Looking for Anemia (and Other Disorders) in SNOMED CT: Comparison of Three Approaches and Practical Implications

Fleur Mougin, PhD[1], Olivier Bodenreider, MD, PhD[2], Anita Burgun, MD, PhD[3]

[1]LESIM, INSERM U897, ISPED, University Victor Segalen Bordeaux 2, France
[2]National Library of Medicine, Bethesda, Maryland, USA
[3]INSERM U936, EA3888, School of Medicine, University of Rennes 1, IFR 140, France
fleur.mougin@isped.u-bordeaux2.fr

Abstract

Health professionals are faced with challenges when they have to exploit the semantics of concepts present in clinical terminologies in support of research activities. The difficulty lies in the fact that this semantics is represented not only through the labels of concepts, but also their position in the hierarchy, and, when available, their logical and textual definitions. We investigate and contrast the lexical, hierarchical, and logical representations of concepts in SNOMED CT through the example of Anemia *and three other disorders. The four use cases we developed suggest that the lexical, hierarchical, and logical representations of concepts have a limited degree of overlap, but are complementary. Finally, we draw practical implications from our findings for SNOMED CT users and developers.*

Introduction

The semantics of a concept in a terminology is represented through multiple facets [1]. The concept labels (preferred term and synonyms) provide a lexical representation of the meaning. The position of the concept in the hierarchical structure of the terminology is a surrogate for the extensional representation of the concept, as it determines the set of its descendants (i.e., its extension). Finally, definitions (textual and logical), when available, represent the intension of the concept, through genus and differentiae in Aristotelian textual definitions, and through the set of defining roles (or properties) in description logic formalisms [2].

Ideally, the intensional and extensional representations are equivalent, while the lexical representation provides a convenient reference to the concept. In practice, however, the representations are not always equivalent, and users of biomedical terminologies are likely to obtain different results depending on which query methods (lexical, hierarchical, or logical) they use for the terminology.

The objective of this work is to investigate and contrast the lexical, hierarchical, and logical representations of concepts in SNOMED CT through the example of the concept Anemia (271737000). We perform a similar analysis on three additional concepts in order to confirm our original findings, for use cases including eligibility screening for clinical trials, decision support, and quality assurance. We conclude by drawing practical implications from our findings.

Looking for anemia

A clinical research group wants to realize a clinical research study involving patients hospitalized for anemia, based on diagnoses recorded as SNOMED CT concepts in the hospital's electronic health records (EHR) system.

A naive approach would be to search for those patients whose diagnosis is exactly Anemia. However, this approach would miss patients suffering from specific forms of anemia (e.g., "acquired aplastic anemia"). Identifying these patients requires a more sophisticated query, such as "patients whose diagnosis is Anemia or any of its descendants in the SNOMED CT hierarchy".

Another straightforward approach is a "Google-like" method that consists in the identification of all diagnostic terms containing the word "anemia". Here again, this approach appears suboptimal, as it would also fail to identify specific forms of anemia (e.g., "pancytopenia").

Finally, in order to cast a larger net, it might be desirable to include patients diagnosed with a disorder related to anemia, but not itself necessarily a specific type of anemia. Such disorders include, for example, diseases *due to* or *associated with* anemia. This kind of query is more complex and leverages the logical definition of concepts in SNOMED CT. For example, diseases like Pericarditis associated with severe chronic anemia (43742007)[1] would be retrieved through to the role *Associated with*[2], whose value is Chronic anemia (191268006). It must be noted that the query for values associated with roles must not be restricted to the concept of interest, but must rather include any of its descendants. (No concept would have been retrieved here if the associated value had been restricted to the concept Anemia itself, not including its descendants).

None of the three approaches above is totally satisfying, since they all produce different sets of diagnoses, and, as a consequence, different sets of patients selected for inclusion in the clinical research study. The following examples suggest that these approaches can be complementary.

Specific types of anemia are retrieved by the hierarchical approach, but may be ignored by the lexical approach if the concept name does not explicitly contain "anemia" (e.g., "pancytopenia").

Disorders related to anemia without being themselves a specific form of anemia can be retrieved by the logical approach, but are ignored by the hierarchical approach. For example, depending on the requirements of the clinical research study, Sickle cell retinopathy (11603001) could be a disorder of interest, since it is *due to* a specific form of anemia: Hereditary hemoglobinopathy disorder homozygous for hemoglobin S (127040003). However, it would not be included among the disorders selected, because it is (legitimately) not a descendant of Anemia.

Finally, the logical definitions are not always complete and queries based on logical definitions are therefore also likely to result in false negatives. For example, Megaloblastic anemia due to chronic hemolytic anemia (47844003) is neither defined as *Due to* Chronic anemia nor *Due to* Hemolitic anemia (61261009). As a result, a query based on the logical definitions will fail to include patients suffering from Megaloblastic anemia due to chronic hemolytic anemia in the set of patients selected for the study on Anemia.

