deriving and maintaining the residual classes and exclusions required in the LINs and their relation to the FC.

3. Conclusion

Although ICD has been criticised for not conforming to ontological or other principles of well-formedness, there is, nevertheless, a clear advantage of harmonising the ICD with SNOMED CT, arising from computable re-use of structured clinical data for several purposes. However, on the way towards such a harmonisation it has become clear that ICD cannot be understood as an ill-formed clinical ontology. Statistical classifications are closer to clinical models and must meet specific criteria for their use in statistical reporting and epidemiology. Therefore, if harmonisation is to be achieved, a multi-layer architecture is required. For the layer that corresponds most closely to the hierarchies in the existing ICD versions – the linearizations – there are good reasons that their codes should be considered information entities, linked to but distinct from domain entities, that are represented in the subset of SNOMED CT which will form the IHTSDO / WHO common ontology (CO). Further work will focus on the formalization of the binding between ICD codes and CO classes, almost certainly based on queries against the ontology as illustrated in the SELECT statement above.

Acknowledgments

This work was supported by the World Health Organization (WHO) and the International Health Terminology Standards Development Organization (IHTSDO).

References

[1] The International Classification of Diseases 11th Revision is due by 2015. The World Health Organization http://www.who.int/classifications/icd/revision/en/ (last accessed 1 May 2014).

[2] Tudorache T, Falconer S, Nyulas C, Storey MA, Ustün TB, Musen MA. Supporting the Collaborative Authoring of ICD-11 with Web Protégé, Proceedings of the AMIA Annual Symposium 2010, 802-806.

[3] Rodrigues JM et al. Sharing Ontology between ICD 11 and SNOMED CT will enable seamless re-use and semantic interoperability. Studies in Health Technology and Informatics 2013;192:343-346.

[4] International Health Terminology Standards Development Organisation (IHTSDO): SNOMED CT: http://www.ihtsdo.org/snomed-ct/ (last accessed 1 May 2014).

[5] Rector A, Rossi-Mori A, Consorti MF, Zanstra P. Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation, Int. Journal of Medical Informatics 48 (1998), 71-84.

[6] W3C OWL Working Group: OWL 2 Web Ontology Language Document Overview 2009, http://www.w3.org/TR/owl2-overview/ (last accessed 1 May 2014).

[7] Baader F et al. The Description Logic Handbook: Theory, Implementation, and Applications, 2nd Edition. Cambridge University Press, 2nd ed. 2007.

[8] Smith B. Applied Ontology: A New Discipline is Born, Philosophy Today 12 (29), 1998,5-6.

[9] Ashburner M et al. Gene Ontology: tool for the unification of biology. Nature Genetics 2000, 25:25.

[10] OBO foundry principles: http://obofoundry.org/ontologies.shtml (last accessed 1 May 2014).

[11] Schulz S, Cornet R, Spackman K. Consolidating SNOMED CT's ontological commitment. Applied Ontology 6, 2011, 1-11.

[12] Schulz S, Rector A, Rodrigues JM, Spackman K, Competing interpretations of disorder codes in SNOMED CT and ICD. AMIA Annu Symp Proc. 2012, 819-827.

[13] Bodenreider O, Smith B, Burgun A. The ontology-epistemology divide: A case study in medical terminology. Proceedings of FOIS 2004: IOS Press, 2004, 185-195.

[14] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods of Information in Medicine 37(4-5), 1998, 394-403

[15] Rector AL, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. Applied Ontology 4 (1), 2009, 51-69

[16] Schulz S, Martinez-Costa, C. How Ontologies Can Improve Semantic Interoperability in Health Care. Lecture Notes in Computer Science Volume 8268, 2013, 1-10.