Digital Healthcare Empowering Europeans
R. Cornet et al. (Eds.)

doi:10.3233/978-1-61499-512-8-130

Auditing of SNOMED CT's Hierarchical Structure using the National Drug File - Reference Terminology

Aleksandr Zakharchenko, BS[1], James Geller, PhD[1]
[1]*New Jersey Institute of Technology, Newark, NJ*

Abstract. With the ongoing development in the field of Medical Informatics, the availability of cross-references and the consistency of coverage between terminologies become critical requirements for clinical decision support. In this paper, we examine the possibility of developing a framework that highlights and exposes hierarchical incompatibilities between different medical terminologies in order to facilitate the process of achieving a sufficient level of consistency between terminologies. For the purpose of this research, we are working with the Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) and the National Drug File - Reference Terminology (NDF-RT) – a clinical terminology focused on drugs. For discovery of inconsistencies we built an automated tool.

Keywords. SNOMED CT, Terminology as Topic, Biomedical Ontologies, Health Care Quality Assurance.

Introduction

SNOMED CT (formerly an acronym for Systematized Nomenclature of Medicine) – (Clinical Terms) is a large controlled medical ontology governed by the International Health Ontology Standards Development Organization (IHTSDO) [1, 2]. It is currently considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. In the July 2014 US release, SNOMED CT contained over 300000 active concepts divided into 19 *is-a* hierarchies, represented as Directed Acyclic Graphs (DAGs). SNOMED CT's attribute relationships are not considered in this paper.

In the past, we have conducted extensive research on auditing and quality assurance of terminologies with a focus on SNOMED CT [3-6]. In this paper, we are combining our SNOMED CT work with the National Drug File - Reference Terminology (NDF-RT), produced by the U.S. Veterans Health Administration (VHA) [7]. By comparing the *is-a* hierarchical structures of these two terminologies, it becomes possible to further improve the quality of the coverage of both terminologies.

As different terminologies use various sources of information, differences in coverage are inevitable. Thus, the *is-a* hierarchies of these terminologies are likely to differ significantly in certain areas. An attempt to combine the hierarchies of two terminologies could help with identifying missing concepts as well as suggest better ways to classify concepts in the *is-a* hierarchies. The idea of performing a comparative classification analysis between terminologies has been investigated [8-11], with some approaches focusing on SNOMED CT and NDF-RT. The most detailed analysis of

hierarchical structures of these terminologies is by Mortensen and Bodenreider [10]. In their research, the authors attempted to analyze the direct mapping of classes between SNOMED and NDF-RT based on the concepts those classes include. While their research has shown a low correlation between the two terminologies, the authors suggested that an ingredient-based approach should provide better results. In this paper, we take their recommendation into account and apply ingredient-based semantic mining as our way of harmonizing the terminologies.

1. Methods

For the purposes of this research, we focus on the following similar hierarchies of two terminologies - SNOMED CT's *Pharmaceutical biologic product hierarchy (*further referred to as *Product hierarchy),* SNOMED CT's *Substance hierarchy* and NDF-RT's INGREDIENT_KIND hierarchy. In order to perform a structural audit of SNOMED CT's hierarchies (24870 and 17256 concepts, respectively) through usage of the hierarchical structure of NDF-RT's hierarchy (10118 concepts), we transformed the Directed Acyclic Graphs (DAGs) into trees. This was done through creation of multiple copies of concepts for all the concepts that have multiple parents. While this conversion has added redundancy, it provided the possibility to define the concept of a "layer" of the terminology as a set of concepts that all have the same number of concepts between them and the root (Figure 1). These layers are used to simplify the processing of concepts with multiple parents, improving visualization of the results.