In addition, the burden on the user to exploit the various representations of a concept can be non-trivial. Access to the lexical representation is generally covered by browsers and terminology servers, but with various degrees of refinement [3]. Exploiting the hierarchical representation requires access to the transitive closure of hierarchical relations (precomputed or computed on the fly), which is not systematically provided by browsers and terminology servers[3]. In order to fully take advantage of the logical representation, it is necessary to use an ontology editor, such as Protégé, which provides access to reasoning services through a Description Logics (DL) classifier, such as Fact++ or Pellet. Finally, the burden on users is compounded by the fact that several approaches must be used and their results combined.

Background

Lexico-syntactic approaches have long been proposed for extracting taxonomic relations from text [4,5], and specific patterns (e.g., adjectival modification) have been applied to biomedical terminologies, where the interplay between lexical and hierarchical representations has been investigated [6]. Dolin et al. studied a method that leverages both the lexical and logical representations. They proved that the combination of the two method was superior to their "sum" [7]. Finally, the difference between the hierarchical and logical representations has been examined in [8], in which the authors investigated the roles defined for a concept and its parent concept. To our knowledge, the three representations, lexical, hierarchical, and logical, have never been studied together in the context of a biomedical terminology.

The Systematized Nomenclature of Medicine - Clinical Terms or **SNOMED CT** (SNCT) is a description logic-based biomedical terminology, which is owned, maintained, and distributed by the International Health Terminology Standards Development Organisation (IHTSDO) [9]. Version 2010/01/31 of the SNCT is used in this study, which was downloaded from the UMLS Knowledge Source Server[4]. It contains 291,205 current[5] concepts and 467,214 synonyms. Concepts are related through hierarchical relations (431,616) and defining relations (716,767 roles). Figure 1 illustrates the concept Anemia with its preferred term and its synonyms. Four of its descendants are displayed and the role *Associated with* between Pericarditis associated with severe chronic anemia and Chronic anemia is also presented.

[1] Sans serif font is used for SNOMED CT concepts

[2] *Sans serif italic font* denotes SNOMED CT roles

[3] Some SNOMED CT browsers provide access to the transitive closure (e.g., through a click on "All descendents and related subtypes" in http://snomed.vetmed.vt.edu/sct/menu.cfm)

[4] http://umlsks.nlm.nih.gov/

[5] The status of active SNCT concepts is defined as "current"

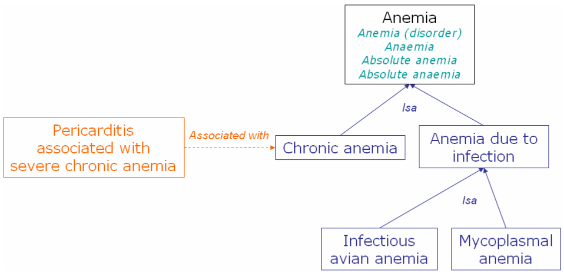

Figure 1. Excerpt of Anemia's SNCT concept representations. The lexical, hierarchical, and logical representations are respectively displayed in green, blue, and orange.

Methods

Defining the three representations. We restrict the *lexical representation* of a concept to its label(s). In practice, labels include the preferred term of the concept and all its synonyms.

The *hierarchical representation* corresponds to the descendants of a concept in the SNCT hierarchy. The set of descendants of a concept consists of the first-generation descendants of this concept (i.e., its children) and their descendants, recursively, all the way to the bottom of the SNCT hierarchy. In graph theory, this operation is called the transitive closure of hierarchical relations.

The *logical representation* is composed of the set of associative relations of a given concept to other concepts. We restrict this representation to the defining roles of a concept, i.e., those relationships representing essential features of the concept. (In this work, we ignore such DL-specific aspects of the logical representation as the fact that the concept is primitive or fully defined).

Building an extension for each representation. For each representation, we created a set of SNCT concepts corresponding to the "extension" of the original concept of interest (Anemia), i.e., the set of concepts retrieved from the original concept using a particular representation. The *lexical extension* is composed of all SNCT concepts having at least one label containing the word "anemia". The *hierarchical extension* corresponds to the descendants of the SNCT concept Anemia. The *logical extension* is the set of concepts whose defining role's value is Anemia, or one of its descendants. Finally, we restricted the extensions by keeping only those SNCT concepts corresponding to disorders.

Results

The results obtained for each representation of the SNCT concept Anemia are displayed in Figure 2.

Extensions. We found 331 concepts containing "anemia" in at least one of their labels. In most cases (95.2%), "anemia" was present in the preferred term of the concept. 16 concepts were however part of the lexical extension through a synonym, e.g., Selective malabsorption of cyanocobalamin (234363001) with "Imerslund-Grasbeck anemia".

