ABSTRACT

METHODS FOR EXTENDING BIOMEDICAL REFERENCE ONTOLOGIES AND INTERFACE TERMINOLOGIES FOR EHR TEXT ANNOTATION

by
Vipina Kuttichi Keloth

Biomedical ontologies and terminologies are a cornerstone in various electronic health record systems (EHRs) for encoding information related to diseases, diagnoses, treatments, etc. Ontologies in general represent entities (concepts) and events along with all interdependent properties and relationships in an efficient way to facilitate easy access, retrieval and sharing. With the landscape of medicine rapidly changing, biomedical ontologies and terminologies need to rapidly evolve to support interoperability, medical coding, record keeping, and healthcare activities in general, and to facilitate interdisciplinary research. Extending ontologies by identifying new and missing concepts plays a vital role in the maintenance of ontologies to keep up with the constant changes.

Even though different biomedical ontologies capture knowledge in a wide variety of medical domains, they still have substantial overlap in their conceptual content. This dissertation explores various methodologies that can be used to enrich the content of biomedical ontologies and terminologies.

The dissertation is divided into two parts. The first part addresses how cross-ontology topological patterns can be designed and used to identify missing concepts. The methods presented involve comparing horizontal and vertical density differences between identical concepts in two ontologies. Horizontal density studies identified cases of missing child concepts, alternative classifications, synonyms, and errors in ontologies. A deeper analysis of alternative classification is performed. These alternative classifications are

analyzed, and a metric is presented for identifying likely cases of such alternative classifications. Vertical density differences occur when a concept is missing on a path in one ontology but exists in the other one. Furthermore, topological patterns involving three terminologies are presented. A pattern named "fire ladder" incorporates both vertical and horizontal density differences among three terminologies supporting concepts import.

Biomedical ontologies are developed with great investment of time, effort, and budget. Are biomedical ontologies regularly maintained? If not, what are the root causes behind this? A detailed investigation of these questions is conducted both from a quantitative and qualitative perspective.

Ontologies and terminologies are not the only sources of medical concepts. Large repositories of unstructured medical text exist in EHRs. Preliminary studies reveal that reference ontologies and terminologies do not contain many of the frequently recorded fine granularity concepts in EHRs. Recently, with the COVID-19 pandemic, EHRs have been accumulating information regarding new symptoms, procedures and tests that are not all currently present in existing reference ontologies and terminologies. To overcome these issues, in the second part of the dissertation, natural language processing techniques to mine concepts from clinical text are presented. The mined concepts are incorporated into interface terminologies that are catering to the annotation of EHR text in different medical specialties. Mining clinical text to create a COVID interface terminology and a Cardiology interface terminology are discussed. EHR annotation enables secondary use of EHR text for clinical research, such as identifying eligible patients for clinical trials.

METHODS FOR EXTENDING BIOMEDICAL REFERENCE ONTOLOGIES AND INTERFACE TERMINOLOGIES FOR EHR TEXT ANNOTATION

by
Vipina Kuttichi Keloth

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2021

APPROVAL PAGE

METHODS FOR EXTENDING BIOMEDICAL REFERENCE ONTOLOGIES AND INTERFACE TERMINOLOGIES FOR EHR TEXT ANNOTATION

Vipina Kuttichi Keloth

Dr. James Geller, Dissertation Advisor — Date
Professor of Computer Science, NJIT

Dr. Yehoshua Perl, Dissertation Co-Advisor — Date
Professor of Computer Science, NJIT

Dr. Michael Halper, Committee Member — Date
Professor of Informatics, NJIT

Dr. Zhi Wei, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Chunhua Weng, Committee Member — Date
Professor of Biomedical Informatics, Columbia University

Dr. Zhe He, Committee Member — Date
Associate Professor of Information, Florida State University

BIOGRAPHICAL SKETCH

Author: Vipina Kuttichi Keloth

Degree: Doctor of Philosophy

Date: May 2021

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2021

- Master of Technology in Systems Analysis and Computer Applications,
  National Institute of Technology, Karnataka, India, 2014

- Master of Computer Applications,
  Mahatma Gandhi University, Kerala, India, 2010

- Bachelor of Science in Physics,
  Kannur University, Kerala, India, 2007

Major: Computer Science

Publications:

Published Journal Papers

Keloth VK, Geller J, Chen Y, Xu J. Extending Import Detection Algorithms for Concept Import from two to three Biomedical Terminologies. BMC Medical Informatics and Decision Making. 2020 Dec 15;20:272 (2020).

Zheng L, Min H, Chen Y, Keloth VK, Geller J, Perl Y, Hripcsak G. Outlier Concepts Auditing Methodology for a Large Family of Biomedical Ontologies. BMC Medical Informatics and Decision Making. 2020 Dec 15;20:296 (2020).

Zheng L, He Z, Wei D, Keloth VK, Fan JW, Lindemann L, Zhu X, Cimino JJ, Perl Y. A Review of Auditing Techniques for the Unified Medical Language System. Journal of American Medical Informatics Association (JAMIA). 2020 Oct 1;27(10):1625-1638.

Keloth VK, He Z, Elhanan G, Geller J. Alternative Classification of Identical Concepts in Different Terminologies: Different Ways to View the World. Journal of Biomedical Informatics (JBI). 2019 Jun;94:103193.

Submitted Journal Papers

Keloth VK, Geller J. Inquiry into Common Assumptions in Ontology Concept Name Processing. Journal of Biomedical Informatics. Under revision.

Zheng L, Perl Y, He Y, Ochs C, Geller J, Liu H, Keloth VK. Visual Comprehension and Orientation into the COVID-19 CIDO Ontology. Journal of Biomedical Informatics. Under revision.

Published Conference Papers

Keloth VK, Zhou S, Lindemann L, Elhanan G, Einstein AJ, Geller J, Perl Y. Mining Concepts for a COVID Interface Terminology for Annotation of EHRs. In IEEE International Conference on Big Data (Big Data) Proceedings (pp. 3753-3760). 2020.

Keloth VK, Zhou S, Einstein AJ, Elhanan G, Chen Y, Geller J, Perl Y. Generating Training Data for Concept-Mining for an 'Interface Terminology' Annotating Cardiology EHRs. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Proceedings (pp. 1728-1735). 2020.

Geller J, Klein ST, Keloth VK. Measuring and Avoiding Information Loss During Concept Import from a Source to a Target Ontology. In 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) Proceedings (pp. 442-449). 2019

He Z, Keloth VK, Chen Y, Geller J. Extended Analysis of Topological-Pattern-Based Ontology Enrichment. In IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Proceedings (pp. 1641-1648). 2018.

Keloth VK, He Z, Chen Y, Geller J. Leveraging Horizontal Density Differences between Ontologies to Identify Missing Child Concepts: A Proof of Concept. In AMIA Annual Symposium Proceedings (pp. 644-653). 2018.

Geller J, Keloth VK, Musen MA. How Sustainable are Biomedical Ontologies? In AMIA Annual Symposium Proceedings (pp. 470-479). 2018.

Presentations:

Generating Training Data for Concept-Mining for an 'Interface Terminology' Annotating Cardiology EHRs. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), December 15, 2020.

Mining Concepts for a COVID Interface Terminology for Annotation of EHRs. IEEE International Conference on Big Data (Big Data), December 11, 2020.

Leveraging Horizontal Density Differences between Ontologies to Identify Missing Child Concepts: A Proof of Concept. AMIA Annual Symposium. San Francisco, CA, USA, November 5, 2018.

മാതാ പിതാ ഗുരു ദൈവം

To amma and achan. For all the love and support.

To all healthcare workers, who have sacrificed their lives,
while on the front lines of the pandemic.

ACKNOWLEDGMENT

There are many people I must thank for their support and assistance during the five exciting years of my life as a Ph.D. student.

I would like to express my deep gratitude to my advisor Dr. James Geller, whose guidance, support, and encouragement is invaluable in my life. I also express my heartfelt gratitude to my co-advisor Dr. Yehoshua Perl for guiding me both professionally and personally. I will forever be thankful to them for always making themselves available, whether it be for research discussions, editing my terrible drafts and much more. Their dedication and hard work have always been a source of inspiration for me.

I am extremely grateful to Dr. Michael Halper, Dr. Zhi Wei, Dr. Chunhua Weng and Dr. Zhe He for being part of the dissertation committee and providing their valuable insights and feedback to various projects.

I would like to thank the Department of Computer Science at NJIT for providing teaching assistantship for the past four years. Some of the research projects in this dissertation were also partially supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779.

A big thank you to Dr. Gai Elhanan, Dr. Yan Chen, Dr. Mark A. Musen, Dr. Julia Xu, Dr. Andrew J. Einstein, Dr. Luke Lindemann, and all other collaborators who I have had the pleasure to work with and learn from. Special thanks to Dr. Christopher Ochs, Dr. Ling Zheng, Dr. Hao Liu, Ms. Angel J Butler and Ms. Shuxin Zhou for being a part of my NJIT life over the years and making it a pleasant experience.

I would like to acknowledge with gratitude, the love and support of my family – my parents, Chitra NK and Satheendran UK; my sister, Vinisha Rahul and her family,

especially my nephew, Virat Rahul, who with those big, beautiful eyes and gummy smile has made my 2020 a lot better. Warmest thank you to my brother, Arun C Kamal, for inspiring me and always believing in me more than I do. Finally, I would like to express a huge thank you to my friend, Rajarshi Bhowmik for the great companionship and for being there through my laughs, tears, fights, dreams, and celebrations.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(Continued)

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Motivation

Over the past two decades, the biomedical research community has made tremendous progress in developing ontologies and terminologies that encode biomedical knowledge about entities (concepts) and their relationships to each other [1, 2]. Biomedical ontologies and terminologies are important in Electronic Health Record (EHR) systems (e.g., Epic [3], Cerner [4], Allscripts [5], etc.) and play a major role in facilitating clinical practice, healthcare applications, biomedical research, and enable reasoning with biomedical knowledge. For example, the International Classification of Diseases (ICD- 9/10) [6] diagnosis code set is used for billing, the Current Procedural Terminology (CPT) [7] is used to code procedures, SNOMED CT [8] is used to encode the clinical data in patient records, and Logical Observation Identifiers Names and Codes (LOINC) [9] to encode lab and other observations.

One of the major challenges for biomedical ontologies is keeping up with the pace of the rapidly changing biomedical sciences. Hence, a significant amount of effort has gone into providing the right infrastructure for the maintenance and interoperability among the ontologies. Extending biomedical ontologies and terminologies by adding new concepts is a vital part of this maintenance effort. With the COVID-19 pandemic, many ontology and terminology developers had to take immediate actions by issuing interim releases and updating their ontologies and terminologies by adding new concepts enabling clinicians and researchers to code and analyze huge volumes of EHRs of COVID-19 patients [10, 11].

The distinctions between ontologies and terminologies have been ably discussed by other researchers [12-14]. In short, an ontology is concerned with the study of classes of entities and the relations among them and is more than just a list of terms, while the purpose of a terminology is to collect all the entities in a particular domain. The words "terminology" and "ontology" are used in an inclusive sense in this dissertation, whether a source is an ontology or "only" a terminology (vocabulary, etc.).

The number of ontologies and terminologies has grown rapidly over the past decade. BioPortal [15, 16], which is widely regarded to be the world’s most comprehensive repository of biomedical ontologies, reported around 300 ontologies in 2013 [17] and has grown to 841 ontologies to date [18]. The Unified Medical Language System (UMLS) Metathesaurus [19] was designed to enable effective retrieval of information and understanding of the meaning of different entities across different ontologies and terminologies. The UMLS Metathesaurus contains over 200 general and specialized biomedical terminologies in about 25 different languages; however, the majority of these terminologies are in English [20]. The concepts in these terminologies are organized in hierarchical structures based on their relationships to one another. Synonymous terms are clustered into a unique concept, identified by a Concept Unique Identifier (CUI).

Terminologies vary from one another in several aspects like the domain or subject area they cover, the level of abstraction, the level of detail, the modeling philosophy, and what language its terms are taken from. For example, the National Cancer Institute Thesaurus (NCIt) [21] is a reference terminology that includes broad coverage of the cancer domain, whereas the Gene Ontology (GO) [22] is an ontology for describing genes and their functions. Even though ontologies differ in their domains, there is a significant overlap in the conceptual content among many pairs of ontologies. For instance, according to the 2018AB version of the UMLS, NCIt has 20% overlap with MeSH [23], 17.4% overlap with SNOMED CT, and 7.1% overlap with MEDCIN [24].

In prior work, the overlap in the conceptual content was studied in detail [25-27] and it was found that there are substantial differences in the "vertical density" of the conceptual content among different ontologies. Rector et al. [28] defined density as "The number of semantically 'similar' concepts in a particular conceptual region" and further states that "High local density in an ontology usually co-occurs with high levels of specialisation and degree of detail, …..". Specific topological patterns called *m:n trapezoids* [27] and *Cross-Ontology Diamonds* [25] (Figure 1.1) were proposed to demonstrate the differences in vertical density. Detailed analyses of these patterns were conducted to identify missing concepts, synonyms, errors, etc.