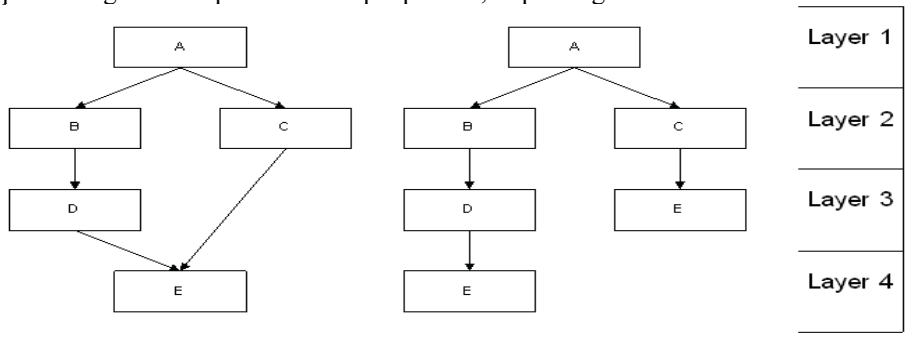


Figure 1. This example illustrates removal of multiple parent links and introduction of layers. Graph on the left side shows the original hierarchy; graph on the right side shows the resulting hierarchy.

After the conversion, concepts with multiple parents appeared as concepts with single parents in each of their parents' hierarchies. In some cases this led to such concepts appearing on different levels. This representation of the concept hierarchies makes it easy to see that the selected SNOMED CT hierarchies have approximately the same height (18 and 12 levels, respectively) as the selected NDF-RT hierarchy (14 levels). This observation was important for our research, as it indicates that all three hierarchies are of the same order of vertical magnitude, an assumption that was subsequently used in our approach.

After extracting the data in the format described above, we were able to analyze the concept names in SNOMED CT's Product and Substance hierarchies and NDF-RT's INGREDIENT_KIND hierarchy. As a result of this analysis, we have collected cases of concept pairs with the same (or close) meaning that do not have a precise match of their names. For example, the concept *Salicylic acid gel (product)* in

SNOMED CT's hierarchy has a matching *Salicylic Acid* concept in NDF-RT's hierarchy. Another example would be *Carbomers (product)* from SNOMED CT, as compared to its matching concept *carbomer* from NDF-RT. In order to overcome this obstacle, we have implemented a partial string match algorithm for the concept names. This partial match algorithm requires that if the name of one of the matched concepts contains the name of the other concept, but is different from it, the smaller of the concept names should constitute at least 75% of the larger concept name. Switching to a partial match algorithm from a precise match algorithm has improved our results by increasing the number of matched concepts by 48 for the Product hierarchy and by 187 for the Substance hierarchy. The end-user is able to review the concepts from the two terminologies that were paired (using partial match) in our software tool, to correct erroneously matched concepts.

As a result of this approach, we were able to efficiently match pairs of concepts having different drug dosages or different forms and we identified a total of 1535 matches for the Product hierarchy and 3294 for the Substance hierarchy with NDF-RT. For each of these concept pairs an automated analysis of the hierarchical paths has been performed in order to detect similarities and discrepancies.

2. Results

We implemented our methodology in a software tool (HierarchyDIFF). Applying our methodology to the SNOMED CT and NDF-RT hierarchies provided us with results (Figure 2) that can be subdivided into three categories. 1) Matching Hierarchies, 2) Missing Concepts, and 3) Additional Granularity of the Classification.

Category 1: Despite having a significant number of similarly named concepts, the hierarchical classification was different for most of them – only 61 Product concepts and 378 Substance concepts have the same parent concepts both in NDF-RT and in SNOMED CT.

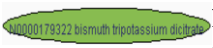
Category 2: Performing a path comparison we observe the following. A significant number of NDF-RT concepts, such as N0000000002, Chemical ingredient ==> N0000175090, Unclassified Ingredients ==> N0000171572, Cetaphil and Batroxobin, compared to Chemical ingredient ==> Enzymes and Coenzymes ==> Enzymes ==> Hydrolases ==> Peptide Hydrolases ==> Endopeptidases ==> Serine Endopeptidases ==> Venombin A ==> Batroxobin are missing in SNOMED CT's hierarchy. In total, we have identified 8583 such concepts for the Product hierarchy and 6824 for the Substance hierarchy and have shown a few of them as green ovals in Figure 2 in a convenient hierarchical way.

