An Analysis of SNOMED-CT Using Structural Symmetry and Lexical Matching

Emilio R. Gonzalez
Computer Information Systems, Undergraduate LSAMP Research Program

CUNY – Borough of Manhattan Community College
New York, New York

National Science Foundation
Arlington, Virginia

December 31, 2015

Prepared in fulfillment of the 2015 Fall Progress Report for CUNY's LSAMP program under direction of Yan Chen in the Computer Information Systems department at City University of New York's Borough of Manhattan Community College

Participant: Emilio R. Gonzalez

____

Research Mentor: Yan Chen

____

TABLE OF CONTENTS

ABSTRACT

An Analysis of SNOMED-CT Using Structural Symmetry and Lexical Matching.

EMILIO R. GONZALEZ[1], YAN CHEN[1]
[1] CUNY: Borough of Manhattan College, New York, NY 10007

Two concepts, C1 and C2, are symmetric if concept C1 can be obtained from concept C2 by replacing a single modifier such as “left” with “right” or “upper” with “lower”. For example, concepts Active upper limb movements and Active lower limb movements are symmetric, since changing “upper” from the former concept with “lower” resulted in the latter concept. Likewise, concepts anterior perineal hernia and posterior perineal hernia are also symmetric. We consider these two groups of concepts as consistent, as each group of concepts share the same parent concept. However, we find that the concept pneumonia has a child congenital pneumonia, but there is no concept named as acquired pneumonia. We consider this as an inconsistency. [1]

We have selected the following ten modifier pairs for our symmetric concepts study: (acute, chronic), (acquired, congenital), (primary, secondary), (unilateral, bilateral), (first, second), (upper, lower), (superior, inferior), (left, right), (anterior, posterior), (ascending, descending).

1. INTRODUCTION

The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is the most comprehensive, multilingual clinical healthcare terminology in the world. [5] It contains more than 311,000 active concepts with unique meanings and formal logic-based definitions organized into 19 hierarchies. [7] SNOMED CT is currently used in 27 countries and 5000 governments, hospitals and organizations. [4] SNOMED CT provides for consistent information interexchange in clinical healthcare and is essential to an interoperable electronic health records. It also reduces the variability of data presentation in clinical healthcare practice and research. [2]

SNOMED CT is updated biannually and has oversight from the International Health Terminology Standards Development Organisation (IHTSDO). Each release involves a significant amount of manual editing by a team of Terminologists. Due to SNOMED's broad scope and inherent complexity, it is inevitable that inconsistency and errors will find their way into SNOMED's knowledge content. Quality assurance is a key component of the terminology development and maintenance, the same way as software development.

SNOMED CT has proven to be beneficial in a variety of ways including: Enhancement of overall patient care, help with monitoring health problems and trends worldwide, cost-effectiveness due to elimination of redundancy testing, narrowing down patients that are eligible for vaccines and cures and expanding knowledge of diseases, treatments and outcomes. [6]

2. MATERIALS AND METHODS

i. SNOMED-CT Conceptual Model

SNOMED-CT's components have a fairly simple data-structure, where a concept has a Description, Concept Identifier and a Relationship. (*see Figure 1)* Concepts are clinical meanings that are organized into hierarchies. Concepts include a concept identifier, which is a string of digits 8-16 characters in size. Concept descriptions link human readable terms to the terminology and usually include one preferred language referred to as a Fully Specified Name, or FSN, and one or many Synonyms, which specialize the concept. (*see Figure 2*) Relationships link concepts to related concepts. Each concept must have at least one |is-a| relationship and can include many attribute relationships to specialize the concept. (*see Figure 3*) [6, P.14-17]

Due to SNOMED's broad scope and inherent complexity, it is inevitable that inconsistency and errors will find their way into SNOMED's knowledge content. The IHTSDO ensures this by testing their knowledgebase and looking for inconsistencies – a.k.a. quality assurance. Terminologists work manually to find these common errors, which include and are not limited to: concepts that lack synonyms, concepts with no synonyms, concepts with wrong or inaccurate parents, concepts with no parents, concepts with wrong or inaccurate children, concepts with no children, duplicate concept entries and missing or inaccurate relationships between concepts. See Figure 4 for examples of consistencies and inconsistencies.