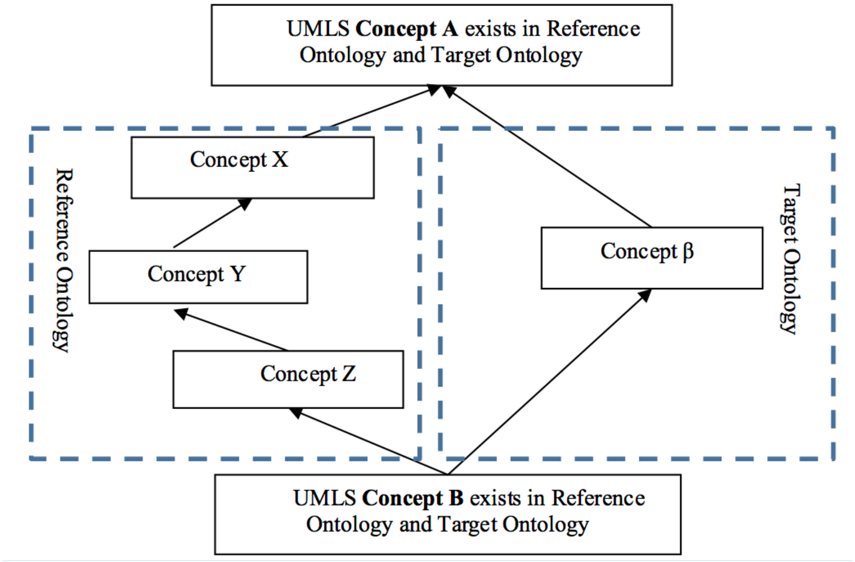

UMLS **Concept A** exists in Reference Ontology and Target Ontology
Concept X
Reference Ontology
Concept Y
Concept Z
Concept β
Target Ontology
UMLS **Concept B** exists in Reference Ontology and Target Ontology

Figure 1.1 An example of a Cross-Ontology Diamond. The arrow represents IS-A hierarchical relationship.
Source: [25]

This dissertation extends the prior work by identifying patterns based on "horizontal density" differences. These patterns reveal cases of missing child concepts in an ontology and alternative classifications of identical concepts in different ontologies. A mathematically expressed criterion to identify potential cases of alternative classifications is introduced. Additionally, the methodology is extended from two ontologies to three ontologies to explore a new pattern named "fire ladder."

The previously discussed methods explored the structural differences among pairs of ontologies to identify missing concepts that could be imported from another ontology. However, a huge amount of medical data also exists in the form of unstructured text in EHRs. This data contains finer levels of information that are usually not present in the standard ontologies. Two natural language processing techniques named "concatenation" and "anchoring" were used to mine fine granularity concept names from unstructured clinical text data. The benefits of creating an interface terminology that has comprehensive coverage of a particular medical specialty to annotate EHR data are presented. These techniques are demonstrated for the COVID-19 and the cardiology domain.

1.2 Dissertation Overview

Chapter 2 provides brief information about different biomedical ontologies, integrated terminological systems, and clinical databases. In addition, Chapter 2 presents a detailed review of the literature in the areas of semantic harmonization and extension of ontologies. A distinction between reference ontologies and interface terminologies is also provided.

Chapter 3 presents a study exploring the horizontal density differences in pairs of ontologies in the UMLS. The study introduces an algorithm and a metric that automatically suggests child concepts that are likely to be imported into SNOMED CT and NCIt from eight other ontologies in the UMLS.

Chapter 4 explores the idea of "alternative classifications" of identical concepts in different terminologies. The study demonstrates a revised algorithm and metric for importing child concepts into NCIt from MEDCIN. The metric automatically identifies likely cases of "alternative classifications," that were discovered in this research. Furthermore, a detailed analysis of different types of alternative classifications is provided.

Chapter 5 demonstrates how topological patterns can be extended from two terminologies to three terminologies leading to the discovery of "fire ladder" patterns.

Chapter 6 reports a study conducted on the ontologies in BioPortal. This study analyses the life cycle of ontologies, specifically those ontologies that have more than a thousand concepts. The root causes for not updating the ontologies are studied and they are categorized into groups, based on the phenomenological approach to qualitative data analysis.

Chapter 7 discusses two studies on identifying new fine granularity concepts from medical text to create interface terminologies for annotation of EHRs. The techniques are demonstrated for the cardiology and COVID-19 domains.

Finally, Chapter 8 presents future work and Chapter 9 contains concluding remarks about the studies discussed in previous chapters. The studies in Chapter 3 and Chapter 6 were published in the Proceedings of the American Medical Informatics Association (AMIA) 2018 Annual Symposium [29, 30], Chapter 4 in the Journal of Biomedical Informatics [31], and Chapter 5 in BMC Medical Informatics and Decision Making [32]. The two studies in Chapter 7 were published in the Proceedings of the IEEE International

Conference on Big Data [33] and the IEEE International Conference on Bioinformatics and Biomedicine [34].

CHAPTER 2

BACKGROUND

2.1 Biomedical Ontologies

Biomedical ontology research encompasses a variety of entities ranging from dictionaries of names for biological products to controlled vocabularies to knowledge bases and processes for the acquisition of ontological relations [2]. Biomedical ontologies are widely used to facilitate research in many other domains like knowledge and data mining [35, 36], natural language processing tasks [37, 38], and other health care applications [39-41]. This section will introduce some large biomedical ontologies and terminologies that are relevant to this dissertation.

2.1.1 SNOMED CT

SNOMED CT is a standardized vocabulary of clinical terms used by healthcare staff all over the world for the electronic exchange of clinical health information [42]. SNOMED CT is managed by IHTSDO (International Health Terminology Standard Development Organization), which in 2017 adopted the trading name of "SNOMED International." The number of concepts in SNOMED CT continues to grow, and the January 2021 International Edition of SNOMED CT contained 354,448 active concepts and 1,178,592 relationships [43]. SNOMED CT is a global clinical ontology, and for improving interoperability and healthcare it is being translated into local languages by the SNOMED CT members.

Concepts in SNOMED CT are arranged in 19 hierarchies under a root concept named *SNOMED CT Concept*. Some of the hierarchies are *Body structure, Clinical finding, Specimen, and Procedure.* Each concept in SNOMED CT has at least one IS-A relationship with a parent concept except for the root concept. A concept can have more than one parent concept. Also, a concept can have one or more "attribute relationships." For example, the concept *Fracture of tarsal bone* IS-A *Fracture of foot* and has *finding site – Bone structure of tarsus* and *associated morphology – Fracture,* where *finding site* and *associated morphology* are attribute relationships.

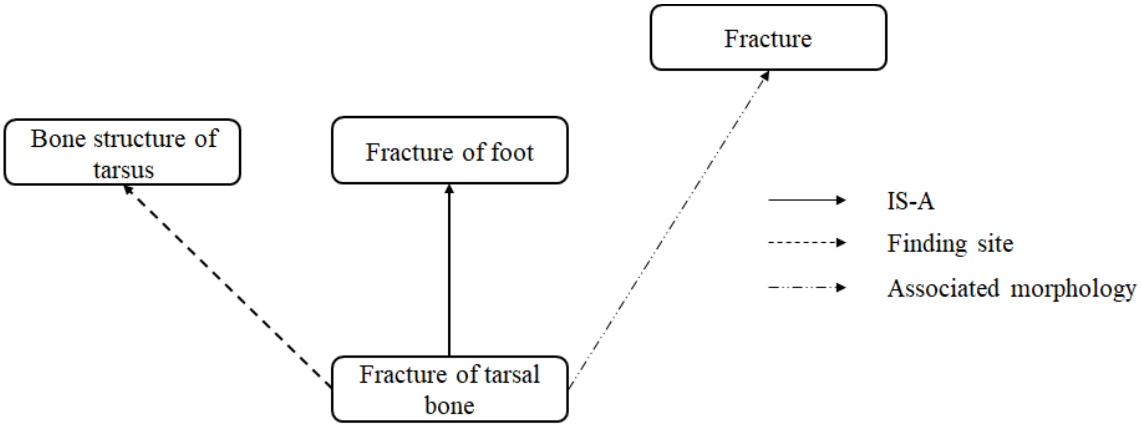


Figure 2.1 An example of different types of relationships in SNOMED CT.

The U.S. SNOMED CT Content Request System (USCRS) allows SNOMED CT users to submit change requests to the curators of SNOMED CT [44]. Online forms provided by the system allow users to submit requests for adding new concepts, relationships, etc., changing descriptions and relationships and retiring concepts, etc.

2.1.2 National Cancer Institute Thesaurus (NCIt)

The National Cancer Institute Thesaurus (NCIt) is a reference ontology for many systems and is a widely recognized standard for biomedical coding. NCIt enables retrieval of information across a wide range of domains used in cancer research, facilitating the process of migrating basic research into clinical research and practice in the cancer research domain [21]. NCIt has been developed by NCI Enterprise Vocabulary Services (EVS) [45] to facilitate standardization of terminology use across the biomedical community. NCIt is updated monthly and a recent version of NCIt is the January 2021 release version 21.01d. NCIt has stable, unique codes for biomedical concepts with over 100,000 textual definitions and 400,000 relationships between concepts [46]. NCIt provides cross-links and semantics for a broader range of EVS terminology resources. SNOMED CT and NCIt are two of the most widely used ontologies with large numbers of concepts. The concepts in these ontologies overlap with each other and with several other UMLS ontologies.

NCIt is modeled based on the description logic (DL) paradigm [21, 47]. NCIt can be accessed on the Web [48], or by file download in any of three formats: Ontylog XML, OWL Lite, and ASCII flat file. NCIt is maintained by a multidisciplinary team of editors, who in the past have added about 700 new entries each month. NCIt also provides facilities that allow users to place requests for adding, updating, and retiring concepts and relationships [45]. The concepts in NCIt are structured into 19 logically disjoint classes; for example, *Disease, Disorder or Finding*, *Abnormal cell*, *Gene*, and *Molecular Abnormality*. The concepts are organized in multiple parent-child IS-A hierarchies within each class and have over 100 distinct role relationships that provide asserted and inherited logical links between pairs of concepts. Figure 2.2 shows 13 neoplasm concepts in the *Disease, Disorder or Finding* hierarchy of NCIt.

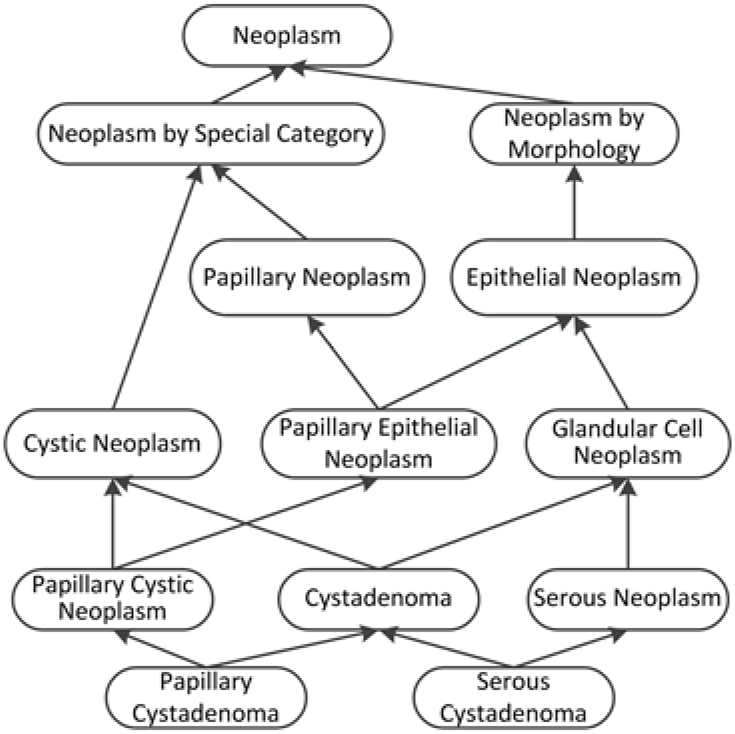


Figure 2.2 An excerpt of 13 neoplasm concepts in the Disease, Disorder or Finding hierarchy of NCIt. Concepts are represented by rounded rectangles and the arrows represent IS-A relationship between the concepts.

2.1.3 MEDCIN

MEDCIN was created and is maintained by Medicomp Systems Inc [49]. It is a medical terminology that encompasses symptoms, tests, diagnoses, physical examinations, etc. It was designed to allow for rapid entry, retrieval, and correlation of relevant clinical information at the point of care, to enable applications to store medical information as coded data elements, and to produce narrative reports from the same data. The approach used in developing MEDCIN was to organize the individual data elements (findings) into six broad categories namely symptoms, history, physical examination, tests, diagnoses, and treatments which reflect the types of information acquired during clinical processes [24].

MEDCIN defines each finding with a degree of sensitivity and clinical specificity so that it can be used for differential diagnoses and is suitable for documentation and research. The findings are organized giving due consideration to the number of steps required to record a finding and the number of search elements required to locate it for future analysis. For this purpose, MEDCIN uses a data structure that embodies the concept of inheritance between levels in a hierarchy. For example, MEDCIN has more than 80 findings for *"numbness,"* starting with the general finding and proceeding in a structured hierarchy through all areas of the body, increasing in detail at each level [24].

MEDCIN is updated regularly throughout the year and the updated files are released at least twice a year. In the 2020AB release of the UMLS, MEDCIN has around 358,221 concepts. MEDCIN has about 3% source overlap with NCIt and about 13.9% with SNOMED CT.

2.1.4 Coronavirus Infectious Disease Ontology (CIDO)

The Coronavirus Infectious Disease Ontology (CIDO) [50] was created to provide a standardized representation of various coronavirus infectious diseases. CIDO provides standardized human- and computer-interpretable annotations and representations of various infectious coronavirus diseases, including their etiology, transmission, epidemiology, pathogenesis, diagnosis, prevention, and treatment. Its development follows the OBO Foundry Principles [51]. CIDO is released in OWL format and is freely available on the GitHub repository [52]. CIDO is available on BioPortal [53] and Ontobee [54]. Currently, CIDO is at version 1.0.184 with a total of 6,938 concepts and 371 properties of which 201 are relationships. Concepts are interconnected by 201 lateral relationship types such as *caused by, infection with,* and *treatment for*.

