Preliminary results using Snorocket to detect errors in the post-coordination of SNOMED CT

Pilar Avila[1], Laura Gambarte[1], Hernán Berinsky[1], Sonia Benítez[1], Hee Park[1], Daniel Luna[1], David Pérez[1], Fernando Campos[1], Sofía Zanetti[1], Soledad Díaz[1], and Alejandro López[2],

[1] Departamento de Informática en Salud, Hospital Italiano de Buenos Aires
mariad.avila@hospitalitaliano.org.ar
[2] TermMed, Buenos Aires
alejandro.lopezosornio@gmail.com

Abstract. We present preliminary results for the application of a procedure that detects and corrects errors in concept definitions of a local interface vocabulary with SNOMED CT as its reference vocabulary. Using the relations inferred by SNOROCKET we detected redundant fully defined concepts, but also we detected suspected patterns where concepts had redundant inferred relations. Our procedure detected errors in 1.63% of the whole vocabulary, the primary type of error was produced by duplications since these concepts did not exist when the knowledge modeler asserted them. Using these results, we implemented a GUI to track patterns and correct errors. Our procedure contributes to the quality assurance of our local interface vocabulary since errors in the hierarchies can compromise interoperability and meaningful use of the vocabulary. Our approach could be used by thesaurus implementers to detect suspected patterns, grouping them, and offer a centralized interface to correct them.

Keywords: Classification; Controlled Vocabulary; Systematized Nomenclature of Medicine; Quality Control; Knowledge representation; User-Computer Interface

1 Introduction

SNOMED CT is a clinical terminology that allows storage and retrieves healthcare information based on its meaning[1], information definitions are build using their semantic relations. When a single concept is not enough to define the information is possible to build a new one using post-coordination, understood as the representation of a clinical meaning using a combination of two or more SNOMED CT concept identifiers[2]. The Hospital Italiano de Buenos Aires (HIBA) has a local interface vocabulary where each term is mapped via a direct link or a post-coordinated expression to SNOMED CT as its reference vocabulary. The local interface vocabulary was implemented from 2002, and by 2016 already had 520,000 post-coordinated concepts in its terminology system, this system gives services to the

Hospital Italiano Healthcare facilities as well as other health care institutions in Argentina, Uruguay, and Chile.

A major benefit of the local interface vocabulary is its big size and coverage but also is the biggest obstacle for its use and maintenance. It's a requirement for a terminology system that the represented knowledge should be faithful to reality and its quality must be assured, especially in concepts that are post-coordinated for other concepts because the information is propagated whether is right or wrong, therefore a single relation could have a broad impact and unintentional effects[3]. In order to facilitate the inconsistency detection, classifiers are used to obtain a formal representation that provides the explicit semantics of the represented knowledge[4].

A description logic classifier is used here to process all SNOMED CT concepts and HIBA post-coordinated expressions, and based on their definitions organizes the concepts into hierarchies[1]. The results are used to detect redundant concepts and two types of suspicious error definitions. Between the available classifiers, SNOROCKET was selected, since it was specifically designed to be efficient with the logic descriptor used by SNOMED CT[5].

This article analyzes preliminary results for applying a procedure to detect errors in the definitions of the concepts through their inferred form, and to correct them in an iterative cycle of quality assurance. The large size of the local interface vocabulary require a combination of automated and manual checks and reviews, our contribution is a computational approach that helps the users(knowledge modelers) with the audit task in errors automatically found. This approach could be extended to identify more patterns and be presented to users in a consistent GUI, as well as to be applied in others thesaurus with formal definitions.

2 Background

SNOMED CT hierarchies are formulated using a subset of first-order logic known as "description logics" that specifies their semantics. SNOMED CT hierarchies are comprehensive and universal. All and only concepts satisfying the definition of a higher-level "ancestor", concepts are classified under it as "descendants", and all the definitions that apply to a concept also apply to all of its descendants.[3]

When new definitions are created, they must be formulated in a description logic in two steps[3]:

- Modelers assert manually each new concept using SNOMED CT as reference vocabulary, creating a "stated form" with defining relations to SNOMED CT concepts.
- A "classifier" software organizes the concepts logically into hierarchies based on their stated definitions; the result is an "inferred form" of the concepts.

