MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-940

Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration

Stefan Schulz[a], Jean-Marie Rodrigues[b,c], Alan Rector[d], Christopher G. Chute[e]

a IMI, Medical University of Graz, Austria, b INSERM U1142, LIMICS, Paris, France
c Department of Public Health and Medical Informatics, Univ. Jean Monnet of Saint Etienne, France
d University of Manchester, United Kingdom
e Division of General Internal Medicine, Johns Hopkins University, Baltimore, MD, 21287, USA

Abstract

The time has come to end unproductive competitions among different types of biomedical terminology artefacts. Tools and strategies to create the foundation of a seamless environment covering clinical jargon, clinical terminologies, and classifications are necessary. Whereas language processing relies on human interface terminologies, which represent clinical jargon, their link to reference terminologies such as SNOMED CT is essential to guarantee semantic interoperability. There is also a need for interoperation between reference and aggregation terminologies. Simple mappings between nodes are not enough, because the three kinds of terminology systems represent different things: reference terminologies focus on context-free descriptions of classes of entities of a domain; aggregation terminologies contain rules that enforce the principle of single hierarchies and disjoint classes; interface terminologies represent the language used in a domain. We propose a model that aims at providing a better flow of standardized information, addressing multiple use cases in health care including clinical research, epidemiology, care management, and reimbursement.

Keywords:

Terminology as Topic; Dictionaries as Topic; Knowledge Bases

Introduction

The evolution of electronic health records has been accompanied by the development of numerous and increasingly sophisticated lexical, terminological and ontological tools and resources, supporting structured data entry and processing of unstructured narratives [1]. A major challenge is to preserve meaning from descriptions of individual subjects of care to descriptions of populations, as well as from medical care to biomedical research. Attempts to do so have often neglected the fundamental reconciliation between different genres of biomedical terminologies used for these different purposes and use cases [2]. These genres include ontologies, reference terminologies, interface terminologies, and classifications (in this paper referred to by the more general term "aggregation terminologies" [3]). The reconciliation approach in [4] proposes to address this issue in two steps, links to and from:

- Interface terminologies and reference terminologies
- Reference terminologies and aggregation terminologies.

The two-step approach is required to prevent "confusion of concepts and the words used to express those concepts" [5], which is essential in order to achieve better interoperability at a time where clinicians, documentation specialists, epidemiologists, health care administrators, payers and health service researchers increasingly require that clinical data captured at one place be processed and analysed in different application contexts.

Methods

In the following we analyse the different terminology genres (interface, reference, and aggregation terminologies) and provide arguments that justify this distinction.

Interface vocabularies

Mainstream work on terminology and ontology during the last twenty years has been guided by a normative perspective, primarily driven by the English speaking community. Apart from increasingly incorporating principles of Applied Ontology [6] into terminologies and thesauri (with the Gene Ontology [7] and SNOMED CT [8] being the most prominent examples), the labelling of the nodes in these system has mainly followed a top-down strategy, with naming conventions emphasizing maximally self-explanatory and unambiguous labels such as "Malignant tumour of thyroid gland (disorder)". However, these labels, as clearly understandable as they are, do often not represent the language used by clinicians (e.g., "Thyroid Ca").

This gap is typically filled by (human) interface terminologies [9], i.e. collections of language expressions that actually occur in medical documentation. Such interface terms are typically the building blocks of clinical narratives but also are used as text values for structured data entry. However, there are several issues with interface terms, which often make them unsuitable for labelling reference terminology content:

- Interface terms tend to be as short as possible, and therefore ambiguous out of context. Abbreviations and acronyms play a major role, e.g. "CA" may mean "calcium", "cancer", and "cholic acid".
- Interface terms have different meanings in different user groups, characterised by medical professions and medical specialties, regional dialects and geo- graphic

names (e.g. "GWB": "general well-being", but in New York hospitals also: "George Washington Bridge").

- The meaning of interface terms may change across time, e.g., the acronym "AIDS" has been used, for a long time, for "Acquired Immunodeficiency Syndrome", although other expansions such as "Acquired Iatrogenic Death Syndrome" can be found in the literature.

As a consequence of the dynamics of clinical and scientific language, good interface terminologies require continuous maintenance. Interface terms need to be harvested from "living" language sources. They need to be set in a context, which makes their different meanings transparent, e.g. "Ureter Ca", "Ca level", instead of just "Ca". Only under these conditions, they can be reliably anchored within reference terminologies.