As CIDO follows the OBO Foundry [51] principles, there is extensive reuse of concepts from about 20 other ontologies including the Chemical Entities of Biological Interest (ChEBI) [55] and the Human Phenotype Ontology (HPO) [56]. For example, drug concepts are reused from ChEBI, the National Drug File - Reference Terminology (NDF-RT) [57], and the Drug Ontology (DrON) [58]. CIDO contains 244 original COVID-19-specific concepts.

2.2 Integrated Terminological Systems

With the rise in the number of biomedical ontologies, research was also directed towards integrating data from different ontologies and providing platforms for the same. The Unified Medical Language System (UMLS) [19, 59], and BioPortal [16] were created towards achieving this goal. The UMLS aims at merging existing vocabularies. The National Center for Biomedical Ontology (NCBO) BioPortal provides access to commonly used biomedical ontologies and also tools for working with them. This section will describe these two systems in detail.

2.2.1 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS), initiated in 1986, was designed and is maintained by the US National Library of Medicine [59]. It brings together many health and biomedical vocabularies and standards to enable interoperability and can be used to enhance and develop applications, such as Electronic Health Records, classification tools, dictionaries, and language translators. The various subdomains integrated into the UMLS include biomedical literature, genome annotations, anatomy, genetic knowledge bases, model organisms, clinical repositories, and many more [19]. The UMLS has three components namely, the Metathesaurus, Semantic Network, and SPECIALIST Lexicon with Lexical Tools. The UMLS Metathesaurus contains terms and codes from over 200 vocabularies including CPT, ICD-10, LOINC, MeSH, RxNorm [60], and SNOMED CT. The Semantic Network provides the semantic types and the semantic relations between the concepts. A large syntactic lexicon of biomedicine and tools for normalizing strings, generating lexical variants, and creating indexes are provided by the SPECIALIST Lexicon with Lexical Tools component.

The Metathesaurus organizes concepts by their meaning and synonymous terms are grouped into concepts identified by a Concept Unique Identifier (CUI). Different concepts are linked to each other using various types of relationships and each relationship is identified by a Relationship Unique Identifier (RUI). Relationships (REL) have labels, describing their nature, such as, *has parent relationship* (PAR), *has a broader relationship* (RB), etc. Some relationships also have an additional annotation (RELA) that gives a more detailed explanation of the relationship such as *consists_of*, *isa*, *part_of*, *location_of*, etc.

The UMLS is updated in May and November of each year. The May release is denoted as the AA release and the November release is denoted as the AB release. The 2020 AB release of the UMLS is distributed in the RRF format (Rich Release Format). This release of the UMLS has 4,413,092 unique concepts from 215 source vocabularies in 25 different languages [20]. In total, 70.78% of concepts in the Metathesaurus are in English.

2.2.2 NCBO BioPortal

BioPortal is a repository of biomedical ontologies – the largest of its kind, with 841 ontologies (on 02/24/2021) and growing. The ontologies in BioPortal fall under different categories namely those developed in OWL (Web Ontology Language) format [61], others in OBO (Open Biomedical Ontologies) format, several medical terminologies from the UMLS in the RRF format, and a few in SKOS (Simple Knowledge Organization System) format [62]. Ontologies in BioPortal vary widely in the topic areas that they cover. For example, some of the topics are health, anatomy (plant, animal, fish), phenotype, chemical, genomic, and proteomic subject areas, etc.

BioPortal maintains information about various aspects of an ontology. BioPortal has a submission page for each ontology. Available data on the submission page includes the release date, upload date, ontology format, submission id, short description, URL of the homepage, version number, etc. of all the submissions of the ontology to BioPortal. Other items of information provided in BioPortal for individual ontologies include the number of classes, the number of properties, etc.

BioPortal SPARQL [63] is a service from BioPortal to query biomedical ontologies using the SPARQL standard. All the BioPortal's ontologies have been transformed into RDF triples from their different original formats which helps users with uniform access to key properties of the ontologies [17]. Furthermore, it also contains cross-ontology mappings of different types, generated both manually and automatically.

2.3 Clinical Databases

Clinical data in the form of unstructured text present in discharge summaries, lab reports, progress notes, etc. contain relevant information and detailed descriptions of patient conditions. Many fine granularity concepts that are frequently used by medical professionals are recorded in these unstructured clinical text segments. The standard ontologies and terminologies often do not contain such concepts. To mine these concepts, data from two clinical databases was used which are described in detail in this section.

2.3.1 MIMIC-III Intensive Care Database

MIMIC-III (Medical Information Mart for Intensive Care) is a freely accessible, de-identified critical care database comprising information relating to patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts [64]. The data is very diverse ranging from vital signs, medications, and laboratory measurements to procedure codes, diagnostic codes, billing information, and survival data. As an extension to the database, MIMIC III also contains waveform data from ECG and EEG measurements [65]. The data is de-identified following the Health Insurance Portability and Accountability Act (HIPAA) regulations [66]. A schematic overview of the database along with the data extraction process is shown in Figure 2.3. The database contains data from patients admitted to the hospital from 2001 to 2012 with a total of 61,532 ICU stays and 46,476 unique patients.

The MIMIC-III Database provides a table named "NOTEEVENTS" that records free text "notes" produced by the hospital staff during the course of the patients’ stay in the ICU. Nurses write "nursing notes," which summarize the events that occurred during their shift. When a patient is discharged from the hospital, the physician summarizes the entire hospitalization period in the form of a "discharge summary," which is also recorded in the database table. In addition to the progress or nursing notes and discharge summaries, the table also records reports of diagnostic tests including X-rays, echocardiograms, and ECGs. Currently, the table "NOTEEVENTS" has 2,083,180 entries, including 59,652 discharge summaries.

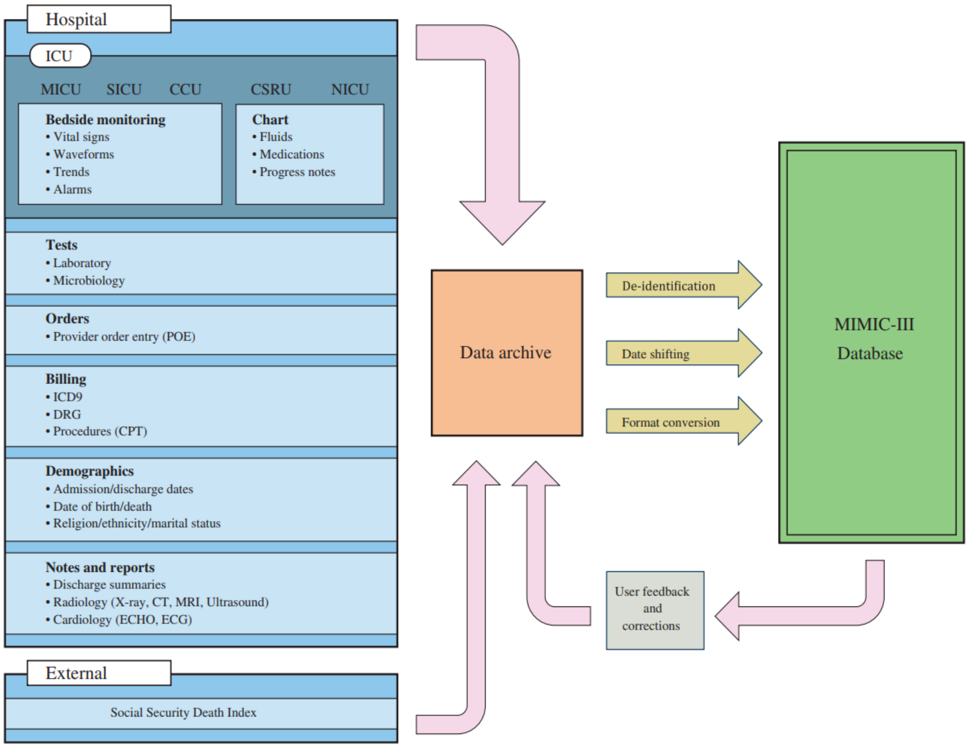


Figure 2.3 Overview of MIMIC-III Database.
Source: [64]

2.3.2 COVID-19 Radiology Case Studies

Radiology imaging reports, including Computed Tomography (CT) imaging and X-rays play a major role in the diagnosis and management of COVID-19 patients. Radiology case studies are chronicles of patient progress describing classic and unusual presentations of diseases with a focus on findings in CTs and X-rays. SIRM - the Italian Society of Medical and Interventional Radiology has provided a COVID-19 Database [67]. This database contains COVID-19 radiological case studies. There are 115 case studies to date. Each case study describes patient demographic information such as age and sex, prior medical history, symptoms, detailed CT and chest X-ray findings, and medications. X-ray and CT images are also provided with a detailed description of the findings.

2.4 Semantic Harmonization and Extension of Ontologies

Biomedical knowledge is in constant evolution and growth. Hence, developing and extending biomedical ontologies is a process that can never be considered complete. With several ontologies sharing similar domains, semantic harmonization and interoperability add another dimension to this challenge. This section introduces prior work that addresses challenges in achieving semantic harmonization. Prior research focusing on the structural differences among ontologies with similar concepts, and how these differences can be used in extending ontologies, is also discussed in detail.

Semantic harmonization is a process of combining data from multiple sources and representations into a form that facilitates sharing the meaning of data across these sources and representations. Cunningham et al. [68] identify "re-using standard vocabularies wherever possible" with a mechanism to add new terms, as one of the nine principles of semantic harmonization. Weng and Fridsma [69] proposed a conceptual design for collaborative semantic harmonization that identified three key design principles, namely, reuse, collaboration, and harmonization as modeling. Research areas dealing with ontology matching [70, 71], ontology mapping [72] and alignment [73-76] that strive to achieve semantic harmonization and interoperability, are described in detail in Section 2.4.1.

Issues that hinder semantic harmonization and interoperability are due to the modeling policies adopted by different ontologies and the level of detail they require when representing the knowledge in the same domain to support their applications. Literature in this area uses either "granularity" or "density" to denote this difference in the level of detail in different ontologies. In Section 2.4.2, previous work related to granularity/density differences in ontologies that enables the extension of ontologies is described.

2.4.1 Ontology Matching/Alignment

Ontology alignment or ontology matching – the process of finding correspondences between concepts in ontologies – has been studied for a long time [70-72, 77-79]. Many publications have surveyed the state-of-the-art in this field [72, 79, 80]. There are different ontology matching techniques based on the approaches used to implement the matching algorithms. Euzenat and Shvaiko [70] proposed a matching technique classification as shown in Figure 2.4.

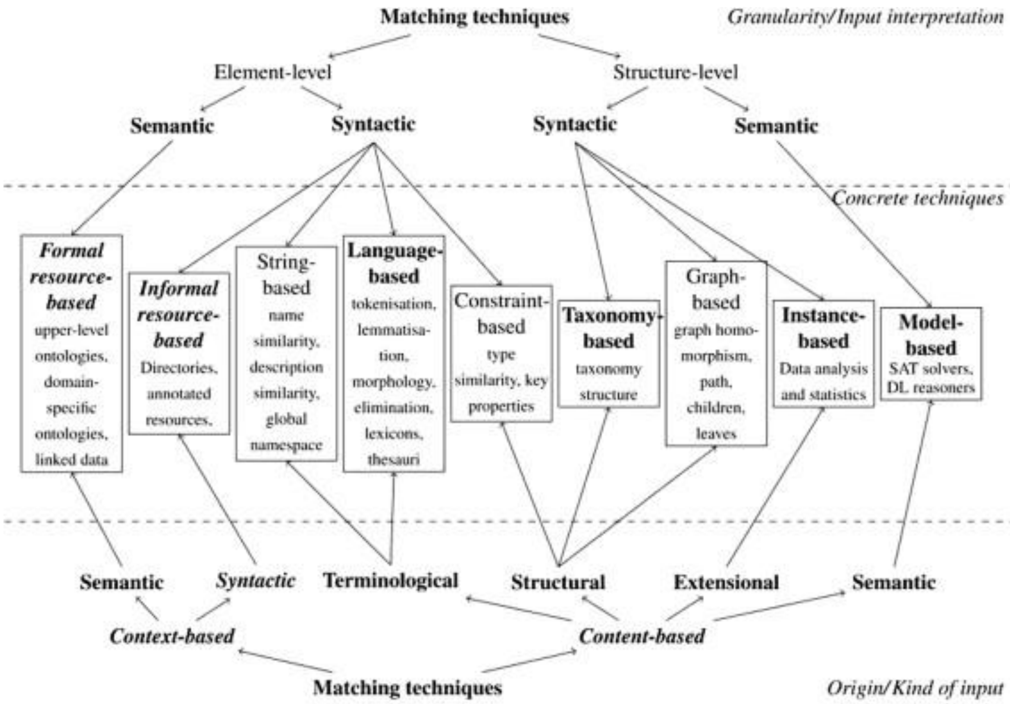

Matching techniques
Granularity/Input interpretation
Element-level
Structure-level
Semantic
Syntactic
Syntactic
Semantic
Concrete techniques
Formal resource-based
upper-level ontologies, domain-specific ontologies, linked data
Informal resource-based
Directories, annotated resources,
String-based
name similarity, description similarity, global namespace
Language-based
tokenisation, lemmatisation, morphology, elimination, lexicons, thesauri
Constraint-based
type similarity, key properties
Taxonomy-based
taxonomy structure
Graph-based
graph homomorphism, path, children, leaves
Instance-based
Data analysis and statistics
Model-based
SAT solvers, DL reasoners
Semantic
Syntactic
Terminological
Structural
Extensional
Semantic
Context-based
Content-based
Matching techniques
Origin/Kind of input

Figure 2.4 Classification of ontology matching techniques.
Source: [70]

The classification hierarchy can be read top-down as well as bottom-up. The top-down interpretation is based on whether the matching techniques use:

1. *Element-level matches* (obtain the correspondences by considering the entities in isolation) or

2. *Structure-level matches* (obtain the correspondences by analyzing how the entities fit in the structure of the ontology).