Quality assurance is an important part of a terminology's lifecycle, SNOMED CT has been specially studied because of its importance. Many techniques have been proposed for identifying errors[6], these could be summarized in 3 groups: (a) evaluating the semantic completeness[7], (b) from an ontological and logical perspective[8-10], (c) and using Description Logic modeling and concept classification[1,3]. We will rely on the latter approach in this paper.

An important feature of description logics is they allow concepts to be either fully or partially defined. To be fully defined means to be defined by necessary and sufficient conditions, in the case of partially defined, called “primitive” by SNOMED CT, is only defined by necessary conditions[1]. Consequently, only with fully defined concepts is possible to find redundant concepts using a classifier, this redundancy occurs because the inferred form of the post-coordinated concepts is equivalent or because the concept already exists in the reference vocabulary and it should be pre-coordinated instead of post-coordinated.

Using the inferred form is possible to find another type of errors, SNOMED CT publishes descriptive statistics in a draft[11] and implements in its GitHub repository [12]the detection of 22 patterns that could be considered as suspected error definitions. In the beta version of this procedure is detected a general pattern as suspected definition that happens when the inferred form has at the same time two relations to concepts where one is subsumed by the other. In a correct classifier execution, the most general relation should be deleted, and only the most specific should appear. These errors were classified into two types:

- By group interchanges occurs between general and specific relations. For example, figure 1 shows a graphical comparison of the stated and inferred form for “Right main coronary artery thrombosis”. Here the finding site “Right coronary artery structure” is in the same group with the associated morphology “Thrombus”. There is also another group with a more general finding site “Coronary artery structure” but a more specific associated morphology “Occlusive thrombus”. This type of error is called crossover.
- When is inferred a more specific attribute relation. In the case of the example of “Allergy to antiseptic agent” that is shown in the figure 2, the inferred form has a causal agent more specific “Antibacterial agent” that the one asserted in the stated form “Anti-infective agent”. There is a suspected error pattern because the attribute relation was inherited from a concept that has the most specific definition.

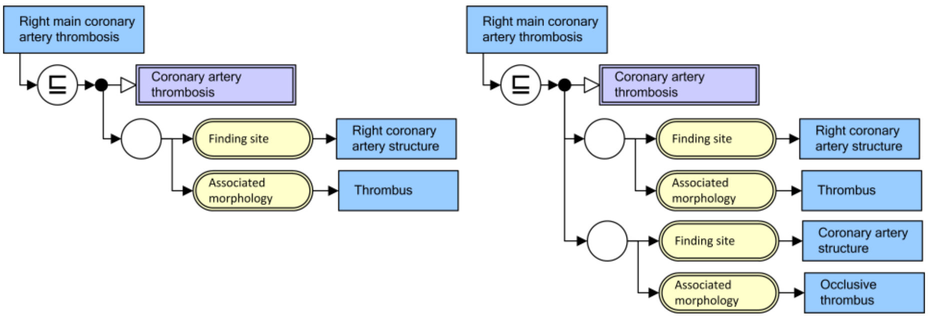


Fig. 1. Stated (left) and inferred (right) model comparison of the concept “Right main coronary artery thrombosis”.

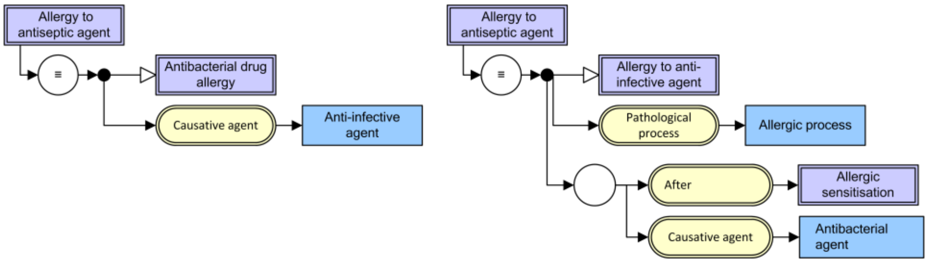


Fig. 2. Stated (left) and inferred (right) model comparison of the concept "Allergy to antiseptic agent"

3 Materials and methods

This study uses the SNOMED CT Spanish edition available on 31 October 2015. This edition has 310,000 concepts defined by 1530,000 relations, and the HIBA's local interface vocabulary that has 520,000 post-coordinated concepts defined by 830,000 relations.