Reference terminologies

In contrast to interface terminologies, reference terminologies should provide stable and well-defined representational units (aka "concepts", "classes", "descriptors" or – confusingly – "terms"). The stability of these units relies not only on unambiguous textual labels, but also on textual definitions or scope notes, links to external standards, as well as on formal-ontological definitions usually based on Description Logics [10], typically using or referring to the OWL [11] language, like in SNOMED CT. E.g., the SNOMED CT concept Pancreatitis is defined as being logically equivalent to a disorder with inflammatory morphology that is located at some pancreas structure.

Connection of reference terminologies with other terminology system must address epistemically-"infested" content [13], i.e. reference to a concept within a discourse context that expresses negation, doubt, intention or risk. Interface terminologies may include terms like "suspected leukaemia" (in some languages like German even fused in a single term, "Leukämieverdacht"). The same is found in aggregation terminologies like ICD-10 "Glaucoma suspect" (H40.0) or "Observation for suspected tuberculosis" (Z03.0). This requires that the reference terminology provides a mechanism to deal with epistemic contexts. SNOMED CT's attempt to this is the context model (the "Situation with explicit context" hierarchy branch), which, however, exhibits several weaknesses under ontological scrutiny [14]. An ontologically founded model to represent both clinical entities and information entities and to connect them with each other was proposed in SemanticHealthNet [15]. This approach used OWL expressions to describe the compositional structure of information models, all of which under the BioTopLite class "information object". The relation between information objects and clinical entities or classes thereof is done via the object property "represents".

Aggregation terminologies

Aggregation terminologies contain rules that enforce the principle of single hierarchies and disjoint classes. This makes them mostly suited for statistical analyses. The most important aggregation terminology is the International Classification of Diseases.

Our experience of linking reference terminologies with aggregation terminologies is based on the ICD – SNOMED CT harmonization process, including a preliminary SNOMED CT

– ICD 10 mapping based on expert knowledge of both coding and medicine [16], and on extensive work on the 11th revision of ICD [17]. ICD-11 was designed on top of a multi-component architecture [18]. We here focus on the component to be released first, probably in 2018, viz. the Mortality and Morbidity Statistics [19].

Fig.1 shows the three building blocks of a terminology ecosystem constituted by clinical language resources (left), reference terminologies with or without ontological foundation (centre) and the building blocks of advanced or aggregation terminologies.

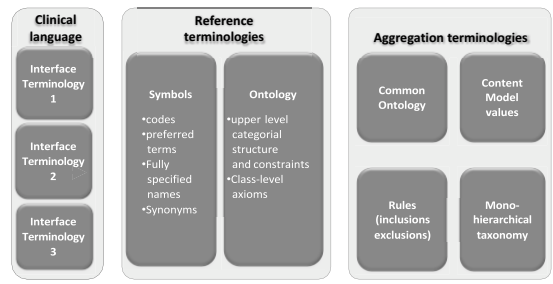

Figure 1- Three building blocks of a terminology ecosystem constituted by clinical language resources (left), reference terminologies with or without ontological foundation (centre) and building blocks of (advanced) aggregation terminologies.

Results

In this section we present the results of our scrutiny of current terminology systems, propose a general typology and provide some recommendations for their further evolution.

Lack of interface terms in reference terminologies

The need of interface terms becomes obvious when matching terminologies with clinical narratives such as clinical notes or discharge summaries. In an experimental study on the coverage of English and Swedish SNOMED CT releases used as annotation vocabulary for a corpus created out of hybrid clinical document samples in several languages [21], a nearly equivalent rate of concept coverage (87%) contrasted with a neatly different, and generally lower coverage of terms; with 47% for Swedish compared to 68% for English. This difference is due to the fact the Swedish SNOMED CT has only one term per concept, whereas the English SNOMED CT version has more than two, on average. These results demonstrate the need for language-specific interface terminologies, and they put in question whether a simple enrichment of a clinical reference terminology with interface terms, as practiced by IHTSDO for English and Spanish is really sufficient.

Local interface terminology efforts are needed

We support ongoing national terminology building efforts as decentralised bottom-up activities, starting with a systematic collection of commonly used words and phrases in daily communication between patients and health professionals. For instance, an effort to build an interface terminology for German linked to SNOMED CT codes semi-automatically has already resulted in more than 1.8 million interface terms, not including short forms like acronyms and abbreviations for which methods of disambiguation and resolution are currently tested [22]. If such efforts are costly, they are more helpful for guaranteeing the use of reference terminologies and the seam-

less flow of meaningful clinical information than huge top-down reference terminology translation efforts.