There are many ways to find and sniff out these inconsistencies. Programmers and terminologists create their own programs to accomplish this task using methods such as: lexical matching, whole word matching, Prefix matching, Snippet matching, Concept Code matching, Symmetric matching, Sentence matching, Suffix matching, extension matching, and sub-set matching.

ii. Theoretical Model

In 2012, the Structural Informatics Group at the University of Washington, Seattle, analyzed the FMA ontology using structural symmetry and they concluded that this approach is a "valuable method to ensure quality assurance". [3] Using this conclusion, we decided to use structural symmetry to audit SNOMED CT. The difference between my work and their work is the fact that my code will be used exclusively for SNOMED CT and will be more efficient, highly customizable and very adaptive to other languages.

C = Starting Concept – This can be any SNOMED CT hierarchy/sub-hierarchy.

```
IdentifyInconsistency(C)
     {
          for each child of concept CH of C
               if CH contains modifier
                    Let CH_r be a string with the modifier
                    replaced by its symmetric one in the
                    modifier pair.
               For each sibling CH_s of CH
               If CH_s is the same as CH_r
                    CH is a consistent case.
               If no CH_s is the same as CH_r
                    report CH as an inconsistent case.
               IdentifyInconsistency(CH)
     }
```

This algorithm can be applied to any SNOMED CT hierarchy. If the whole hierarchy needs to be audited, then the algorithm can be called from the root concept, for example, the Clinical Finding hierarchy. It can also be applied to a specific area that a researcher is interested in, for example, the sub-hierarchy rooted at "bleeding". In this study, we will audit the consistency of SNOMED CT by automatically generating symmetric concept pairs by using linguistic structure in the concepts names.

The algorithm I have developed is highly customizable and very proactive with the ability to be modified to serve other purposes such as: use with any language (for SNOMED INTL releases,) adding, removing and switching modifier pairs, search for missing or inaccurate parents and/or children, report non-existent concepts that should exist, report missing or inaccurate relationships, report missing or inaccurate descriptions, report duplicate concept entries and even report misspelled concepts using custom threshold rates defined by the user.

iii. Apparatus

To make this algorithm work we will be using a computer workstation that will run Netbeans Java Integrated Development Environment (IDE) and SnAPI Application Program Interface (API) from Dataline's Snoflake. We will also be utilizing the IHTSDO's SNOMED CT international release files along with SnAPI's programmer's manual that describes all SnAPI methods. Using Netbeans IDE we will be able to connect to Snoflake's SNOMED-CT database API and run specified commands and programs such as my algorithm. To make my program run SnAPI, a GUID license code will be acquired from Dataline's Snoflake.

iv. Component Methods

Several Java methods will be created to facilitate quick data retrieval and comparisons, these methods include: a main program to run all methods and scan all concepts and sub-concepts using recursion, a method to get a concept's FSN, a method to determine if a child concept contains a modifier, a concept to determine a child concepts modifier pair, a method to search for a concept string using a lexical approach, a method to get a concepts children, a method to get a concepts parent, a method to report any errors and inconsistencies and a final method that compiles all statistical data into a single file. Other methods include core SnAPI components that are needed to connect to certain methods within the API infrastructure.

v. Statistics and Error Reporting

I have added in a highly efficient statistical reporting system along with an error log that details the type of error, how this error occurs and where this error can be found including all the information needed to correct it. In this study I have made 5 basic errors with detailed statistics and reporting. These errors are: error 0 - No information for this error, error 1- modifier pair does not exist, error 2 - Parent mismatch, error 3 - modifier pair does not exist or might be mistyped (TH>=85.00%) and error 4 - parent for one or more concepts do not exist.