The classification task is made using SNOROCKET, which is the SNOMED CT's preferred classifier[3,13] This classifier returns two outputs: (1) equivalent concepts, to help to deduct which concepts should be pre-coordinated and which ones should be merged into a single one; (2) inferred form, useful to detect suspected error patterns in the definitions.

The implemented method is schematized in the figure 3, and consisted of a cycle of detection and correction error for achieving quality assurance of the local interface vocabulary of HIBA:

1. Modelers define the concept's stated form.
2. SNOROCKET is executed. The output is used to detect suspected error definitions and redundant concepts.
3. Modelers identify which concepts should be merged and which ones should be redefined.
4. Modelers manually correct detected errors in the concept's stated form.

A platform built "in-house"' with a GUI was developed to do the last two steps of the algorithm, here modelers can track down errors to the source and merge the concepts keeping the referential integrity in the terminology system.

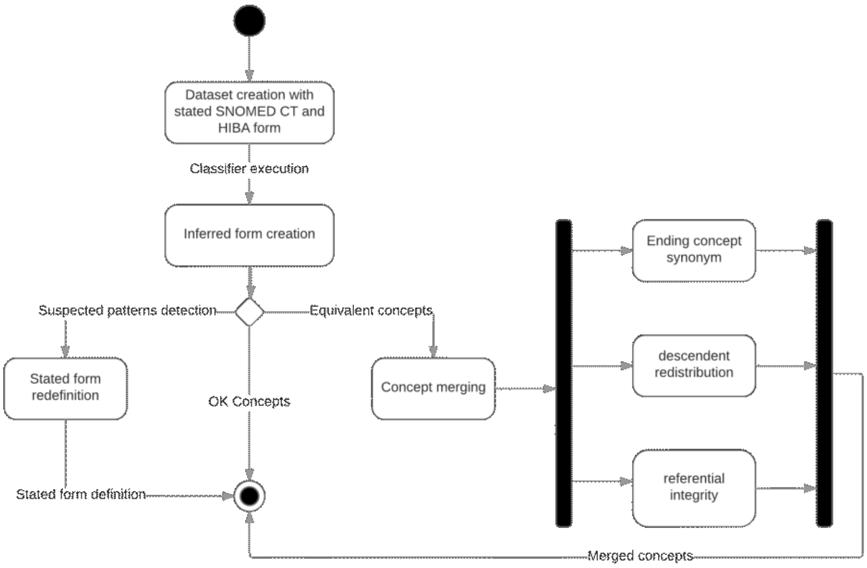


Fig. 3. Implementation procedure

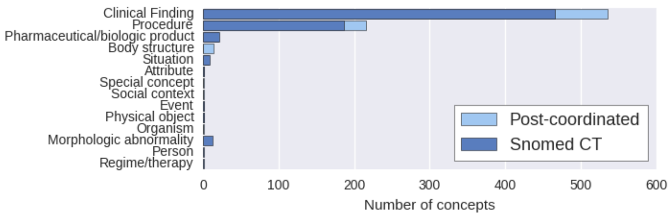
4 Result analysis

After the classifier was run, suspicious patterns were detected in 8,532 HIBA's concepts, representing 1.63% of the whole local interface vocabulary. Of these patterns, 85.62% were redundant concepts, 12.31% were crossover type, and the remaining 2.07% were patterns where the inferred form has relations of type attribute more specific than the stated form.