Reading the classification bottom-up, the matching techniques are initially divided into two categories based on the origin of the information considered for the matching process as

1. *Content-based* (focus on the internal information coming from the ontologies to be matched) or

2. *Context-based* (focus on external information that comes from relations between ontologies or other external resources).

The methodologies described in different studies in this dissertation use the structural classification of the content-based matching techniques. The Ontology Alignment Evaluation Initiative (OAEI) [73] works towards achieving consensus for the evaluation of these methods. This initiative contains various tracks with different data sets that evaluate different features of the system to be tested.

Most research in this field focuses on methods for finding simple *1:1* correspondence between concepts in two ontologies. Only a very few alignment systems have focused on finding complex correspondences. Recently, Zhou et al. [74] proposed a complex alignment benchmark based on the real-world GeoLink dataset. The alignments in this dataset not only cover *1:1* correspondences but also contain *1:n* and *m:n* complex relations. They have identified 12 different kinds of simple and complex correspondence patterns and made available the alignments in both rule and EDOAL syntax [81]. Gouveia et al. [78] proposed a method for ensuring the quality of ontology alignments by identifying ambiguous *1:1*, *1:n*, *n:1*, and *m:n* ontology matching scenarios and providing sets of resolution strategies.

Oliveira and Pesquita [76] proposed a set of algorithms to create ternary compound alignments (compound matching of three distinct ontologies) for large biomedical ontologies. Another work, related to ontology integration based on mapping, repair, and conservative alignment proposed by Stoilos et al. [82] focuses on building a framework that integrates several medical ontologies to support real-world health care services. The integration starts with a seed ontology to which new ontologies are added to enrich and extend the seed ontology. They also developed algorithms to deal with structural incompatibilities.

2.4.2 Granularity/Density Differences and Ontology Enrichment

Ontologies often adopt different levels of detail when representing the same piece of knowledge to support different applications. Many different terms such as granularity, specialization, degree of detail, density, etc. are used in literature to describe these differences. Rector et al. [28] provided a detailed account of the terms granularity, scale, collectivity, specialization, degree of detail, density, and connectivity with the help of examples taken from biomedicine.

Prior research has explored structural incompatibility in granularity (granularity differences) among ontologies in the UMLS as a tool to enhance the conceptual content of the ontology. This research mainly uses a *rule-based approach* or a *topological pattern-based approach*. As an example of the rule-based approach, Sun and Zhang [83] identified granularity differences as well as similarities between large biomedical ontologies through rules. They investigated the examples of correspondence across two anatomical ontologies (the Adult Mouse Anatomical Dictionary (MA) [84] and NCIt) and synthesized patterns, and constructed rules that were then fed into a rule inference engine to distinguish among different subclasses and classifications. An example of multiple subclasses for an anchor concept (*MA: digestive system fluid/secretion, NCIt: Gastrointestinal Fluid Or Secretion*) is shown in Figure 2.5.

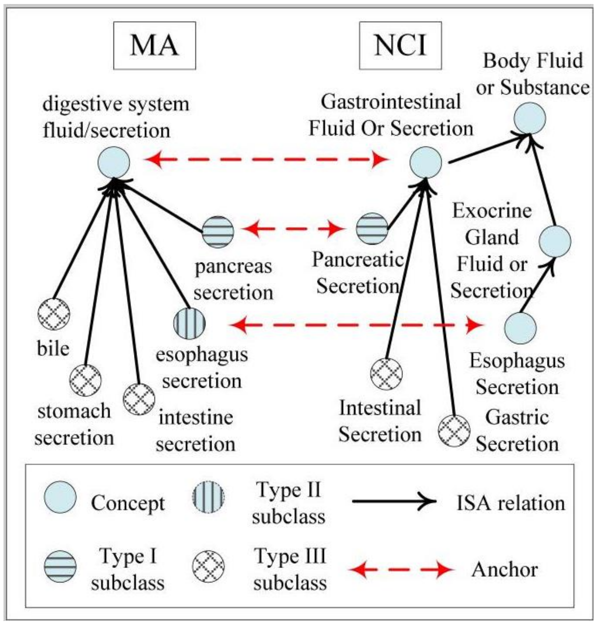


Figure 2.5 Subclass classification similarity and differences across MA and NCIt.
Source: [83]

In another work with this approach, the same authors [85] conducted a parallel study to construct rules to systematically identify the differences as well as similarities in a partonomy (a hierarchy of part-whole relationships) between two biomedical ontologies (MA and NCIt) instead of using IS-A relationships. The rules used in these two studies were constructed after a manual investigation of some examples of anchor concepts and hence do not cover all cases of structural incompatibility among ontologies. The main limitation of this approach is that every time a new mismatch pattern is identified new rules need to be added.

Luo et al. [86] proposed a method for evaluating the granularity balance of IS-A and part-of relationships *within* a biomedical ontology. They used "parallel concept set (PCS)" (two concepts that share a similar level of conceptual knowledge) and the length and strength of the paths between the PCSs to design evaluation models to improve the quality of one ontology.

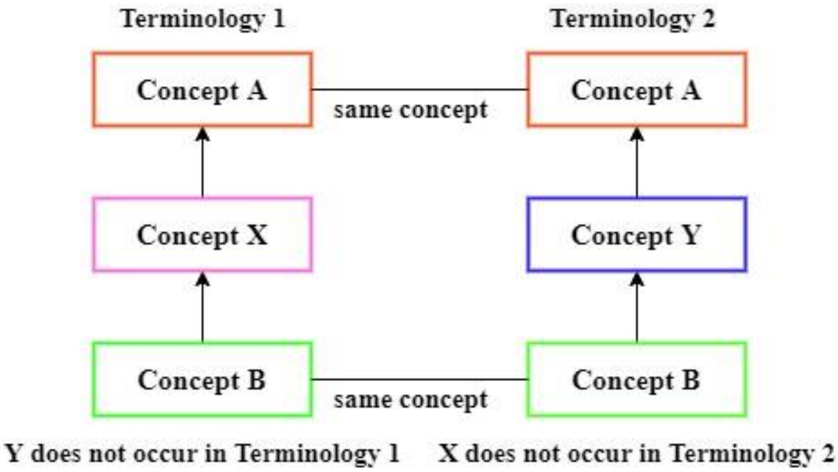


Figure 2.6 An abstract layout of structurally congruent concepts.
Source: [26]

The topological pattern-based approach mainly emphasizes identifying the vertical density differences among different biomedical terminologies. In this work, He et al. [26] used "structurally congruent concepts" in pairs of terminologies as a method for harmonizing the terminologies. Two concepts X (from Terminology 1) and Y (from Terminology 2) are called "structurally congruent" if X and Y have the same parent and the same child in both the terminologies and concept X does not appear anywhere in Terminology 2 and Y does not appear anywhere in Terminology 1 and X and Y are not synonyms. Figure 2.6 shows an abstract layout of two structurally congruent concepts (X and Y). They used six UMLS terminologies to pair with SNOMED CT in their study. The structurally congruent concepts were interpreted in six possible ways, including alternative classification, synonyms, structural errors, etc.

The work on structurally congruent concepts was extended to develop structure-based algorithmic methods to identify the concepts that could enrich the conceptual content of SNOMED CT [27]. More complex topological patterns like *m:n trapezoids* were extracted with the help of the proposed trapezoid identification algorithm, and potential concepts for inclusion as parents, children, synonyms, etc. into SNOMED CT were identified. Analogous methods were tested to locate potentially missing concepts for NCIt, using eight source terminologies from the UMLS [87]. The usefulness of the NCI Metathesaurus instead of the UMLS Metathesaurus for enriching NCIt was also studied in detail [88]. Figure 2.7 shows the layout of a 2:3 trapezoid pattern.

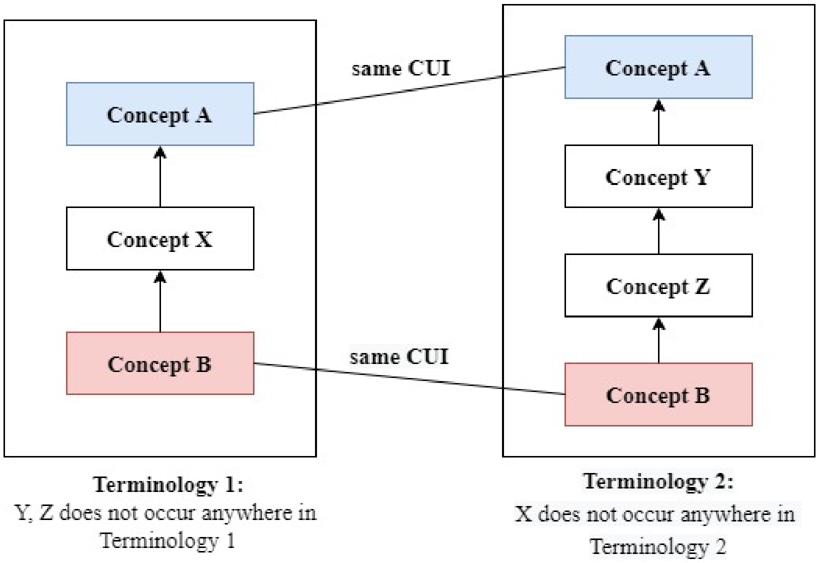


Figure 2.7 The layout of a 2:3 trapezoid topological pattern. The arrow represents IS-A hierarchical relationship.
Source: [27]

The main limitation of the studies that use topological patterns is that although the process of identifying the potentially missing concepts is automated, it still requires a human expert to review the suggestions and make the decision as to whether to import a concept or not. A study has been conducted to estimate the difficulty of this task for a domain expert and it proved that it is still challenging even with the support of the algorithm that offers suggestions for import [25, 89].

2.5 Reference Ontologies vs. Interface Terminologies

Biomedical ontologies represent classes of entities, focused on the definition of classes and the relations among them. Ontologies represent knowledge in a formal, principled way. Burgun in a paper titled "Desiderata for domain reference ontologies in biomedicine" [90] describes domain reference ontologies as ontologies that "represent knowledge about a particular part of the world in a way that is independent from specific objectives, through a theory of the domain represented." Five desirable characteristics are suggested for reference ontologies which include good lexical coverage, good coverage in terms of relations, compatibility with standards, modularity, and ability to represent variation in reality.

Terminologies are generally built to serve applications such as document retrieval, resource annotation, and healthcare billing. Biomedical terminologies do not use formal and well-defined descriptions; they usually define the terms by human language expressions. Although terminologies can be successfully used in representing abstract meaning, they are not precise and expressive enough for more knowledge-intensive applications [91].

Rosenbloom et al. [92] defined a ***clinical interface terminology*** as "a systematic collection of healthcare-related phrases (terms) that supports clinicians' entry of patient related information into computer programs, such as clinical 'note capture' and decision support tools." It is an interface between the users and the standard reference terminologies required by clinical information systems [93]. Interface terminologies are designed with the end-users in mind and hence consist of relatively common clinical phrases and colloquial usages as opposed to a standard concept-based aggregation of clinical information in a reference ontology.

2.6 Biomedical Annotation Tools

A suite of biomedical annotation tools is available for use. Some of these are general-purpose tools in that they detect a wide variety of entity types, and in most cases, they link terms to concepts in the UMLS. The cTAKES [94], MetaMap [95], QuickUMLS [96], and NCBO Annotator [97, 98] are examples of general-purpose annotation tools. The NCBO Annotator (previously known as the Open Biomedical Annotator) is an ontology-based web service, available on the BioPortal platform [15], which tags biomedical text automatically with ontology terms. With NCBO Annotator it is possible to annotate text with concepts from an ontology of the user's choice (from 840+ BioPortal ontologies). A user can also upload a proprietary ontology to BioPortal, and utilize this ontology to annotate text, thereby customizing the annotations based on the requirements of different studies. MetaMap was developed by the National Library of Medicine (NLM) and can annotate clinical text with appropriate concepts from the UMLS. cTAKES is another system for the extraction of information from clinical text with the UMLS. It is available as open-source software from Apache.

Another category of annotation tool is trained to identify specific entity types like genes, disease, chemicals, etc. PubTator [99] and BERN [100] are examples in this category. Yet another category of annotation tool is trained on manually annotated datasets using machine learning techniques and can work on a wide variety of entity types based on the training data available. CLAMP [101] is an example of such an annotation tool. CLAMP incorporates several machine learning components and the latest release, 1.6.0, contains multiple deep learning modules.

CHAPTER 3

HORIZONTAL DENSITY DIFFERENCES IN ONTOLOGIES

Two concepts that occur in different ontologies are known to be identical if they have the same Concept Unique Identifier (CUI) in the UMLS Metathesaurus. It was observed that such concepts rarely have exactly the same sets of children (=subconcepts) in the two ontologies. The number of common children was found to vary widely. While there were some cases in which the identical concepts had exactly the same child concepts in both the ontologies, in a majority of the cases the number of common children differs widely. This difference in the sets of child concepts for identical parent concepts in two different ontologies is referred to as horizontal density difference. It was also observed that in many cases there were no common children.

This chapter presents a study on analyzing horizontal density differences and exploring how these differences can be utilized to enrich the conceptual content of an ontology. The study leverages the horizontal density differences among pairs of ontologies in identifying missing child concepts in the target ontology. An algorithm for identifying missing child concepts along with a metric that filters these concepts is introduced. The identified concepts are reviewed by domain experts for inclusion into the ontology and results are presented.

3.1 Concept Import and Density Differences

The existing biomedical ontologies differ widely in the domain they cover. With the medical knowledge continuously expanding, new concepts are being added regularly. As discussed previously, even though ontologies differ in their domain knowledge, there is substantial overlap in the conceptual content of the ontologies. This overlap can be utilized for identifying missing child concepts by exploring the vertical and horizontal density differences. Firstly, a distinction between vertical density difference (which was previously explored in the literature and described in Section 2.4.2) and horizontal density difference (which is the subject of this study) is provided.