Category 3: Although a significant number of the NDF-RT concepts were identified as "missing concepts," for 749 NDF-RT concepts that were missing, we were able to map their hierarchical structure to the hierarchical structure of SNOMED CT Product hierarchy (1076 for the Substance hierarchy), by identifying common concepts and marking them in our visualization. For example, the concept Metyrosine is classified as N0000000002, Chemical ingredient ==> N0000007833, Amino Acids, Peptides, and Proteins ==> N0000006806, Amino Acids ==> N0000007703, Amino Acids, Cyclic ==> N0000011248, Amino Acids, Aromatic ==> N0000006152, Tyrosine ==> N0000007502, Methyltyrosines ==> N0000005770, alpha-Methyltyrosine ==> N0000179717, Metyrosine in NDF-RT and as 373873005, Pharmaceutical biologic

product (product) ==> 14833006, Cardiovascular drug (product) ==> 1182007, Hypotensive agent (product) ==> 86131002, Metyrosine (product) in SNOMED CT's Product hierarchy. In such cases, the concept itself is marked in black, and its NDF-RT hierarchical path is displayed as concepts with cyan-colored, rounded-corner rectangles. If a common parent is identified, it is marked with a red oval. A total of 13 and 16 such common parents and 1308 and 2587 common descendants have been identified in Product and Substance hierarchies correspondently.

LEGEND:

Start of a new hierarchical path. The concept is present in both terminologies, but some descendants are classified differently



NDF-RT Concept is missing in SNOMED. None of its descendants are present in SNOMED, resulting in a new concept/sub-hierarchy being added to SNOMED's hierarchy.

NDF-RT concept is missing in SNOMED. At least one of its descendants is present in SNOMED CT under different classification, allowing building new hierarchical paths to already existing concepts

Concept did not require any hierarchical changes after merge

End of a new hierarchical path. The concept is present in both hierarchies, but has different hierarchical paths

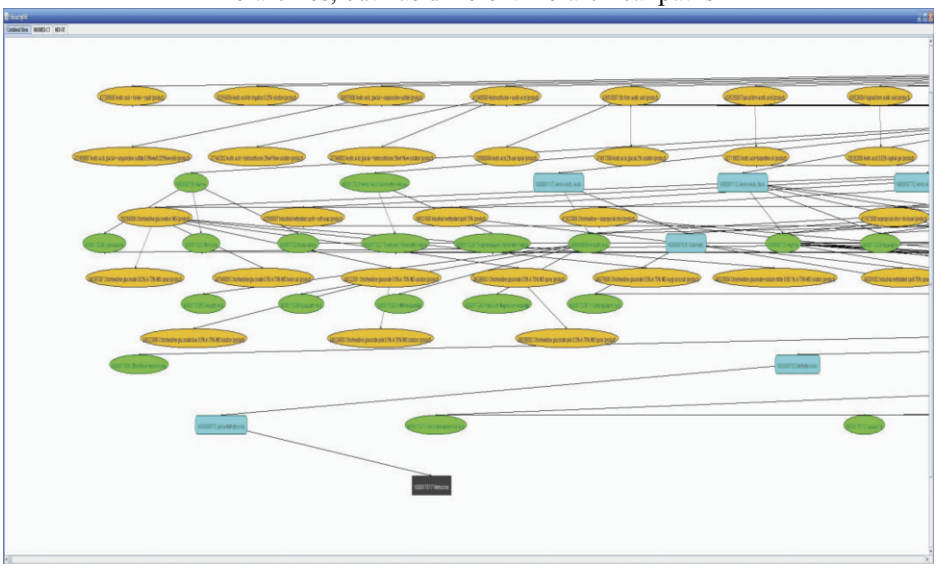


Figure 2. View of the merge of the terminologies with edges enabled. Users are provided with an option of disabling some or all of the edges to avoid overloading the visualization. Above the figure is the Legend.