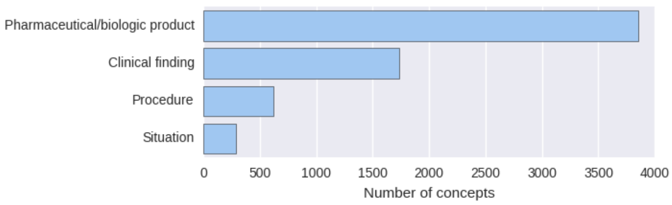
As it can be seen in the figure 4, in both cases the redundant concepts focus on clinical finding, procedure, and pharmaceutical/biologic product hierarchies. These three hierarchies are the 88.82% of the total of all the post-coordinated concepts. There were two types of redundant concepts:

1. Concepts that should be pre-coordinated since already exists a single SNOMED CT's concept that defines the HIBA's concept entirely.
2. Post-coordinated concepts that were equivalent and must be merged or redefined.

Only in the fully defined concepts is possible to detect redundancies, and must have been corrected manually by a knowledge modeler. In the case of primitive concepts, it was used two criteria to detect patterns with suspected errors. The pattern called crossover were focused on concepts from the clinical finding hierarchy; and the concepts of the pattern where the inferred form had relations of type attribute more specific than the stated form, were uniformly distributed in the three major hierarchies previously mentioned, as it is shown in the figure 5.

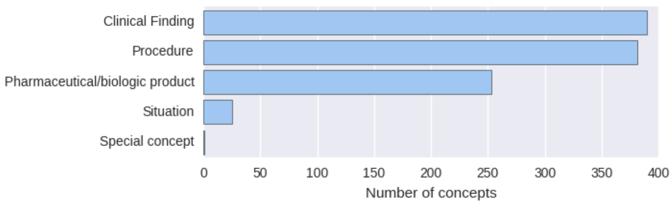


(a) Concepts that must be pre-coordinated

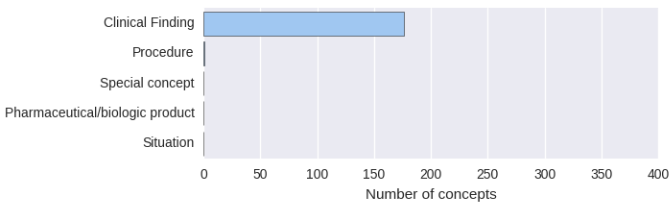


(b) Concepts that must be merged

Fig. 4. Redundant concepts



(a) Concept with inferred form that had relations of type attribute more specific than the stated form



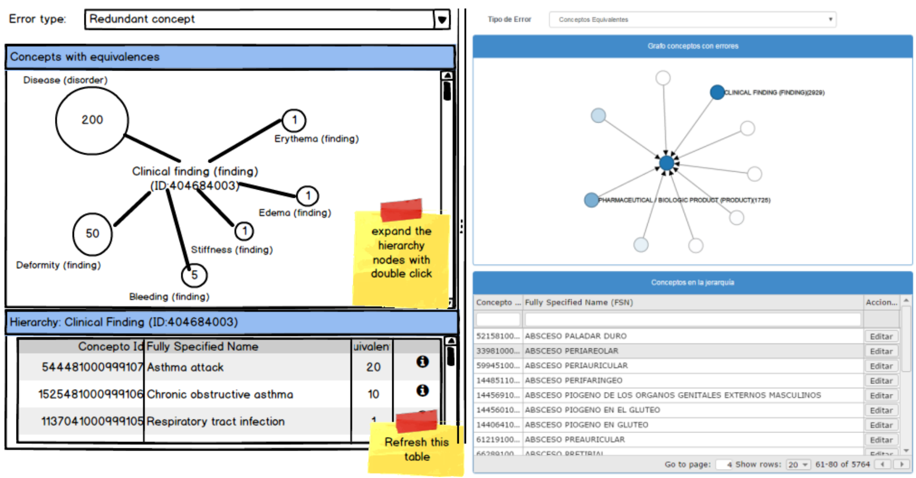
(b) Concepts with crossover

Fig. 5. Concepts with suspected error patterns

From the results obtained with SNOROCKET, we designed software that allows correcting errors following top-down and bottom-up strategies. The top-down strategy allows getting all the auditable concepts from an ancestor. In this way, for example, pharmacology, procedure or clinical finding experts can retrieve the concepts they want to correct or navigate to more specific hierarchies. Figure 6 shows how to navigate through the hierarchies and filter the concepts using the GUI. Afterward, when is selected a concept, a modeler can execute the actions to amend the errors.