Consider pairs of ontologies with local similarities where corresponding "IS-A" paths in those ontologies are of different lengths. For example, in Figure 3.1 concept GP (grandparent of the focal concept F) and concept GC (grandchild of F) are common in both ontologies. The ontology on the left side is the source ontology (the ontology which supplies missing child concepts). The ontology on the right side is the target ontology (the ontology into which the concepts are imported). In the source ontology, between the concepts GP and GC, there are three other concepts P (parent), F, and C (child), whereas, in the target ontology, there is only one concept δ between GP and GC. This difference in the number of intermediate children which leads to "IS-A" paths of different lengths is what is called *vertical density difference*. The resulting topological patterns were named *m:n trapezoids* [27, 87] and *m/n cross-ontology diamonds* [25]. Prior work explored the possibility of importing additional concepts into the target ontology whenever there are more concepts on the source ontology side [25, 27, 88]. Notice that these topological patterns are highly constrained and used to examine concepts within the boundaries of identical ancestors and descendants in both ontologies.

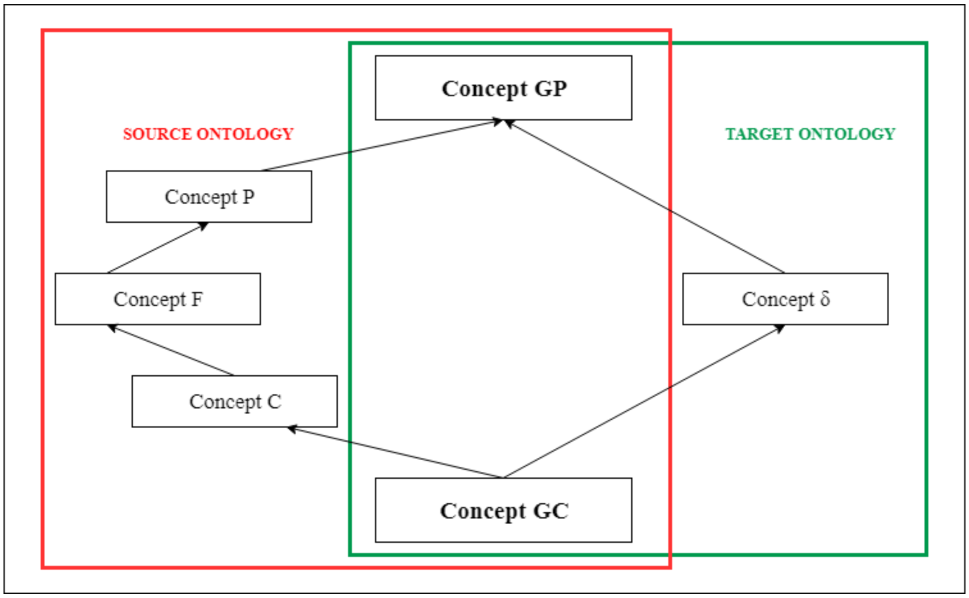


Figure 3.1 An abstract layout for demonstrating vertical density differences.
Source: [30]

Relaxing this vertical constraint leads to horizontal density differences which are explored in this study. Consider the concept *Gas gangrene* in Figure 3.2. It is present in both SNOMED CT and MEDCIN. There are 10 child concepts common between the two ontologies, of which three are shown in Figure 3.2. The concept *puerperal gangrene gas* in MEDCIN does not exist as a child of *Gas gangrene* in SNOMED CT. This difference in the sets of child concepts for identical parent concepts is called horizontal density difference. This leads to the possibility that *puerperal gangrene gas* is missing in SNOMED CT and could be imported into it. Every ontology has its own policies concerning what to include and what to omit. However, the possibility exists that this concept would be included by the SNOMED CT curators if they were previously not aware of the fact that it is potentially missing and were later informed that it might be.

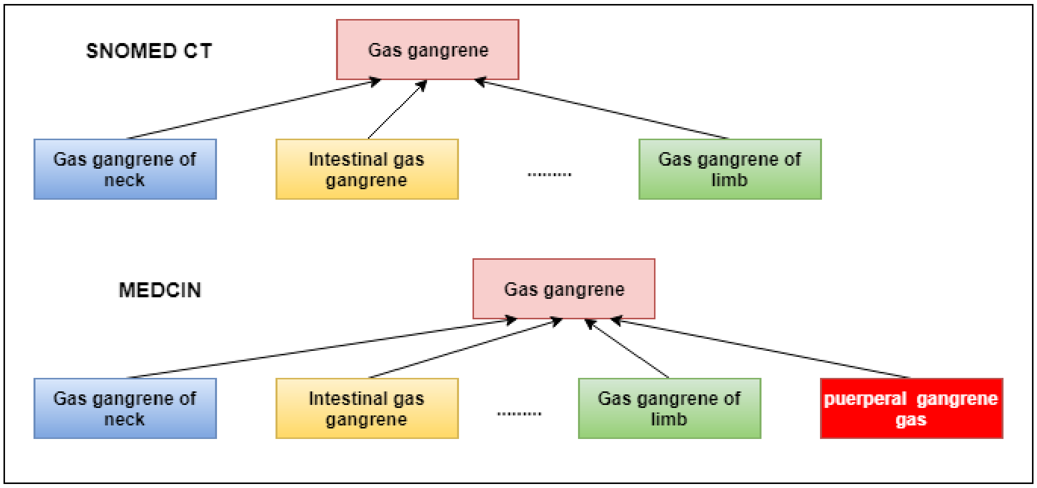


Figure 3.2 An example demonstrating horizontal density differences in SNOMED CT and MEDCIN. The arrow represents IS-A hierarchical relationship.
Source: [30]

With an understanding of the difference between vertical and horizontal density differences this study asks the question - Given a concept occurring in two different ontologies with two different but overlapping sets of children, does this indicate the possibility of importing children from one ontology into the other? A quantitative analysis of the phenomenon of horizontal density differences in 10 ontologies from the UMLS Metathesaurus is presented.

3.1.1 Identifying Source Ontologies

The UMLS is used to identify source ontologies that could potentially supply missing child concepts to the two target ontologies SNOMED CT and NCIt. Two criteria are applied for selecting the source ontologies – 1) the ontology should be in the English language; 2) the ontology should have a PAR relationship (parent relationship in a Metathesaurus source vocabulary) with an *inverse_isa* annotation (additional relationship). There are 10 such ontologies in the 2017 AA release of the UMLS in addition to SNOMED CT and NCIt, namely the Anatomical Therapeutic Chemical Classification System (ATC) [102], Medical

Entities Dictionary (CPM) [103], Current Procedural Terminology (CPT) [7], Foundational Model of Anatomy Ontology (FMA) [104], Gene Ontology (GO) [22], Human Phenotype Ontology (HPO) [105], MEDCIN [24], the Veterinary Extension to SNOMED CT (SNOMEDCT_VET) [106], Universal Medical Device Nomenclature System (UMD) and University of Washington Digital Anatomist (UWDA). Two ontologies were excluded from the study – SNOMEDCT_VET, and UWDA. SNOMEDCT_VET was developed and is maintained to extend SNOMED CT with additional animal-specific content [55] and UWDA consists of only some selected components of FMA and hence it was also not considered, to avoid overlap. When SNOMED CT is the target ontology, then NCIt is joined to the group of source ontologies and vice versa.

3.1.2 Finding Potentially Missing Child Concepts

To find potentially missing child concepts, a process needs to be developed that will alert ontology curators concerning concepts that they might want to import into their ontologies, as in the example in Figure 3.2, where the SNOMED CT curator might be interested in importing *puerperal gangrene gas* from MEDCIN. For this purpose, there is a need to

1. establish that there are sufficiently many ontology pairs with adequate numbers of pairs of identical concepts to make this endeavor practically useful.

2. develop an algorithm that will automatically identify horizontal density differences for finding appropriate concepts to import.

3. verify with a human expert that concepts suggested for import, at least in principle, qualify, independent of the concept inclusion policies of the ontology curators.

The case for imports is not always clear-cut. For example, if a concept P has 20 children in ontology T_1 and 18 children in T_2, such that all 18 are also children of P in T_1

then it would be highly likely that the two remaining child concepts could or even should be imported into T2. On the other hand, if the concept P has 20 children in ontology T1 and 20 children in T2 and only one of the concepts is common, then the remaining 19 child concepts should almost certainly not be suggested for import. The question arises then how to deal with intermediate cases between these two extreme examples.

To provide a rational method to decide when concepts should be suggested for import, a "similarity metric," based on the number of identical children and the total number of children is used. A curator could be guided by a threshold value of similarity. Child concepts with a value of the metric that falls below the threshold would not be considered for import, and vice versa. Figure 3.3 demonstrates the basic approach. In Figure 3.3(a), the concept P exists in the ontologies T1 and T2. There are two common concepts, A and B, and an additional concept C in T1. One can view the two concepts A and B as "two votes" that P is meant to represent the same knowledge in T1 and T2. That indicates that C should be proposed as a concept for import into T2. On the other hand, in Figure 3.3(b) there is only one "vote" and the likelihood that C or B should be imported into T2 is greatly reduced.

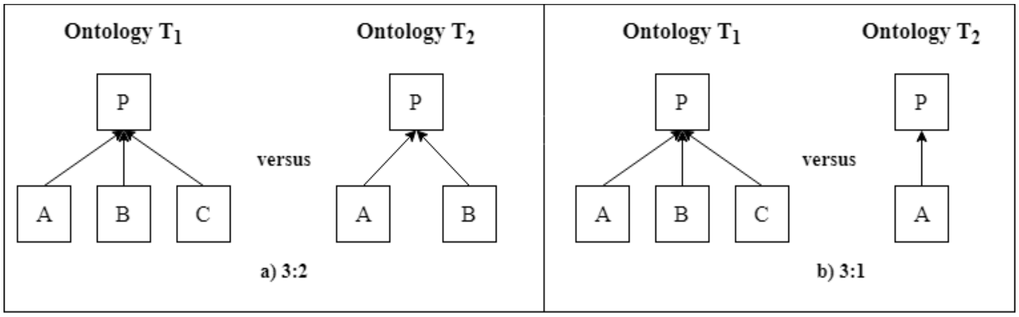


Figure 3.3 Ratio of similarity between identical concepts in ontology T1 and T2. **a)** Relatively stronger evidence for importing into T2; **b)** Relatively weaker evidence for import into T2.
Source: [30]

Thus, the decision of whether the concepts should be imported depends on two variables, the number of common children and the total number of children in the source ontology. The closer the number of common children is to the total number of children the more evidence there appears to be that the children that are not common should be imported. This can be expressed as the requirement that the ratio of these two variables should be as close (but not equal) to 1, as possible. This is expressed by the following metric.

Let $|B|$ denote the number of children that are common in both the ontologies and let $|C_o|$ denote the number of children in the source ontology. Then the ratio of these two variables is given by

$$J = \frac{|B|}{|C\text{o}|} \tag{3.1}$$

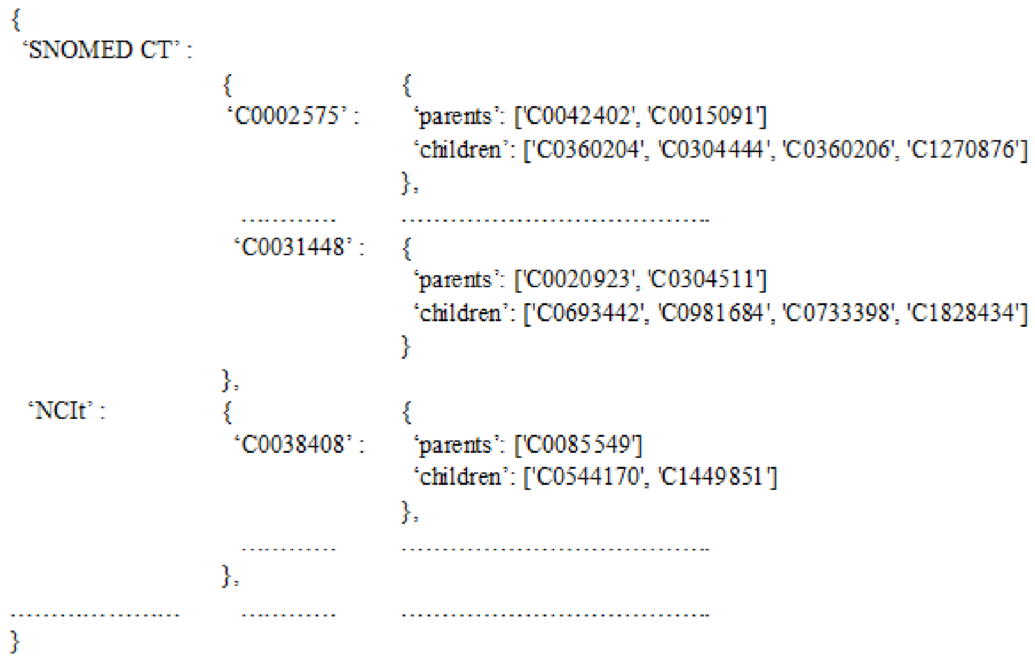
When *J* becomes larger than a threshold, then import should be recommended. The question remains how to choose the threshold (τ) value. If τ is chosen too large (too close to 1) there will be too few concepts that are potentially imported, and the yield of the method will be low. If τ is chosen too small, then there is an increased risk of importing concepts that are not correct in the context of the target ontology and additional work for the curator is caused.