3. Discussion, Conclusions and Future Work

We explored the possibility of auditing and improving the structure of a medical terminology (SNOMED CT) via the hierarchical structure of another terminology (NDF-RT). The approach presented in this paper goes beyond pointing out the concepts that are uniquely only in one terminology and our HierarchyDIFF tool performs an

analysis of the hierarchical relationships, thus allowing us to audit the two terminologies together in a more efficient way. The technique described here can also be considered as one of the steps that would facilitate the process of merging two terminologies, although the usability of this approach is still to be determined.

This study is an initial exploration of combining several hierarchical drug/ingredient terminologies. As such, it used a simple method of partial match of concepts, based on name similarity. While our approach did yield a positive result, we will attempt to combine the class-based and semantic approaches. In addition, we will also include RxNorm into the method, using an approach described by Fung et al. [12]. This addition would allow us to refine our results and improve our semantic methods, as well as to use more sophisticated and precise methods based on concept classification. These two hypotheses are subject to verification and further research.

The detected hierarchical inconsistencies are subject to review and approval by the organizations and curators managing the corresponding medical terminologies. Their input on whether the proposed changes to the hierarchical structures are valid might differ from what the tool suggests. In addition, further studies might help improve the interface and the functionality of the HierarchyDIFF tool.

In this paper, we analyzed the possibility of combining the hierarchical information of two medical terminologies, NDF-RT and SNOMED CT and used this combination to highlight inconsistencies in classification. We presented a visualization tool to support this task, called HierarchyDIFF. Although we succeeded in discovering and extracting useful information, such as proposed missing concepts, a final determination of the correctness of these results will need to be made by domain experts.

References

[1] IHTSDO SNOMED CT http://www.ihtsdo.org/snomed-ct. Accessed July 27, 2014
[2] Cornet R, de Keizer N. *Forty years of SNOMED: a literature review*. BMC Med Inform Decis Mak, 8 Suppl 1,2008, p.S2 http://www.ncbi.nlm.nih.gov/pubmed/19007439
[3] Wang Y, Halper M, Wei D, Perl Y, Geller J. *Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED*. J Biomed Inform. 2012 Feb;45(1):15-29.
[4] Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, Spackman KA. *Analysis of error concentrations in SNOMED*. AMIA Annu Symp Proc. 2007:314-8.
[5] Wang Y, Halper M, Wei D, Gu H, Perl Y, Xu J, Elhanan G, Chen Y, Spackman KA, Case JT, Hripcsak G. *Auditing complex concepts of SNOMED using a refined hierarchical abstraction network*. J Biomed Inform. 2012 Feb;45(1):1-14.
[6] Geller J, Ochs C, Perl Y, Xu J. *New abstraction networks and a new visualization tool in support of auditing the SNOMED CT content.* AMIA Annu Symp Proc. 2012;2012:237-46.
[7] National Drug File–Reference Terminology(NDF-RT™) Documentation March 2014 Version, U.S. Veterans Health Administration http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT%20Documentation.pdf Accessed July 27, 2014
[8] Zhu Q, Freimuth RR, Pathak J, Durski MJ, Chute CG. *Disambiguation of PharmGKB drug-disease relations with NDF-RT and SPL* J Biomed Inform.2013 Aug;46(4):690-6.
[9] Zhu Q, Jiang G, Chute CG. *Profiling structured product labeling with NDF-RT and RxNorm.* J Biomed Semantics. 2012 Dec 20;3(1):16.
[10] Mortensen J, Bodenreider O. *Comparing Pharmacologic Classes in NDF-RT and SNOMED CT*, Available at http://mor.nlm.nih.gov/pubs/pdf/2010-smbm-jm.pdf
[11] McCoy JA, McCoy AB, Wright A, Sittig F. *Automated Inference of Patient Problems from Medications using NDF-RT and the SNOMED-CT CORE Problem List Subset.* AMIA Poster 2011
[12] Fung KW, Jao CS, Demner-Fushman D. *Extracting drug indication information from structured product labels using national language processing.* J Am Med Inform Assoc 2013;20:482-488.