A total of 462 descendants of the SNCT concept Anemia were recovered, including Evans syndrome (75331009) and Acquired stomatocytosis (111576004).

Nine concepts are related to Anemia (or any of its descendants) through the roles *Associated with* (1), *Due to* (7), and *After* (1). Examples include Pericarditis associated with severe chronic anemia *Associated with* Chronic anemia, Neonatal jaundice with glucose-6-phosphate dehydrogenase deficiency (206439006) *Due to* Hemolytic disease of fetus OR newborn due to isoimmunization (387705004), and Acute chest syndrome (372146004) *After* Hereditary hemoglobinopathy disorder homozygous for hemoglobin S (127040003).

Comparing the three extensions. We observed that 317 of the 331 concepts found in the lexical extension also belong to the hierarchical extension. Among the 14 concepts missing from the hierarchical extension, eight would have been expected to be present, e.g., Chronic non-spherocytic hemolytic anemia (234402007) and von Jaksch's anemia (234345001). Three concepts are erroneously recovered, two of which because their label indicates the absence of "anemia": Macrocytosis - no anemia (234339009) and Iron deficiency without anemia (234340006). The third one, Tryptophanemia[6] (237925006), "accidently" contains "anemia" in its label, but has nothing to do with Anemia. The three remaining concepts also appear in the logical extension, including Myasthenic syndrome due to pernicious anemia (193213003), which is *Due to* Pernicious anemia (84027009). The only concept present in the all three extensions is Megaloblastic anemia due to congenital deficiency of intrinsic factor (60504009) because it contains "anemia" in its preferred term, is a descendant of Anemia, and is *Due to* Congenital deficiency of intrinsic factor (234361004).

[6] Elevated level of the amino acid tryptophan in blood

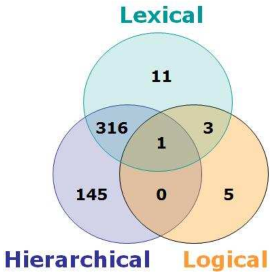

Figure 2. Results for each representation of the SNCT concept Anemia and the intersections of extensions

Applications

The use case we developed with Anemia illustrates the issues and challenges faced by health professionals in their use of clinical terminologies in support of research activities. We now show that our findings on Anemia are relatively typical by investigating three additional concepts with the same methods. We also provide three different scenarios in order to illustrate the applicability of our approach to a wider range of situations.

Eligibility screening for clinical trials (CT). Intracranial haemorrhage (ICH), spontaneous or traumatic, is a devastating event that commonly results in major neurological disabilities. Clinical trials have been designed with several objectives, e.g., improving the conventional neuroradiological evaluation or preventing vasospasm[7]. Screening patients with a particular diagnosis (here, ICH) is often the first step in determining their eligibility to some clinical trial. This step can be greatly facilitated through adequate queries to an EHR system.

Decision support and reminders (DS). Let's consider a medical center whose activities include the follow-up of patients suffering from diabetes mellitus. The diabetologist nurse needs to make sure patients have been tested for the HbA1c in the past three months in order to conform to best practices of care. Reminders will be sent to patients who need to get tested. The date of the last HbA1c test needs to be retrieved from the EHR system for all patients with a diagnosis of "diabetes mellitus".

Quality assurance (QA). In order to evaluate the quality of care in the surgery department where an incident has occurred, the local safety commission has ordered a review of all cases of complications for a given surgical procedure in this hospital. Here again, such cases must be retrieved from the EHR system through queries, e.g., disorders resulting from a "complication of procedure".

The number of concepts present in each of the three extensions for each use case is shown in Table 1. These results confirm what we observed with Anemia: the three extensions of a given concept overlap only partially and appear to be complementary.

Table 1. Results for each representation of the SNCT concepts Intracranial hemorrhage (1386000), Diabetes Mellitus (73211009), and Complication of Procedure (116224001) and the intersections of their extensions

	CT	DS	QA
Lexical	32	172	18
Hierarchical	155	96	1,498
Logical	3	189	10
Lex ∩ Hie	29	90	18
Lex ∩ Log	2	71	0
Hie ∩ Log	0	6	1
Lex ∩ Hie ∩ Log	0	6	0

Discussion

Findings. One of the obvious limitations of the lexical representation has to do with the presence of negation in labels (e.g., Iron deficiency without anemia). This problem is a well-known issue in biomedical natural language processing [10]. The lexical approach is useful nonetheless, as it can help compensate for classification errors and missing taxonomic relations, and thus help complement the hierarchical representation. For example, the concept Ischemic ulcer of toe (429768000) is not described as a child of Ulcer of toe (301021005) in SNCT. This missing taxonomic relation can be uncovered by the lexical approach when searching for "ulcer of toe". It is however not a general solution, especially in the case of complex misclassified concepts. For instance, Hepatorenal syndrome due to a procedure (31005002) should be described as a child of Hepatic failure due to a procedure (22508003).