To overcome this issue, sorting all the available parent concepts (that exist in both ontologies) by the decreasing value of *J* is recommend. If the lowest value of *J* is equal to or below ½ (coin flip), then it is doubtful whether any concepts should be imported. Thus, τ must be chosen between ½ and 1 in a way that balances the potential import yield and the risk. In any event, the algorithm only proposes concepts for import. A curator of the ontology will need to approve every single one of them. Moreover, it makes sense to present parent concepts in a list sorted in decreasing order of *J* to the curator because the highly likely imports will be at the beginning of the list.

3.1.3 Algorithms

In this section, the algorithm for computing *J* is presented. The algorithm is implemented in two parts. The first part creates a structure named ontDAG which is used by the second part to compute *J*. First, a brief explanation of the notations used in the algorithm. Curly brackets {} are used to denote a dictionary. Dictionaries are unordered key-value pairs, where each key is mapped to a value. If there are several levels of dictionaries, then access is done from left to right. Thus, in Figure 3.4, ontDAG{ontName} returns the sub-dictionary with the key ontName. In practice, ontDAG{'SNOMED CT'} would derive a sub-dictionary with all data stored about SNOMED CT concepts. Square brackets [] denote a set. Empty dictionaries appear as {} and empty sets are written as [].

Algorithm 1 creates a multi-level dictionary structure, as shown in Figure 3.4, which stores the ten ontologies from the UMLS (mentioned in Section 3.1.1) and the concepts in the ontologies that have a parent-child relationship. For each concept, a list of its parent concepts and a list of its child concepts are stored. The dictionary structure facilitates easy retrieval of any concept in any of the ontologies. For each concept, it is also possible to easily retrieve its parents and its children or descendants up to any level. The ontology DAG that is created for this study takes into consideration only (English) concepts with PAR/inverse_isa relationships, but the underlying structure is suitable for any relationship that forms a hierarchy.



```
{
 'SNOMED CT' :
                  {               {
                   'C0002575' :    'parents': ['C0042402', 'C0015091']
                                   'children': ['C0360204', 'C0304444', 'C0360206', 'C1270876']
                                  },
                    ............  ........................................
                    'C0031448' :  {
                                   'parents': ['C0020923', 'C0304511']
                                   'children': ['C0693442', 'C0981684', 'C0733398', 'C1828434']
                                  }
                  },
  'NCIt' :        {               {
                    'C0038408' :   'parents': ['C0085549']
                                   'children': ['C0544170', 'C1449851']
                                  },
                    ............  ........................................
                  },
....................    ............  ........................................
}
```

Figure 3.4 Example of content in the ontDAG dictionary. Within SNOMED CT, the concept C0002575 has the parents C0042402 and C0015091 and the children C0360204, C0304444, etc.
Source: [30]

The input D to Algorithm 1 is a collection of rows derived from a reduced UMLS Metathesaurus relational database representation, containing rows of data values (*chid*, *pid*, *ont*), where *chid* is the child concept id, *pid* is the parent concept id and *ont* is the ontology in which the relationship is present. There are two issues to deal with when we construct the ontology DAG (ontDAG). The first one is the well-documented presence of loops (cycles) and self-loops of IS-A links in the UMLS [107, 108]. This happens when the child concept id and the parent concept id are the same. In the UMLS tables, these instances occur due to the differences in the atom unique identifiers (AUI) for the same concept unique identifier (CUI). The second issue is that of multiple edges between the same parent and child concepts. This is again due to differences in the AUI's of the parent and child concepts.

For removing the cycles, an adaptation of the "naïve" (by their own appellation) approach to eliminating cycles by Mougin and Bodenreider [109] was used. This approach performs a depth-first search of the Metathesaurus graph and marks nodes as visited to detect loops. This approach was adapted by using only concepts that participate in an IS-A relationship (PAR, inverse_isa) in the 10 terminologies used in the study, instead of all the hierarchical relationships in the Metathesaurus, and also depth was limited to a maximum of five levels instead of the 50 levels of Mougin and Bodenreider [109], as the patterns described in all the studies in this dissertation would never go beyond five levels for any concept.

Algorithm 1 Build Ontology DAG from UMLS RRF Table

```
procedure ONTOLOGY-DAG(in: $\mathcal{D}$, out: $ontDAG$)
    $ontDAG \leftarrow \{\ \}$
    for each ($chid, pid, ont$ ) in $\mathcal{D}$ do
        if $ont$ not-in $ontDAG$ then
            $ontDAG\{ont\} \leftarrow \{\ \}$                # sub-dictionary created for a new ontology
        end if

        if $chid$ not-in $ontDAG\{ont\}$ then
            $ontDAG\{ont\}\{chid\}\{'parents'\} \leftarrow [\ ]$      # initialize empty sets for storing parent
            $ontDAG\{ont\}\{chid\}\{'children'\} \leftarrow [\ ]$          # and child concepts
            if $chid \neq pid$ then
                $ontDAG\{ont\}\{chid\}\{'parents'\} \leftarrow [pid]$      # add parent concept id
            end if
        else                                       # check for multiple edges between par-child
            if $pid$ not-in $ontDAG\{ont\}\{chid\}\{'parents'\}$ and $chid \neq pid$ then
                $ontDAG\{ont\}\{chid\}\{'parents'\} \leftarrow ontDAG\{ont\}\{chid\}\{'parents'\} \cup [pid]$
            end if
        end if

        if $pid$ not-in $ontDAG\{ont\}$ then
            $ontDAG\{ont\}\{pid\}\{'parents'\} \leftarrow [\ ]$       # initialize empty sets for storing parent
            $ontDAG\{ont\}\{pid\}\{'children'\} \leftarrow [\ ]$           # and child concepts
            if $pid \neq chid$ then
                $ontDAG\{ont\}\{pid\}\{'children'\} \leftarrow [chid]$       # add child concept id
            end if
        else                                       # check for multiple edges between par-child
            if $chid$ not-in $ontDAG\{ont\}\{pid\}\{'children'\}$ and $pid \neq chid$ then
                $ontDAG\{ont\}\{pid\}\{'children'\} \leftarrow ontDAG\{ont\}\{pid\}\{'children'\} \cup [chid]$
            end if
        end if
    end for
end procedure
```

Algorithm 2 takes as input the ontology DAG (ontDAG which is the output of Algorithm 1), the target ontology (SNOMED CT or NCIt denoted by ρ), and the threshold value (τ). The output is a file with the parent concepts and all missing child concepts in the target ontology. The set of all ontologies (both source and target) is represented by O. C_ρ and C_o represents the set of child concepts for a parent concept in the target and source ontology, respectively. Even though only enriching the conceptual content of SNOMED CT and NCIt is described in this study, the algorithm can be trivially extended to other Metathesaurus ontologies as a target ontology.

Algorithm 2 Find children of a concept that do not occur in the target ontology but occur in a source ontology

```
procedure MISSING-CHILDREN(in: $ontDAG, \rho, \tau$ out: file (contains pid (s) and all missing chid(s))
$O \leftarrow$ [ontology_1, ontology_2…]                # Set of all ontologies in ontDAG
for concept in $ontDAG\{\rho\}$ do
  $\mathcal{C}_\rho \leftarrow ontDAG\{\rho\}\{concept\}\{'children'\}$          # Set of all children of concept in $\rho$
  for $o$ in $O - [\rho]$ do
    if concept in $ontDAG\{o\}$ then
      $\mathcal{C}_o \leftarrow ontDAG\{o\}\{concept\}\{'children'\}$
    end if
    $\mathcal{M}_\rho \leftarrow \mathcal{C}_o - \mathcal{C}_\rho$            # missing concepts in $\rho$
    $\mathcal{B} \leftarrow \mathcal{C}_o \cap \mathcal{C}_\rho$            # child concepts in both in $\rho$ and source ontology
    $\mathcal{J} \leftarrow |\mathcal{B}| \div |\mathcal{C}_o|$            # Calculate similarity metric for the sets of children
    if $\mathcal{J} > \tau$ :                # Metric greater than threshold value
      Output : concept, $\mathcal{M}_\rho$ (write to a file)
  end for
end for
end procedure
```

3.2 Randomized Controlled Trial (RCT)

The concepts proposed for import are presented to a domain expert with long work experience in ontology quality assurance. To assure the validity of the tests and reliability of the results, two randomized controlled trials with SNOMED CT and two more with NCIt are performed. This section describes the process of selecting the two control samples for

SNOMED CT. The same procedure was adopted for NCIt. Figure 3.5 elucidates the experimental group and the two control groups.

Group 1 consists of concepts that are children of concept P in the target ontology. For every parent concept P in the sample, one child concept C_1 was removed. When the pair (P, C_1) is presented to the domain expert, she should determine that an IS-A link between C_1 and P should exist. In Figure 3.5, concepts of Group 1 are shown with red dashed outlines. The IS-A link is shown in green, to indicate that it should be reported as a child concept by the domain expert because these concepts are actually children of P in the target ontology. This is control group 1.

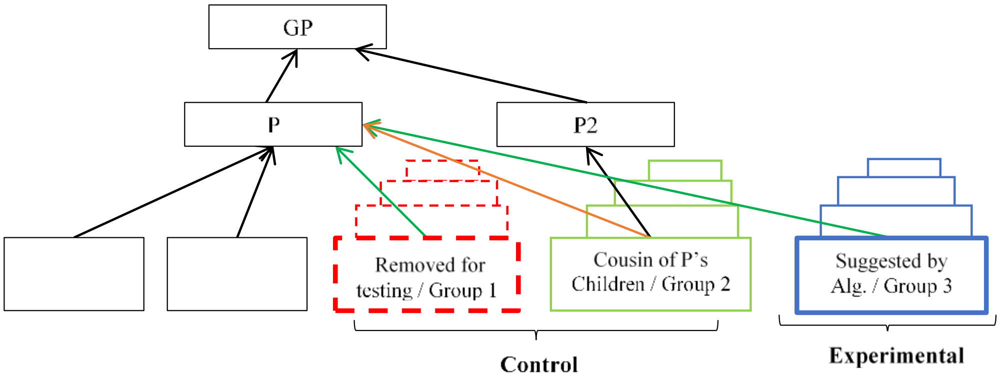

GP
P
P2
Removed for testing / Group 1
Cousin of P's Children / Group 2
Suggested by Alg. / Group 3
Control
Experimental

Figure 3.5 Three groups of concepts for RCT. GP denotes grandparent and P and P2 denote parent concepts.
Source: [30]

Group 2 consists of concepts that are cousins of P's children. In Figure 3.5, P2 is a sibling of P. The children of P2 are cousins of the children of P. Furthermore, it was ascertained that none of the selected children of P2 are also children of P (which is theoretically possible). For every parent concept P in the study, one such concept was selected as cousin C_2. When the pair (P, C_2) is presented to the domain expert, she should determine that no IS-A link between C_2 and P exists, because C_2 appears in the target ontology but is not hierarchically related to P. In Figure 3.5, concepts of Group 2 are shown with green solid outlines. The IS-A links in black accurately show the hierarchical position of C_2. The IS-A link in red should be reported by the domain expert as "not a child concept." This is control group 2.

Group 3 consists of concepts that were suggested by the algorithm as possible children of P in the target ontology because they are children of P in one of the source ontologies. When the pair (P, C_3) is presented to the domain expert she is expected to report that an IS-A link between C_3 and P should exist, which implies that C_3 should be imported into the target ontology. In Figure 3.5, concepts of Group 3 are shown in solid blue. The IS-A link is shown in green, to indicate that it should be reported by the domain expert as a valid child concept if the algorithm works correctly with the source ontologies. This is the experimental group.

Hypothesis 3.1: Concepts from Group 2 and Group 3 will be distinguishable with statistical significance.

The domain expert should be clearly able to distinguish between IS-A links suggested by the source ontology and IS-A links that should not be there.

Hypothesis 3.2: Concepts from Group 1 and Group 3 will be indistinguishable with statistical significance.

The domain expert should not be able to distinguish between concepts that are already in the target ontology and concepts that are suggested by the algorithm for import into the target ontology from the group of source ontologies.

3.3 Results

The parent concepts in SNOMED CT and NCIt were sorted based on the value of *J* computed by the algorithm. Table 3.1 shows the numbers of parent concepts obtained for different ranges of *J* values for both SNOMED CT and NCIt. The numbers of concepts with $J = 1$ and $J = 0$ are not reported.

Table 3.1 Number of Parent Concepts for Different Ranges of *J* in SNOMED CT and NCIt

Range of *J*	**SNOMED CT**	**NCIt**
[0.95, 1.0)	3 concepts	0 concepts
[0.90, 0.95)	34 concepts	3 concepts
[0.80, 0.90)	322 concepts	42 concepts
[0.70, 0.80)	615 concepts	59 concepts
[0.50, 0.70)	2425 concepts	580 concepts
(0.00, 0.50)	3700 concepts	2644 concepts

It was found that in SNOMED CT there were only three parents with *J* between 0.95 and 1. This resulted in too small a sample to be valid for a domain expert study for import, and therefore the threshold was lowered to $J = 0.9$. With a *J* value ≥ 0.9 and < 1 there are 37 concepts. In the case of NCIt, choosing a *J* value of 0.9 would result in only 3 concepts proposed for import, hence, the threshold was lowered to $J \geq 0.8$ in this case, resulting in 45 parent concepts for a domain expert study.

For SNOMED CT, a sample of 37 concepts each in Group 1, 2, and 3 giving a total of 111 concepts were shuffled to avoid bias and presented as the first sample to the domain expert. For NCIt, the same process was followed as with SNOMED CT, but in this case, there are 45 concepts each in the three groups yielding 135 concepts as the second sample. The samples for SNOMED CT and NCIt were analyzed separately, and both the hypotheses were tested for each sample.