It can also be useful to relate interface terminologies with thesauri like MeSH. Thesauri lack formal-ontological foundation [12], but provide precise textual definitions. Here, Pancreatitis is described by the scope note "Inflammation of the pancreas". Although such a description does not use a formal language it may support a mapping to an ontology-based reference terminology such as SNOMED CT.

General typology of terminology genres and interfaces

Based on our experience we propose the following distinction between terminology systems to integrate the three terminology blocks and the two interfaces between the blocks:

(First order) axiom-based systems – generally using description logics (DLs), which provide axioms for sub-class / superclass relations and existential restrictions (e.g. "every instance of A is located in some instance of B"). From these axioms, class hierarchies may be inferred algorithmically by DL reasoners [23]. Typically, this leads to poly-hierarchies. The semantics are that all statements are necessarily true in all possible interpretations. Such systems are open-world [10], and negation means necessarily false, i.e. false in all possible interpretations. No exceptions are allowed. All statements are first order, i.e. about all individuals in a class; statements about the classes themselves are not allowed.

Closed world systems – e.g. logic programming and database systems – have in common that their semantics is based only on what is held explicitly in the system – hence closed world. If hierarchies are present, they must be stated explicitly and cannot be inferred. False means "not provable in the closed world of the system" Therefore, new information about that world can falsify previous conclusions; exceptions may therefore be allowed. Consequently, universal subclass-superclass relations that must hold in any world cannot be inferred, because closed world statements can be proved in the closed world, they just cannot be proved universally for any world. This is typical of rule-based systems. They prescribe what to do in particular situations e.g., in languages for decision support systems. They are the foundation for mono-hierarchical aggregation terminologies and for queries on representations: Mono-hierarchical aggregation terminologies (aka statistical classifications), characterized by single hierarchies and disjoint classes, supported by a large corpus of exclusion and inclusion rules, as well as by coding guidelines, which vary between use cases (e.g. coding for reimbursement vs. coding for mortality statistics) and local contexts. Examples are ICD-9 and ICD-10, the upcoming ICD-11 linearization(s), other WHO classifications, and national catalogues (like, e.g. the German OPS procedure classification) [24].

Queries on representations – which may be used to extract information about any of the above kinds of representation artefacts – as opposed to the knowledge represented. The query languages SPARQL [25] with its specialised extension for DLs and the SNOMED CT Expression Constraint Language[26] are particularly important in medical applications for linking axiomatic systems to aggregation terminologies, be- cause they support formalizations of the characteristic residual classes not classified under and not elsewhere classified. Ex- amples for exclusions are the ICD-10 classes under I10-I15, characterised as "Hypertensive diseases, excluding complicat- ing pregnancy, childbirth and the puerperium".

Knowledge Organization Systems (or “terminologies” in general) provide hierarchies based on loosely defined "broader-than" / "narrower-than" relations between terms or groupings thereof, which are first of all seen as streamlining the navigation among human language terms for which also synonymy and other semantic relations can be asserted. The most popular knowledge organization system in the field of biomedicine is the MeSH thesaurus [12] with its multiple tree structure and entry terms. In MeSH, negation has no formal meaning, nor is it built upon any explicit ontological foundation.

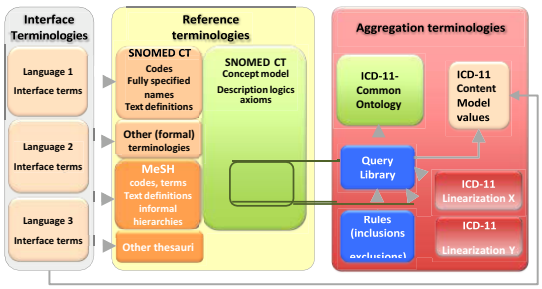

Figure 2- Architecture from Fig. 1, applied to SNOMED CT and ICD-11, with one or more linearizations (e.g. the ICD-11 Mortality and Morbidity Linearization [19] as example of an aggregation terminology as a possible end product. The ICD-11 common ontology is a subset of SNOMED CT. Linearizations are built using language-specific labels from interface terminologies (value sets), which can also be extended by clinical language synonyms, e.g. by additional thesauri like [20]. The colours have the following meaning: Green: Axiom based systems; Pink: Closed world systems including dictionaries; Orange: knowledge organization systems; Blue: queries on representations and annotations; Red: Mono-hierarchical aggregation terminologies .