3. RESULTS

i. Starting Concept Selection

For the first phase of experimentation I have chosen only one relatively small sub-hierarchy that has ideal conditions to test my program with. I chose to audit “Disorder of Skull” which is a sub-hierarchy of the main hierarchy “Disorder” and has the following SNOMED concept id: 118945006. This audit would also allow me to debug any errors within my program and find other forms of making this program more efficient. This entire audit was written into three reporting logs: a general log that records every event, an error log that compiles every piece of data from any given error and a final log which compiles all statistical breakdown and analysis.

ii. Consistency Audit

A total of 732 concepts were scanned in this audit over a total run time of 55 minutes and 18 seconds. On average, about 13 concepts were scanned per minute – this can vary from machine to machine depending on system specs and machine workload. A total of 92 concepts were discovered to contain modifiers, 89 of those being inconsistent and reported to an error log which left 643 remaining concepts consistent. In other words, this audit yielded an 87.84 percent consistency rate with an inconsistency error rate of 12.16 percent.

iii. Error Sources and Breakdown

Out of 89 errors, the majority was caused by child concept’s parent mismatch (57 concepts fell into this category – error 2). Another significant source of error were child concepts that contained a modifier but its concept pair did not exist (29 concepts fell into this category – error 1). Three concepts fell in the 85.00% typo threshold I set and might be mistyped or inaccurate (error 3). Other errors such as parents do not exist (error 4) and lack of significant data (error 0) were not found in this audit.

4. DISCUSSION AND CONCLUSIONS

Auditing of SNOMED CT hierarchies using structural symmetry proves to be quite efficient and effective based on the data collected from the first experimental test audit. We should expect many other hierarchies and sub-hierarchies to yield equal or higher inconsistency rates especially if we adjust the accuracy threshold I have set. All in all we believe this is a valuable approach to quality assurance.

5. FUTURE WORK

Moving forward with this program will require approval from the Dataline to expand our license to allow more requests to the terminology server. Currently we are limited to the amount of concepts we can audit due to the restrictions set on Dataline by the IHTSDO regarding their API and server requests.

6. ACKNOWLEDGEMENTS

This research was supported by the Louis Stokes Alliance for Minority Participation at City University of New York and the National Science Foundation. I would like to thank my mentor Yan Chen for introducing me to this opportunity, Borough of Manhattan Community College’s LSAMP coordinator Nia Rene and CUNY’s LSAMP director Dr. Brathwaite. I would also like to extend a thank you to Focrun Nahar and Isabella Fante for their assistance.

7. REFERENCES

[1] Bodenreider, Olivier, Anita Burgun, and Thomas Rindflesch. "Assessing the Consistency of a Biomedical Terminology through Lexical Knowledge." PubMed. National Library of Medicine,, 4 Dec. 2002. Web. 11 Nov. 2015. <http://www.ncbi.nlm.nih.gov/pubmed/12460634>.

[2] Fenton, Susan, Kathy Giannangelo, Crystal Kallem, and Rita Scichilone. "Data Standards, Data Quality, and Interoperability (Updated)." American Health Information Management Association, 2013. Web. <http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_050482.hcsp?dDocName=bok1_050482>.

[3] Luo, Lingyun, Jose L.V. Mejino JR, and Guo-Qiang Zhang. "An Analysis of FMA Using Structural Self-Bisimilarity." PubMed. The National Center for Biotechnology, 2 Apr. 2013. Web. 11 Nov. 2015. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3690136/>.

[4] "IHTSDO Members." The International Health Terminology Standards Development Organisation, 2015. Web. <http://www.ihtsdo.org/members/>.

[5] "SNOMED Clinical Terms (SNOMED CT)." Unified Medical Language System. U.S. National Library of Medicine, 2007. Web. 11 Nov. 2015. <https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html>.

[6] "SNOMED CT Starter Guide." International Health Terminology Standards Development Organisation, 1 Feb. 2014. Web. <http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_StarterGuide_Current-en-US_INT_20140222.pdf>.

[7] "Summary of SNOMED CT Benefits." International Health Terminology Standards Development Organisation. Web. <http://ihtsdo.org/index.php?id=853>.

[8] "What Is SNOMED-CT." International Health Terminology Standards Development Organisation (IHTSDO), 2004. Web. 11 Nov. 2015. <http://www.ihtsdo.org/snomed-ct/what-is-snomed-ct>.