SNOMED CT: Out of the 37 potentially missing child concepts proposed by the algorithm (Group 3) for SNOMED CT, 22 concepts (roughly 60%) were agreed upon as missing by the domain expert. For Group 1, which consisted of concepts that were already children of the parent concept in both the ontologies, the domain expert disagreed with seven concepts. This basically points to some errors during the modeling of the ontologies. Regarding the second control group which consisted of cousins and not children, the domain expert disagreed with the majority of concepts as expected, but in four cases the domain expert found that the concept could potentially be a child. The results for both the control and test samples are shown in Table 3.2.

Table 3.2 Results of each Group as Analyzed by the Domain Expert for SNOMED CT

Groups	Number of concepts agreed for import by the domain expert	Number of concepts disagreed for import by the domain expert
Group 1	30	7
Group 2	4	33
Group 3	22	15

Fisher’s exact test (two-tailed) was performed on each control sample and test sample for SNOMED CT. For Hypothesis 3.1 a p-value < 0.0001 was obtained. Hypothesis 3.1 was supported with statistical significance. Thus, the concepts proposed by the algorithm as children were distinguishable from known "non-children" with statistical significance.

For Hypothesis 3.2 a p-value of 0.0738 was obtained. Because p > 0.05, it is not possible to conclude that concepts from Group 1 and Group 3 are from distinguishable groups. However, this does not prove that they are from the same group. Quoting David Howell, "We know that we cannot conclude from a non-significant difference that we have proved that the mean of a population of scores … is the same as the mean … of control subjects" [110]. Larger sample sizes, possibly in connection with a different statistical method, will be needed to get support for Hypothesis 3.2.

Table 3.3 Examples of Missing Child Concepts in SNOMED CT Identified by the Algorithm and Verified to be Missing by the Domain Expert

Parent concept	Missing child concept
C0004661: Bacteroides	C2614420: Bacteroides xylanisolvens
C0018805: Heart Injuries	C2836204: heart injury without hemopericardium
C0154409: Recurrent major depressive episodes	C2062772: recurrent major depression without melancholia

NCIt: For NCIt, 25 concepts (roughly 56%) were agreed upon as missing by the domain expert out of the total 45 concepts for Group 3. For Group 1 only three concepts were disagreed with by the domain expert and for Group 2 (control group), one concept which is a cousin was found to be a potential child concept by the expert. Table 3.4 shows the results for both the control and experimental groups.

Table 3.4 Results of each Group as Analyzed by the Domain Expert for NCIt

Groups	Number of concepts agreed for import by the domain expert	Number of concepts disagreed for import by the domain expert
Group 1	42	3
Group 2	1	44
Group 3	25	20

For Hypothesis 3.1 a p-value < 0.0001 was obtained. Hypothesis 3.1 was supported with statistical significance. Hence, for NCIt, the concepts proposed by the algorithm as children were distinguishable from known "non-children" with statistical significance. For Hypothesis 3.2 a p-value < 0.0001 was obtained. Hypothesis 3.2 was not supported for the sample. Thus, the existing child concepts and the algorithmically suggested child concepts were not indistinguishable.

Table 3.5 Examples of Missing Child Concepts in NCIt Identified by the Algorithm and Verified to be Missing by the Domain Expert

Parent concept	**Missing child concept**
C0010266: Cranial nerve diseases	C0154733: Multiple cranial nerve palsy
C0042672: Vinca Alkaloids	C0059752: Vinpocetine
C0012611: Disaccharides	C0007630: Cellobiose

3.4 Discussion

The reasons why the expert did not agree to import some concepts suggested by the algorithm were analyzed. For SNOMED CT, the number of rejected concepts was 15 and for NCIt it was 20. Different reasons suggested for not agreeing to import the child concept are explained below with an example.

1. The algorithmically suggested concept was observed to fit as a grandchild instead of a child concept. Such scenarios indicate the finer level of detail captured by the target ontology. There were six concepts in SNOMED CT and two concepts in NCIt falling into this category.

For example, consider a parent concept *Intensive Care Unit*. The missing concept suggested by the source ontology is *Intensive Care Unit, Neonatal*. There exists a concept *Pediatric Intensive Care Unit* in the target ontology as the child of *Intensive Care Unit.* The missing concept *Intensive Care Unit, Neonatal* would qualify better as a child of *Pediatric Intensive Care Unit* than *Intensive Care Unit*.

2. The algorithm suggested a concept that was found to be a synonym of the already existing child and was not recorded as a synonym (SY) in UMLS. Eight concepts from NCIt and one concept from SNOMED CT belonged to this case.

The concept *Cervical Vertebrae* has children *C1 Vertebra*, *C2 Vertebra*, etc. in the target ontology. The concept suggested by the algorithm is *Axis Vertebra* and according to the domain expert, it is a synonym of *C2 Vertebra*.

3. The algorithmically suggested concept is a combination of two concepts that exist separately in the target ontology. There was one concept each from SNOMED CT and NCIt reported in this category.

The concept in the source ontology that was suggested as missing in the target ontology by the algorithm is *Spina bifida of thoracolumbar region*. This concept is a combination of two concepts existing in the target ontology namely *Spina bifida of lumbar region* and *Thoracic Spina bifida.*

4. The algorithmically suggested concept could possibly be a child concept but then the already existing children should be moved one level down to become the children of the suggested concept.

The concept *Thoracic spinal ganglion* in the target ontology has children *T1 spinal ganglion, …., T12 spinal ganglion*. The algorithmically suggested concept *Variant thoracic spinal ganglion* could be a child of *Thoracic spinal ganglion* but then *T1 spinal ganglion, …., T12 spinal ganglio*n should become the children of *Variant thoracic spinal ganglion*.

If a concept is reported as missing by the algorithm that does not imply that it should be imported automatically. In many cases, curators of ontologies do not include concepts on purpose. One reason for this is that such concepts have no known use case and would "clutter up the ontology" and make maintenance more difficult. Larger ontologies also result in slower responses of user-facing applications using those ontologies. In fact, some curators tend to include new concepts only if there is an explicit request from a user as was found out in extensive prior work on ontology quality assurance.

Another possible reason for rejecting the import of a concept could be that it is at a too fine-grained level of detail, which contradicts the overall philosophy behind the design of the ontology. However, these design philosophies are rarely made available to researchers outside of the organization and in many cases appear not to exist in a written format. The final decision concerning an import always must be made by an expert, who has to balance factors of completeness against efficiency, usability, and maintainability.

CHAPTER 4

ALTERNATIVE CLASSIFICATION OF IDENTICAL CONCEPTS

The study in Chapter 3 discussed how horizontal density differences between pairs of terminologies could be used to identify missing child concepts in the target terminology. It was noted that pairs of terminologies with identical concepts do not always have the same set of children in the two terminologies. The number of children was found to vary widely, and a special situation was identified where the children in one terminology relate to the common parent in a very different way than the children in the other terminology. For example, children in one terminology might subdivide a parent concept by anatomical location in one terminology and by disease kind in the other terminology. We coined the term "alternative classification" (of the same parent concept) for such situations. In previous work, only human experts could recognize alternative classifications. In this study, we present a mathematically expressed criterion for likely cases of alternative classifications. Besides alternative classifications, common parent concepts in a pair of terminologies might also indicate a possible import of a child concept missing in one terminology, different granularities, or errors in either one of the two terminologies. We also investigate different kinds of alternative classifications.

4.1 Identification of Alternative Classifications

While conducting the study described in Chapter 3, observations were made where a concept existing in two terminologies has children in both, yet these sets of children do not have a single child in common. In such cases, the two concepts are actually quite different in the two terminologies, even though they have the same CUI. This extreme difference is a possible indicator of an "alternative classification" or an error. While the discovery of such an alternative classification candidate or error candidate should be made based on a formula/algorithm, the final interpretation and decision always must be made by a human expert.

Figure 4.1 exemplifies a case where the two occurrences of the same concept in two different terminologies have no common children. The concept *Benign Nasopharyngeal Neoplasm* has three children in NCIt and four children in MEDCIN. The three child concepts present in NCIt do not exist anywhere in MEDCIN and similarly, the four child concepts in MEDCIN do not exist anywhere in NCIt.

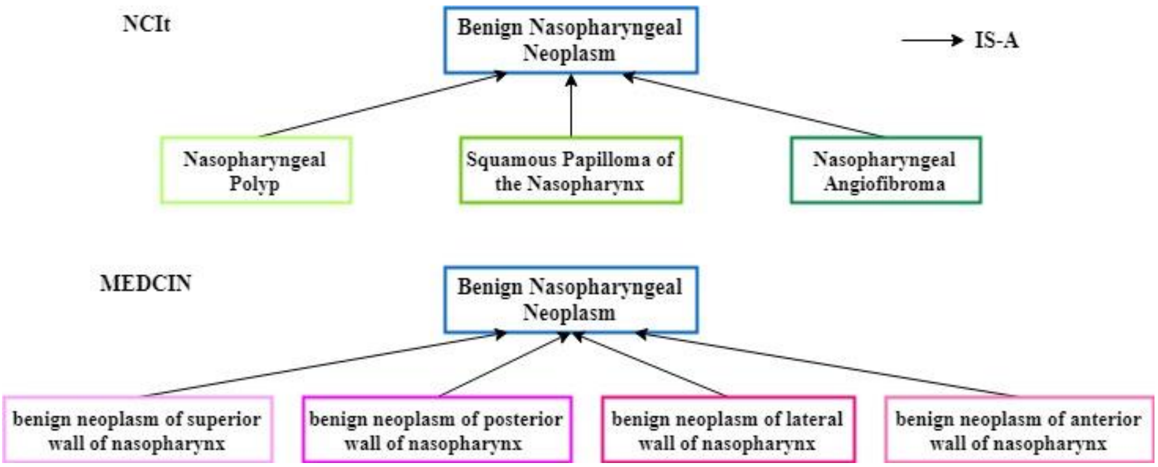

NCIt
Benign Nasopharyngeal Neoplasm
IS-A
Nasopharyngeal Polyp
Squamous Papilloma of the Nasopharynx
Nasopharyngeal Angiofibroma
MEDCIN
Benign Nasopharyngeal Neoplasm
benign neoplasm of superior wall of nasopharynx
benign neoplasm of posterior wall of nasopharynx
benign neoplasm of lateral wall of nasopharynx
benign neoplasm of anterior wall of nasopharynx

Figure 4.1 An example of alternative classification in NCIt and MEDCIN.
Source: [31]

The three child concepts in NCIt specialize the parent concept *Benign Nasopharyngeal Neoplasm* by disease kind (polyp, squamous papilloma, and angiofibroma), whereas the four child concepts in MEDCIN specialize *Benign Nasopharyngeal Neoplasm* by anatomical location (superior wall, posterior wall, etc.).

These two different classifications, which arise because of the difference in the modeling philosophies of the two terminologies, constitute what we call an alternative classification. Although these two hierarchies are individually correct, merging them in a naïve way would lead to a loss of structural information.

The focus of this study is to 1) develop a method to identify highly likely cases of alternative classifications algorithmically; 2) develop a metric that identifies concepts with children that are highly likely to be proposed for import from one terminology into the other; 3) compare how different ranges of the metric might affect the number of concepts that should be considered for import; 4) analyze different cases of alternative classifications in more detail.

4.1.1 Terminology Selection

The first step is identifying terminologies in the UMLS that can be used as source terminologies to supply possible concepts for import into target terminologies. For this the following criteria are used:

1. Only English terminologies could be processed, due to our linguistic limitations.

2. Only terminologies with an IS-A backbone support the hierarchy-based methodology. A necessary but not sufficient condition for this is that the terminology uses the UMLS PAR(ent) relationships.

3. PAR relationships are not always IS-A relationships, thus as an additional condition, only PAR relationships that are marked with inverse_isa annotations, guarantees that IS-A relationships are expressed.

4. If two terminologies have substantial overlap due to common ancestry and/or common domain, then only one of them will be chosen.

5. The chosen terminologies should not have been the subject of our previous investigation on horizontal density (described in Chapter 3).

6. The research is oriented towards terms used in cancer care.

7. As pairs of terminologies are being processed, interesting results can only be expected if there is substantial overlap between the two terminologies of a pair.

It should be noted that criteria 1-4 were applied for identifying source ontologies (Section 3.1.1) in Chapter 3. As a result of this, eight ontologies were obtained in addition to NCIt and SNOMED CT. SNOMED CT was used extensively in the study described in Chapter 3 [30], thus it is excluded by the fifth criterion. But by the sixth criterion NCIt is included as the target terminology of this study as it is a terminology for cancer research. Together with the overlap requirement of the seventh criterion, that leaves only the pair of MEDCIN and NCIt. Thus, in this study, NCIt is our target terminology and MEDCIN is the source terminology.

4.1.2 Algorithm and Revised Metric

The algorithm described in Section 3.1.3 explored horizontal density differences to suggest child concepts that are highly likely to be imported. For the algorithm for computing alternative classification criteria with NCIt as the target and MEDCIN as the source terminology, some modifications were made to the previous algorithm. It is possible that one or more of the child concepts present in the source terminology but missing from the target terminology as immediate children could exist somewhere else in the target terminology. To overcome this issue, it was confirmed that each of the identified missing child concepts does not exist anywhere else in the target terminology by performing a search on the target terminology’s sub dictionary. The metric *J* was completely revised to cover the different cases (alternative classifications, concept import, and errors) more precisely, based on the ranges of the values of the metric.

As a first step, the algorithm identifies all common concepts in NCIt and MEDCIN and then finds all the children of these common concepts in both the terminologies. In the next step, all the children that are common in both the terminologies for each parent concept are identified along with the child concepts present in MEDCIN but missing from NCIt. For each of the missing child concepts identified, the algorithm makes sure that the concept is not present anywhere else in the target terminology. If the concept is present anywhere else, it is removed from the list of missing child concepts. The final step is to develop a formula that computes the evidence for import (EFI). The goal is to distinguish between likely cases of alternative classifications versus predictions when concepts should be imported directly (without further consideration) from the source terminology into the target terminology.