Fig. 2 summarizes the links between three blocks of a specific terminology ecosystem, with different kinds of resources and technologies, regarding the content of SNOMED CT, a subset of it that qualifies as common ontology for ICD-11, and additional resources that secure the architectural principles of aggregation terminology linearizations.

In closed-world systems like ICD-x and other aggregation terminologies it can be asserted that something is true in the world of the representation. For instance, it is sufficient in the above example, to assert a hypertensive disease code as long as there is no evidence that the patient is pregnant or in the perinatal phase. That is, they are usually true, or true under certain conditions, but not necessarily true by definition. The latter would occur when using axiom-based systems were used for axiomatising the content of aggregation terminologies: Using logical negation for representing hypertensive disease as above would entail that any patient classified as hypertensive was not pregnant or in the perinatal phase. This kind of entailment is not intended by aggregation terminologies.

Therefore, if axiom-based systems such as SNOMED CT are to be linked to aggregation terminologies such as ICD-x, queries on representations are needed. It is the underlying axiom-based system in description logic that allows SNOMED CT to fulfil the twelve Cimino criteria for controlled vocabularies [27, 28]. By its Expression Constraint Language, SNOMED CT provides a means of formulating queries on that representation to bridge between its reference terminology and aggregation terminologies.

In the case of the link between SNOMED CT and ICD-11 Mortality and Morbidity Statistics version [19] as one of different possible end products, the ICD-11 Common Ontology

(Fig.2) is a subset of SNOMED CT. Linearizations are built using language-specific labels from specific interface terminologies (value sets), which can also be extended by clinical language synonyms, e.g. by additional thesauri like [20].

For interface terminologies there is currently no agreement on which formalism to use to anchor them on reference terminologies. A possible Semantic Web standard is SKOS [29], currently aligned with ISO standards for thesauri. Interface terminologies supporting aggregation terminologies, but also value sets that provide canonical names raise another issue, viz. imprecision. Interpreting these terms (like "Diabetes mellitus") at face value, is not consistent with their exact meaning, such as Diabetes mellitus in patients that are not pregnant or in the perinatal phase. Shortcut concept "alignments" that firstly rely on lexical criteria and which (falsely) infer equivalence of meaning from string equivalence of identifiers, are therefore inherently imprecise.

Conclusions

A major difficulty in reconciling interface terminologies, reference terminologies (e.g., SNOMED CT) and aggregation terminologies (e.g., ICD-11) was understanding the roles of the different terminology types and how they are related to the distinctions between (i) kinds of terminological/ontological knowledge, (ii) meaning of language expressions, (iii) the things they denote, (iv) necessary truths about classes of things, and (v) use case specific interface terms, explanations and rules. Our suggested typology of representational artefacts could help prevent difficulties in specifying terminology architectures like the one underlying ICD-11. In order to ease understanding of the distinctions between statements of necessary truths, closed world knowledge, navigation associations, and classifications rules, we suggest the following vocabulary for a component-based architecture as sketched in Fig.2:

- The open world component comprising first-order necessary truths, thus constituting the ontological basis of reference and aggregation terminologies in the near future.
- The closed-world component includes rules that assure the architectural constraints of aggregation terminologies. It is the foundation of two other components:
  - Aggregation terminologies, i.e., classification systems in a broad sense, are constituted by single hierarchies for specific purposes, like statistical reporting and billing. They follow the jointly- exhaustive-mutually-exclusive rule.
  - Query libraries express the meaning of nodes of aggregation terminologies, by querying against reference terminologies.
- The knowledge organization foundations include all other kinds of supportive, more loosely specified knowledge resources including lexicons with interface terms. It provides the framework for multiple interface terminologies to be implemented as dynamic plugins for different languages and communities.

Acknowledgements

This work was supported by the World Health Organization (WHO) and SNOMED International (IHTSDO) through their Joint Advisory Group (JAG). It also refers to work that received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 643818.