With a bottom-up strategy, modelers start in the concept that must be corrected. The tool has three types of views, as it is shown in figure 7, stated, inferred and graph view. In the stated view, the concepts are defined using relation that was assorted manually by a modeler; the inferred view is the SNOROCKET's output after the concepts were classified base on their logical definitions; graph view is used to understand the inferences source. In the graph, nodes represent SNOMED CT concepts, and edges represent relations, modelers can expand the nodes to get their ancestors and track the source of unexpected inferences.

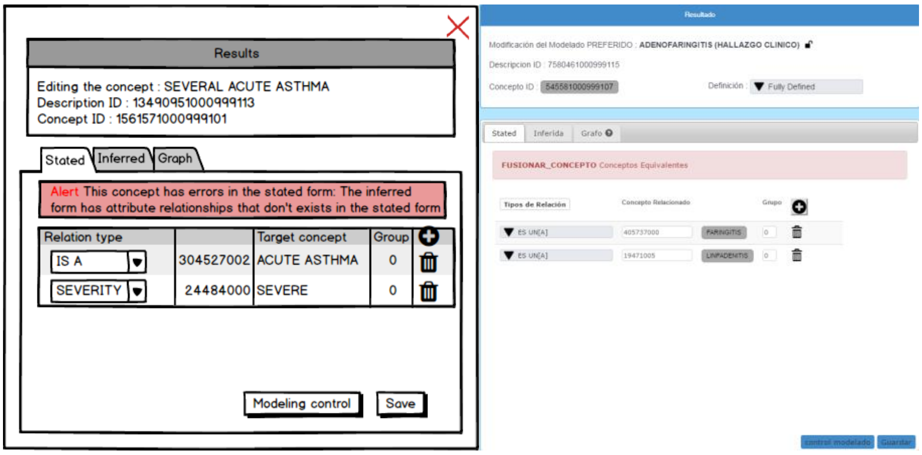
Two types of corrections can be made: 1) to change the definition in stated view 2) to merge concepts when equivalence with others concepts exists. This last case requires defining the new concept's synonym and keeping the referential integrity in the terminology system.



(a) Hierarchy navigator

Fig. 6. GUI to filter concepts with suspected error patterns, mockups(left) and implementation(right)[1]

[1] See all the figures at https://goo.gl/photos/zMN87ckguYzktz8Q7



(a) Stated view

Fig. 7. GUI with different views from a concept that must be corrected, mockups(left) and implementation(right)[2]

5 Discussion

In this paper, we described a procedure to detect and correct errors in the definitions of the concepts from the local interface vocabulary of HIBA, after a classifier was applied. The classifier returns an inferred form for each concept and the equivalent concepts. From this output, software was designed to group concepts with suspected patterns in a centralized interface, that modelers can merge concepts or track and correct the error in the stated view.

In previous studies, using SNOMED CT Editorial Guide directives showed that 3.67% of the most used local interface vocabulary concepts at HIBA were not represented correctly[14]. Using the relations inferred by a classifier, the procedure detected 8,532 errors in the post-coordinated concepts, this is 1.63% of the whole vocabulary.

Detected errors were limited to the redundancies found by SNOROCKET and the patterns that are suspected for errors mentioned for SNOMED CT's draft[11] and GitHub[12]. In this procedure, we implemented two types of patterns that had the biggest probability of being errors, crossover type and when is inferred a more specific attribute relation type. As a consequence of the SNOMED CT updating, the most common errors were produced by redundancies since these concepts did not exist when were asserted by the modeler.

This procedure described in this paper was designed as a part of an integral cycle of quality assurance, and the GUI were addressed to ease the track of error and its source. To achieve this goal, the suspected patterns were divided by types and then for the concept was suggested the action that removes the detected problem.

[2] See all the figures at https://goo.gl/photos/jmqMuP6aLvtUE9bQA

Although the procedure is used in a local interface vocabulary in the field of medicine, it is applicable in another domain thesaurus with formal definitions as well.