FIGURES

Figure 1 – SNOMED CT Conceptual Model

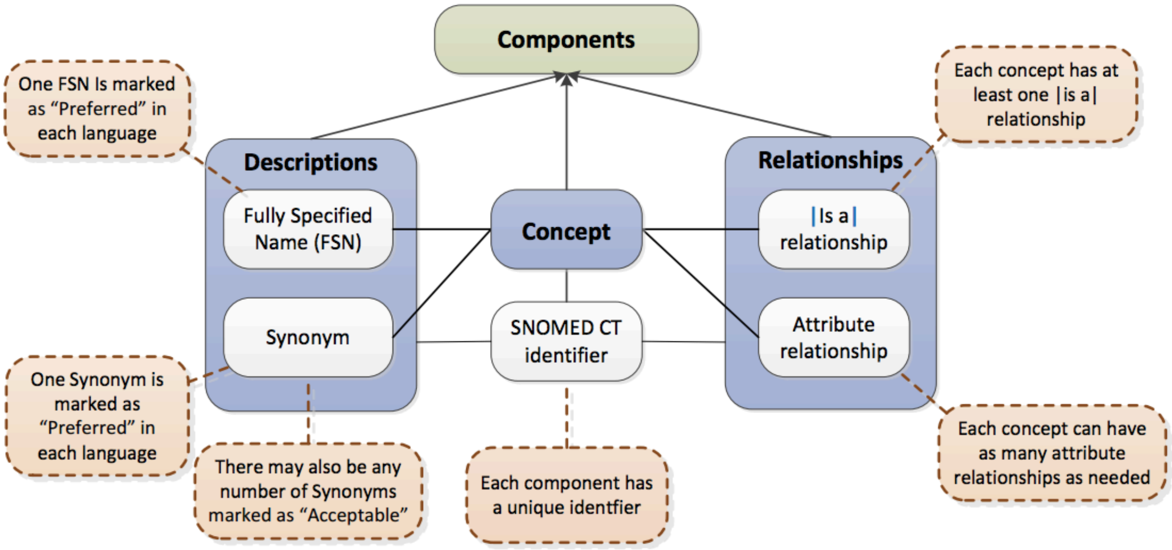

Components
One FSN Is marked as "Preferred" in each language
Each concept has at least one |is a| relationship
Descriptions
Fully Specified Name (FSN)
Synonym
Concept
SNOMED CT identifier
Relationships
|Is a| relationship
Attribute relationship
One Synonym is marked as "Preferred" in each language
There may also be any number of Synonyms marked as "Acceptable"
Each component has a unique identfier
Each concept can have as many attribute relationships as needed