The hierarchical extension often provides the largest result set, and sometimes includes all concepts from the lexical extension. For example, the 18 concepts containing "complication of procedure" in their label(s) are all included in the set of descendants of the concept Complication of procedure. In contrast, the hierarchical extension tends to differ radically from the logical extension. This is somewhat surprising, as the hierarchical structure of a DL ontology is in part automatically inferred from the set of defining relations. However, the performance of a DL classifier is limited by the completeness of the logical definitions

[7] By querying http://clinicaltrials.gov with the keyword "intracranial h(a)emorrhage" on March 11, 2010, we found 86 studies currently recruiting

and the limited expressiveness of the DL dialect used by SNCT. The overlapping cases correspond to concepts, which are descendants related to other descendants through roles. For instance, Diabetes mellitus, juvenile type, with hyperosmolar coma (190330002) is *Associated with* and *Isa* Diabetes mellitus. Finally, as already illustrated above, the hierarchical representation may be incomplete. As an example, the concept Diabetes mellitus with hyperosmolar coma (190329007) does not belong to the descendants of Diabetes Mellitus, although it clearly should.

The logical representation is arguably the most powerful, as it exposes detailed information about the concept. However, roles are inconstantly populated in SNCT. For example, the concept Sequelae of other non-traumatic intracranial hemorrhage (195242008) is a disorder occurring *after* an ICH. However, in SNCT, this concept has no *After* role related to the concept Intracranial hemorrhage. When present, the roles can be very useful, especially to filter results. In the clinical trial we detailed above, it is possible to exclude concepts about seizures (e.g., Seizures complicating intracranial hemorrhage (371114002)) if the trial is not for disorders complicating an ICH. Reasoners can help analyze data encoded in ontologies. However, because of the presence of historical relations between SNCT concepts (i.e., inherited from previous coding systems) and the lack of computable definitions in SNCT, the logical approach may not perform as well as one would expect. For example, the hierarchy of Finding related to pregnancy (118185001) does not fit the requirements for formal ontology. Indeed, Not pregnant (60001007) is a subclass of Finding related to pregnancy, and neither Pregnancy problem (289209003) nor Disorder of pregnancy (173300003) are associated with logical definition that can be used by reasoners to distinguish between them.

Practical implications. A few practical implications emerge from our study for SNOMED CT users, terminology server developers, and SNOMED CT developers.

Implications for users. If SNOMED CT should support clinical decision applications and the selection of patients for clinical research studies, it is better to use multiple approaches to identifying all concepts of interest for the task at hand, as each approach is likely to only provide a partial solution.

Implications for terminology server developers. Interfaces to EHR systems include terminological components. Such components, often terminology servers or browsers, should offer a wide range of services, of which lexical services and navigation are only a part. Access to reasoning services based on the ontology back end is needed for complex queries.

Implications for SNOMED CT developers. The interplay among lexical, hierarchical, and logical approaches can be exploited for quality assurance purposes. However, beside limitations inherent to the limited expressiveness of the DL dialect used for the creation of SNOMED CT, the main issue remains the underspecification of the logical definition of many concepts. The addition of textual definition will be an opportunity for ensuring the parallel enrichment of the logical definitions.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Sowa J. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Course Technology; 1999.
2. Borgida A, Brachman RJ. Conceptual modeling with description logics. The description logic handbook: theory, implementation, and applications. Cambridge University Press; 2003. p. 349-372.
3. Rogers J, Bodenreider O. SNOMED CT: Browsing the Browsers. KR-MED. 2008.
4. Cruse DA. Lexical Semantics. Smelser NJ, Baltes PB, éds. International Encyclopedia of the Social & Behavioral Sciences. Oxford: Pergamon; 2001. p. 8758 - 8764.
5. Hearst MA. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th conference on Computational linguistics - Volume 2. Nantes, France: Association for Computational Linguistics; 1992. p. 539-545.
6. Bodenreider O, Burgun A, Rindflesch TC. Assessing the consistency of a biomedical terminology through lexical knowledge. Int J Med Inform. 2002 Déc 4;67(1-3):85-95.
7. Dolin RH, Huff SM, Rocha RA, Spackman KA, Campbell KE. Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. J Am Med Inform Assoc. 1998 Avr;5(2):203-213.
8. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. Artif Intell Med. 2007 Mar;39(3):183-195.
9. SNOMED CT. IHTSDO, Copenhagen 2007 http://www.ihtsdo.org/.
10. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-310.