6 Conclusion and future work

This paper shows 3 types of errors that can be detected in the definitions of the concepts from a local interface vocabulary using the relations inferred by a classifier. Even though these redundancies and errors unnecessarily impede the work of modelers, the manual search for the causes of the errors can be a tedious task[15]. On the other hand, keeping an interface vocabulary with these errors can propitiate the spread of errors, compromise the interoperability and meaningful use[3]. For example, when the concept with errors was in the internal nodes into the hierarchy the errors are propagated in cascade; or when artificial intelligence is used to automatically make the assertions can compromise the new assertions since the manual assertions are used as input to create models.

The software created to help modelers to correct errors is in its beta version and has not been extensively tested. In further developments, the software is going to be tested using usability metrics to ease the tracking of error in the stated form keeping a design oriented to minimize errors and their impact. As this approach is scalable to be implemented with other types of errors, we are going to implement the detection of all the suspected definitions described by SNOMED CT and measure their recall and precision.

Overall, it can be concluded that the application of this procedure contribute to pointing out where the errors happen and offer a centralized interface to correct them. It's a step to improve the quality of our local interface vocabulary.

References

1. Alan Rector and Luigi Iannone. Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. *Journal of biomedical informatics*, 45(2):199–209, April 2012.
2. T. Benson. *Principles of Health Interoperability HL7 and SNOMED*. Health informatics. Springer London, 2009.
3. Alan L Rector, Sam Brandt, and Thomas Schneider. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *Journal of the American Medical Informatics Association : JAMIA*, 18(4):432–40, January.
4. Ronald Cornet and Ameen Abu-Hanna. Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions. *International Journal of Medical Informatics*, 77(5):336–345, 2008.
5. Michael Lawley Ro Metke-jimenez. Snorocket 2.0: Concrete Domains and Concurrent Classification.
6. J E Rogers. Quality assurance of medical ontologies. *Methods of information in medicine*, 45(3):267–74, 2006.

7. Guoqian Jiang and Christopher G Chute. Auditing the semantic completeness of SNOMED CT using formal concept analysis. *Journal of the American Medical Informatics Association : JAMIA*, 16(1):89–102.
8. Stefan Schulz, Boontawee Suntisrivaraporn, Franz Baader, and Martin Boeker. SNOMED reaching its adolescence: ontologists' and logicians' health check. *International journal of medical informatics*, 78 Suppl 1:S86–94, April 2009.
9. S Schulz, S Hanser, U Hahn, and J Rogers. The semantics of procedures and diseases in SNOMED CT. *Methods of information in medicine*, 45(4):354–8, 2006.
10. Stefan Schulz, Udo Hahn, and Jeremy Rogers. Semantic Clarification of the Representation of Procedures and Diseases in SNOMED((R))CT. *Studies in health technology and informatics*, 116:773–8, 2005.
11. International Health Terminology Standards Development Organisation. Snomed CT Quality Evaluation. `http://qa.snomed.org/docs/SNOMEDCT_Quality_Evaluation_Descriptive_Release_Statistics_20160219_v1_00.pdf`, 2016. [Online; accessed 07-July-2016].
12. International Health Terminology Standards Development Organisation. SNOMED CT Pattern and Changes Evaluation Toolkit. `https://github.com/IHTSDO/sct-statistics-qa/`, 2016. [Online; accessed 07-July-2016].
13. Michael J Lawley and Cyril Bousquet. Fast classification in protégé: Snorocket as an owl 2 el reasoner. *Proc. 6th Australasian Ontology Workshop (IAOA'10). Conferences in Research and Practice in Information Technology*, 122:45–49, 2010.
14. H Navas, A Lopez Osornio, G Elias Leguizamon, N Orrego, SWassermann, D Luna, et al. *Implementación de reglas para el control de modelado con SNOMED CT*. INFOLAC, 2008.
15. Kathrin Dentler and Ronald Cornet. Redundant Elements in SNOMED CT Concept Definitions. pages 186–195. Springer Berlin Heidelberg, 2013.