Figure 2 – SNOMED Concept "Myocardial Infarction" w/ CTID w/out relationships

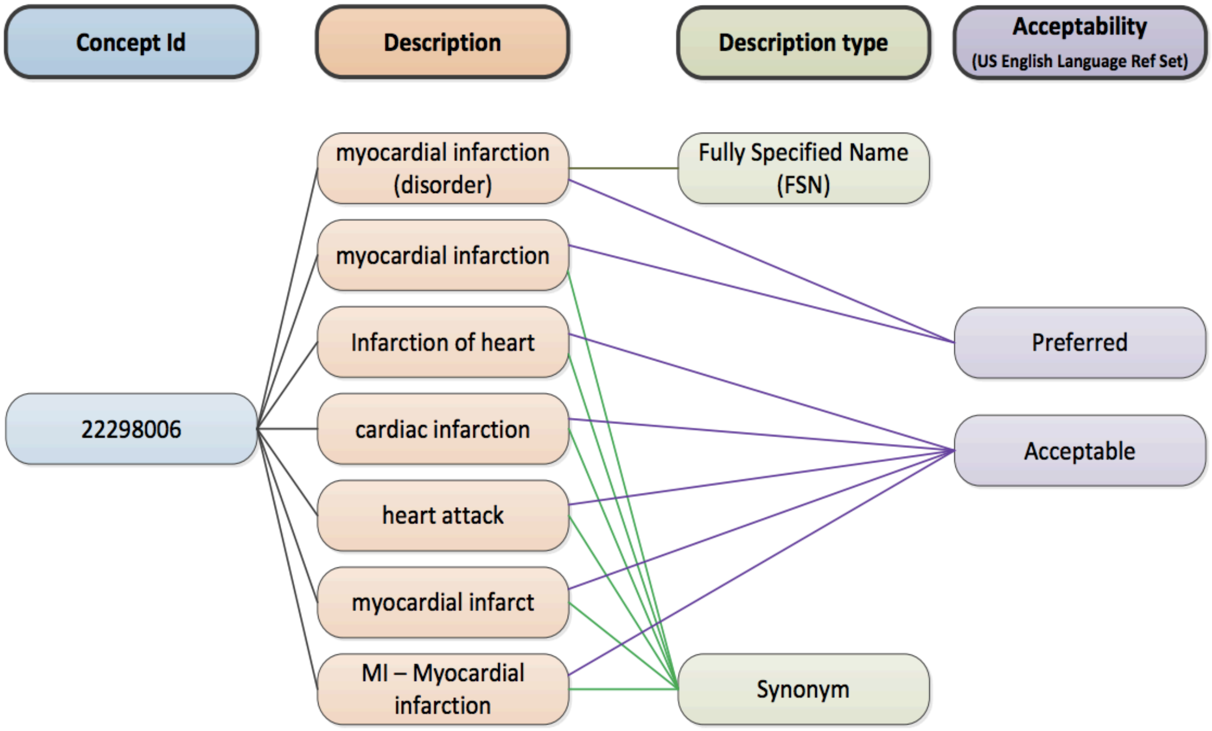
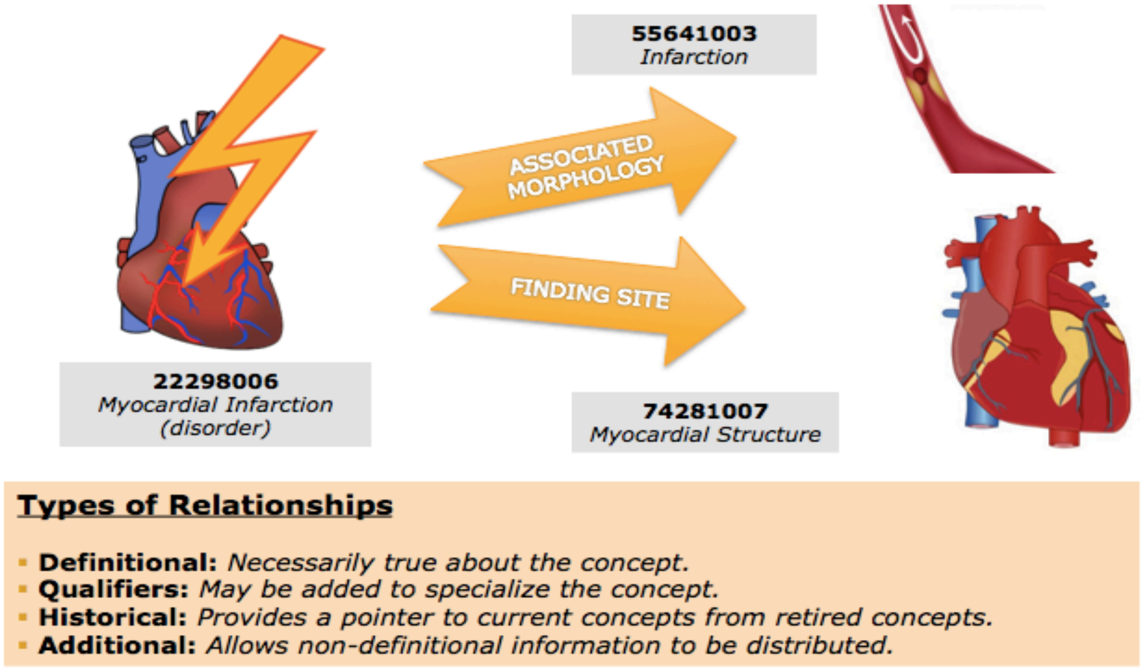