References

[1] O. Bodenreider. Lexical, terminological and ontological resources for biological text mining. S. Ananidou et al, Text mining for biology and biomedicine; Artech House, London, UK, 2006, 43-66.
[2] C. G. Chute. Clinical classification and terminology. JAMIA 7;3 (2000), 298-303.
[3] J. Rogers. Using Medical Terminologies. 2005, http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/terminology-using.htm
[4] R. H. Baud, W. Ceusters, P. Ruch, A. M. Rassinoux, C. Lovis, A. Geissbühler. Reconciliation of Ontology and Terminology to cope with Linguistics, Studies in Health Technology and Informatics 129 (2007), 796-801.
[5] A. L. Rector, Clinical Terminology: Why is it so Hard? Methods of Information in Medicine, 38, (1999), 147-157.
[6] B. Smith. Applied Ontology. A new discipline is born. Philosophy Today 12,29 (1998), 5-6
[7] M. Ashburner et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25;1 (2000),25-9.
[8] SNOMED International (aka International Health Terminology Standards Development Organisation - IHTSDO). SNOMED CT – The Global Language of Healthcare (2016), http://ihtsdo.org/snomed-ct
[9] S.T. Rosenbloom, R.A. Miller, K.B. Johnson, P.L. Elkin, S.H. Brown. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. JAMIA13;3 (2006), 277-288
[10] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, P.F. Patel-Schneider. The description logic handbook, second edition 2, Cambridge University Press, Cambridge, UK, 2010
[11] World Wide Web Consortium (W3C). OWL2 Web Ontology Language (2012). https://www.w3.org/TR/owl2- overview/
[12] U.S. National Library of Medicine. Medical Subject Headings (MeSH) (2016) https://www.nlm.nih.gov/mesh/
[13] O. Bodenreider, B. Smith, A. Burgun: The Ontology- Epistemology Divide: A Case Study in Medical Terminology. 3rd Conference on Formal Ontology in Information Systems (2004), 185–195.
[14] C. Martínez-Costa, S. Schulz. Ontology-based reinterpretation of the SNOMED CT context model. International Conference on Biomedical Ontologies (ICBO) (2013) 90-95.
[15] C. Martínez-Costa, R. Cornet, D. Karlsson, S. Schulz, D. Kalra. Semantic enrichment of clinical models towards semantic interoperability. The heart failure summary use case. JAMIA 22;3 (2015), 565-576.
[16] K. Giannangelo, J. Millar, Mapping SNOMED CT to ICD-10, Studies in Health Technology and Informatics 180, (2012), 83-87.
[17] The World Health Organization, The International Classification of Diseases 11th Revision is due by 2017, http://www.who.int/classifications/icd/revision/en/
[18] J.M. Rodrigues, D. Robinson, V. Della Mea, J. Campbell, Rector, S. Schulz, H. Brear, B. Üstün, K. Spackman, C., G. Chute, J. Millar, H. Solbrig, K. Brand Persson. Studies in Health Technology and Informatics, 216 (2016), 790- 794.
[19] M. Mamou, A. Rector, S. Schulz, J. Campbell, H. Solbrig,J.M. Rodrigues. Representing ICD-11 JLMMS Using IHTSDO Representation Formalisms. Studies in Health Technology and Informatics 228 (2016), 431-435.
[20] DIMDI (German Institute for Medical Documentation and Information). ICD-10 Thesaurus of diagnostic terms (IDT). https://www.dimdi.de/static/en/klassi/icd-10-gm/historie/idt.htm
[21] ASSESS CT. Assessing SNOMED CT for Large Scale eHealth Deployment in the EU. http://assess-ct.eu
[22] S. Schulz, J.A. Miñarro-Gimémez. Using Language Technology for SNOMED CT Localisation. IHTSDO Showcase (2015), Montevideo, Uruguay
[23] N. Matentzoglu, J. Leo, V. Hudhra, U. Sattler, B. Parsia. A survey of current, stand-alone owl reasoners. Informal Proceedings of the 4th International Workshop on OWL Reasoner Evaluation, Vol 1387 (2015)
[24] DIMDI (German Institute for Medical Documenation and Information). OPS – German Procedure Classification. https://www.dimdi.de/static/en/klassi/ops/index.htm

[25] W3C Consortium. SPARQL Query Language for RDF(2008) https://www.w3.org/TR/rdf-sparql-query/
[26] IHTSDO. SNOMED Expression Constraint Specification and Guide (Draft). 2016.
[27] J.J. Cimino, From Data to Knowledge through Concept- oriented Terminologies: Experience with the Medical Entities Dictionary, JAMIA 7;3 (2000), 288-297.
[28] J.J. Cimino, In defense of the desiderata, Journal of Biomedical Informatics 39;3 (2006), 299-306.
[29] W3C. SKOS – Simple Knowledge Organization System (2012), https://www.w3.org/2004/02/skos/

Address for correspondence

Stefan Schulz
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
Auenbruggerplatz 2/V A- 8036 Graz (Austria)
stefan.schulz@medunigraz.at