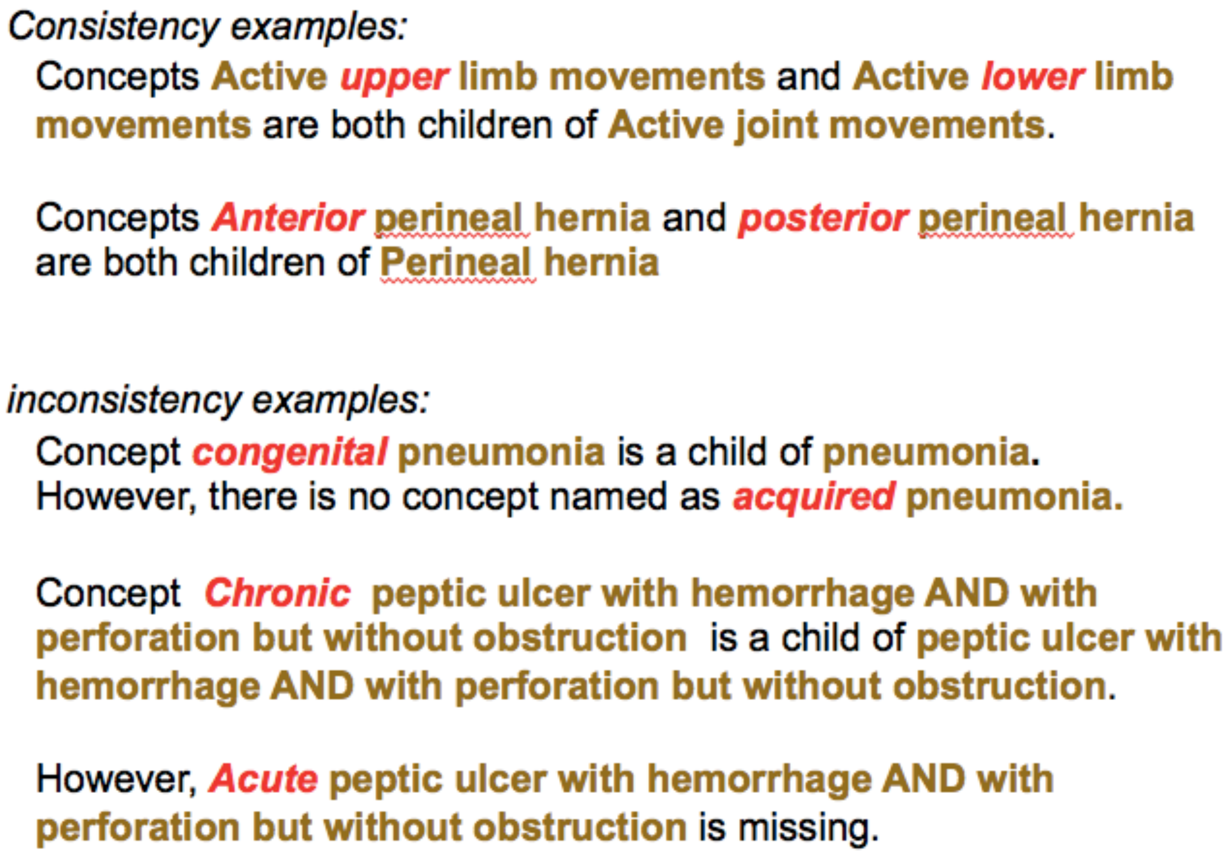
Figure 3 – 2 examples of SNOMED Concept Relationship Model



Types of Relationships

- **Definitional:** *Necessarily true about the concept.*
- **Qualifiers:** *May be added to specialize the concept.*
- **Historical:** *Provides a pointer to current concepts from retired concepts.*
- **Additional:** *Allows non-definitional information to be distributed.*

Figure 4 – Examples of Consistencies & Inconsistencies



Consistency examples:

Concepts **Active *upper* limb movements** and **Active *lower* limb movements** are both children of **Active joint movements**.

Concepts ***Anterior* perineal hernia** and ***posterior* perineal hernia** are both children of **Perineal hernia**

inconsistency examples:

Concept ***congenital* pneumonia** is a child of **pneumonia**. However, there is no concept named as ***acquired* pneumonia.**

Concept ***Chronic* peptic ulcer with hemorrhage AND with perforation but without obstruction** is a child of **peptic ulcer with hemorrhage AND with perforation but without obstruction**.

However, ***Acute* peptic ulcer with hemorrhage AND with perforation but without obstruction** is missing.