Consider a concept *P* that is the same in the source and target terminologies, by the authority of the UMLS Concept Unique Identifier (CUI). Intuitively, the more common children P_{source} and P_{target} have, the higher the evidence that the remaining children that are not common should be imported. Thus, initially

$$\widehat{EFI} = \frac{C_{common}}{C_{source}}, C_{Source} > 0 \tag{4.1}$$

C_{common} is the number of concepts that are children of *P* in both the source and the target terminology, and C_{source} is the number of concepts that are children of *P* in the source terminology. Note that $\widehat{EFI}$ is a number between 0 and 1.

For example, if there are 11 children in the source and eight common children that is higher evidence than if there are 11 children in the source and only five common children. In other words, the ratio of the number of common children and source children needs to be computed, assuming the parent is the same in both terminologies.

When the number of common children becomes "almost" equal to the number of source children, then the evidence should become "almost" 1. When the number of common children becomes equal to the number of source children, then the evidence becomes 1, but then there are no concepts left to import. Thus, this is an uninteresting case; and hence, an assumption is made that the number of common children is strictly smaller than the number of source children.

$$C_{common} < C_{source} \tag{4.2}$$

However, the evidence for importing concepts should also be based on the total number of children, not just on the ratio. For example, if there are 70 common children out of 110 source children, this should provide stronger evidence than if there are only seven common children out of 11 source children, even though the ratio would be the same. Therefore, an additional corrective factor *(F)* is required.

Let the parameter *MAX*, represent the number of children of that parent P that has the most children in the source. (There might be several such parents with equally high numbers of children.) Let $\#C(P_i, S)$ be the number of children of the parent P_i in the source terminology *S*, then,

$$MAX = \#C(P, S)\ such\ that\ \forall i\ \ \#C(P, S) \geq \#C(P_i, S) \tag{4.3}$$

$$\frac{MAX - \ C_{common}}{MAX} \tag{4.4}$$

Equation 4.4 will become smaller when there are more children in common between source and target terminology. But the evidence for import should increase when there are many common children. Hence, the corrective factor *F* is computed as

$$F = 1 - \frac{MAX - \ C_{common}}{MAX} = \frac{C_{common}}{MAX} \tag{4.5}$$

The problem is that this corrective factor *F* becomes too small for terminologies containing concepts with many children. As *F* is between 0 and 1, a convenient way to make it larger but keep it in the same range [0,1] is to apply the square root to it. Thus, applying the square root to *F* (Equation 4.5) and multiplying with Equation (4.1), and observing all the boundary conditions the evidence for import (EFI) is calculated as:

$$EFI = \begin{cases} \widehat{EFI} * \sqrt{\frac{C_{common}}{MAX}} \ \ when \ C_{common} < C_{source} \ and \ \ C_{source} > 0 \\ \qquad\qquad undefined \quad otherwise \end{cases} \tag{4.6}$$

By the above definition, the value of $\widehat{EFI}$ is in the interval [0,1). The corrective factor *F* can only become 1 if $C_{common} = C_{source}$, which was excluded; and therefore, F is also in the range [0,1). Thus, the value of *EFI* will always be in the range [0,1).

In the extreme case, when there are no common concepts, *EFI* becomes zero. As noted above, when there are no children in common, this appears to be an indicator of an alternative classification or an error. In cases when *EFI* is relatively closer to 0 it is more likely to indicate a mix of imports, alternative classifications, and errors. When the value of *EFI* becomes relatively closer to 1, it most likely indicates a possible import.

4.1.3 Sample Preparation

The values of $\widehat{EFI}$ and EFI were computed for all concepts present in both NCIt and MEDCIN. A total of 1049 identical parent concepts in both MEDCIN and NCIt did not have any overlapping children and hence had their EFI = 0. The remaining 917 identical parent concepts had an EFI value >0 and <1.

Sample 1: Sample 1 consisted of 50 parent concepts randomly chosen from those parent concepts that have no common children between the two terminologies (i.e., $C_{common} = 0$). For each of the 50 concepts in Sample 1 (i.e., with no common children) all the children of each parent concept (*P*) in both the source and target terminologies are listed for the domain expert to review. The following four choices were then presented to the domain expert.

1. The children in source and target terminologies form an alternative classification.

2. Error in the source terminology – one or more children in the source terminology should not be children of *P*.

3. Error in the target terminology – one or more children in the target terminology should not be children of *P*.

4. A case of a finer level of detail in the source or target terminology – Suppose the parent concept (*P*) is *Thoracic spinal ganglion*. In the source terminology, the immediate child concepts are *T1 spinal ganglion, T2 spinal ganglion..., T12 spinal ganglion,* whereas in the target terminology the immediate child is *Variant thoracic spinal ganglion,* while *T1 spinal ganglion…, T12 spinal ganglion* are listed as children of *Variant thoracic spinal ganglion*. This is a case of a finer level of detail in the target terminology.

These options are not mutually exclusive. It is possible that for a parent concept *P* there is an error in both the source and target terminology. Thus, the domain expert was asked to consider more than one choice if necessary.

Sample 2: For sample 2, the parent concepts were ordered according to their *EFI* values in decreasing order and the top 50 parent concepts were selected. One missing child concept was randomly selected for each of these 50 parent concepts and made part of the sample. A randomized controlled trial (RCT) was performed for Sample 2 similar to the one described in Section 3.2. For this, a control group was created with the same 50 parent concepts as in Sample 2. For each of these parent concepts (*P*) a sibling of *P* was found and one of its children was chosen, which therefore is a cousin of *P*'s children, and included in the control group. Thus, there were two groups for the RCT – the experimental group and the control group. Figure 4.2 shows this selection process for one parent concept *P*. The order of the concepts in the two groups was randomized. When the domain expert is presented with two concepts for each parent concept (one suggested by the algorithm and the other the cousin of *P*'s children; separated due to the randomization) s/he should be able to distinguish whether there should exist IS-A links between these two concepts and the parent concept *P*.

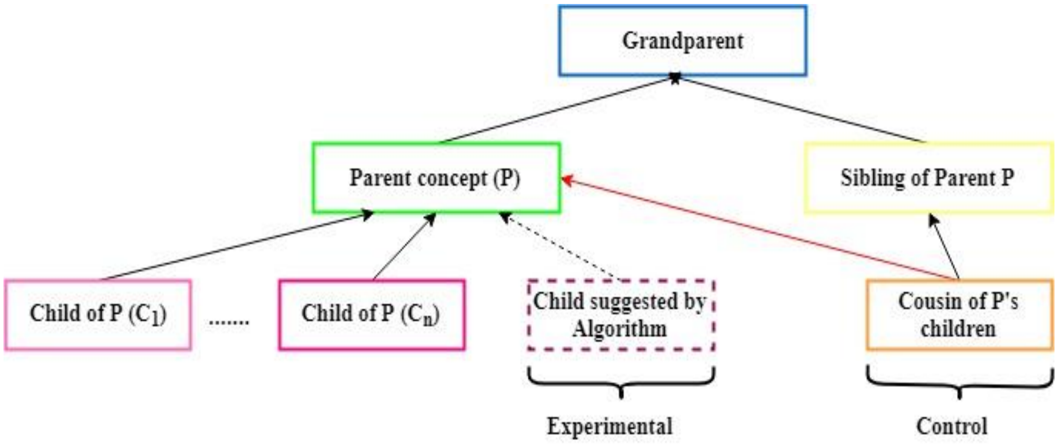


Figure 4.2 Selecting experimental and control group concepts for a parent concept *P*.
Source: [31]

Sample 3: From the ordered list of parent concepts, based on the EFI values in decreasing order, the bottom 50 concepts with an EFI value > 0 were selected. As before, one randomly chosen missing child concept suggested by the algorithm for each parent concept was added to create the sample.

It was observed that the number of common children in Sample 3 is small and in 47 of the cases, there is only one common child between the source and target terminology. The maximum number of common children in Sample 3 was four, which was observed in one case. Also, as discussed before, this sample is likely to contain concepts with errors or cases of alternative classifications. Taking this into account, the domain expert was presented with the following four choices for a parent concept *P*.

1. The suggested child concept should be imported into the target terminology.
2. The suggested child concept should not be imported into the target terminology.
3. Error in the source terminology – the suggested child concept should not be a child of the parent concept P.
4. The common children and the suggested child concept form an alternative classification.

As for Sample 2, an RCT was performed for Sample 3. Two groups were created – the experimental group and control group in the same way as for Sample 2 (Figure 4.2). For both Samples 2 and 3, the domain expert was also provided with all the common children from the source and target terminologies, separately for each parent concept, as part of the sample for a better understanding of the context.

Based on the previous work [30] and preliminary research, three hypotheses were formulated. In preliminary research, it was observed that for the case of no common children "many" parents defined alternative classifications. To quantify "many," there are two cases to be noted to make this observation practically applicable. The first one is distinguishing between a case of an absolute majority, where there are more alternative classifications than other choices. The second is a relative majority, where there are more alternative classifications than cases of the most common second category (error, import, or finer level of detail). Hypothesis 4.1 is formulated for the stronger case.

Hypothesis 4.1: For parent concepts with *EFI*=0 ($C_{common} = 0$), i.e., there are no overlapping children in the two terminologies for the same parent, it is more likely that the children define alternative classifications than the union of the other possible cases. (Possible other cases are error, import, or finer level of detail).

Hypothesis 4.2: Concepts proposed for import as children from a source terminology into a target terminology, based on their *EFI* values' proximity to 1, will be distinguishable with statistical significance from concepts that are known to be cousins and not children.

Note that a concept could be both a cousin and a child at the same time, thus the wording of the hypothesis excludes this case explicitly. The domain expert should be clearly able to distinguish between IS-A links suggested by the source terminology and IS-A links that should not be there.

Hypothesis 4.3: Concepts proposed for import as children from a source terminology into a target terminology, based on their *EFI* values' proximity to 0, will be distinguishable with statistical significance from concepts that are known to be cousins and not children.

4.2 Results

Sample 1: For Sample 1, with no overlapping children for a parent concept between both the terminologies, the domain expert was provided with four choices as discussed in

Section 4.1.3. As the choices were not mutually exclusive, multiple responses were received for each concept. Table 4.1 shows all the possible combinations of choices as analyzed by the domain expert. Overall, out of the 50 concepts, 36 (29+5+2) concepts were found to have children that are alternative classifications in the source and target terminology. Thus, 72% (=36/50) of concepts belonging to the category of alternative classification supports Hypothesis 4.1 that when there are no overlapping children in the two terminologies for the same parent, then it is more likely to be a case of alternative classification.

Table 4.1 Results for Sample 1 as Analyzed by the Domain Expert

Sample 1: *EFI*=0					
Alternative classification	29	Finer level of detail in source or target terminology	3	Error in source terminology + Error in target terminology	3
Error in source terminology	5	Alternative classification + Error in source terminology	5	Error in source terminology + Finer level of detail	1
Error in target terminology	1	Alternative classification + Error in target terminology	2	Error in target terminology + Finer level of detail	1

A review of alternative classifications in Sample 1 uncovered the following. The two parent concepts are identical in each case but are divided up according to different organizational viewpoints. There were four different categories identified.

1. One set of qualifiers (e.g., from the qualifier hierarchy of NCIt) is applied in one terminology, while no qualifiers are applied in the other terminology under the same parent.

An example would be the concept *Visual Impairment*. In one terminology the child concepts are a result of applying the spatial qualifiers left and right (i.e., *Visual impairment* *of right eye* and *Visual impairment of left eye*) whereas no qualifier is used in the other terminology (e.g., *Myopia*). There were five such cases, out of 36, in Sample 1.

2. One set of qualifiers is applied in one terminology, while another set of qualifiers is applied in the other terminology.

As an example, consider the concept *Irradiation of breast*. One terminology uses the qualifiers "left" and "right" (i.e., *Irradiation of left breast* and *Irradiation of right breast*) while the other terminology uses the qualifiers "whole" and "partial" (i.e., *Whole Breast Irradiation* and *Partial Breast Irradiation*). In Sample 1, out of 36 alternative classifications, this happened twice.

3. Different axes of classification are used in the two terminologies.

In the case of the concept *Cardiac Lipoma*, the children in one terminology are subcategorized based on the anatomical structure (e.g., *Epicardial Lipoma*), whereas the children in the other terminology are organized based on the histological finding (e.g., *Fibrolipoma of heart* and *Myelolipoma of heart*). There were 11 concepts of this type in Sample 1.

4. Combinations

In many cases, alternative classifications show that the original modeling of the two terminologies is not consistent, and therefore they do not cleanly fit into one of the first three categories. For example, the parent *Renal cyst* (C3887499) has the children C0268799: *acquired cyst of kidney*, C0268800: *simple renal cyst*, C0403383: *Infected renal cyst*, C0431718: *multiple renal cysts*, and C3812408: *congenital renal cyst* in the source terminology MEDCIN and the children C0521621: *Solitary Multilocular Kidney Cyst* and C4022836: *Solitary Cyst of Kidney* in the target terminology NCIt. Notably, "acquired," "simple," "multiple," and "congenital," used in the source terminology, are qualifiers in the target terminology. The word "solitary," used in the target terminology, is a qualifier there. Yet the term "infected," which is used in a parallel manner to the four qualifiers, is a finding. Furthermore, it is remarkable that two terminologies use qualifiers for the same parent concept, yet do not have a single qualifier in common. In Sample 1, 50% (18/36) of all parent concepts fall into this category.

Solutions are demonstrated for the different categories above from a user perspective as follows. For categories 1 and 3 the default solution is that concept import is still possible. However, straightforward import would lead to a notable loss of information. Thus, in such a case it is advised to create an intermediate level of two concepts that make explicit the different nature of the original child concepts and the imported children. This is demonstrated visually and abstractly in Figure 4.3.

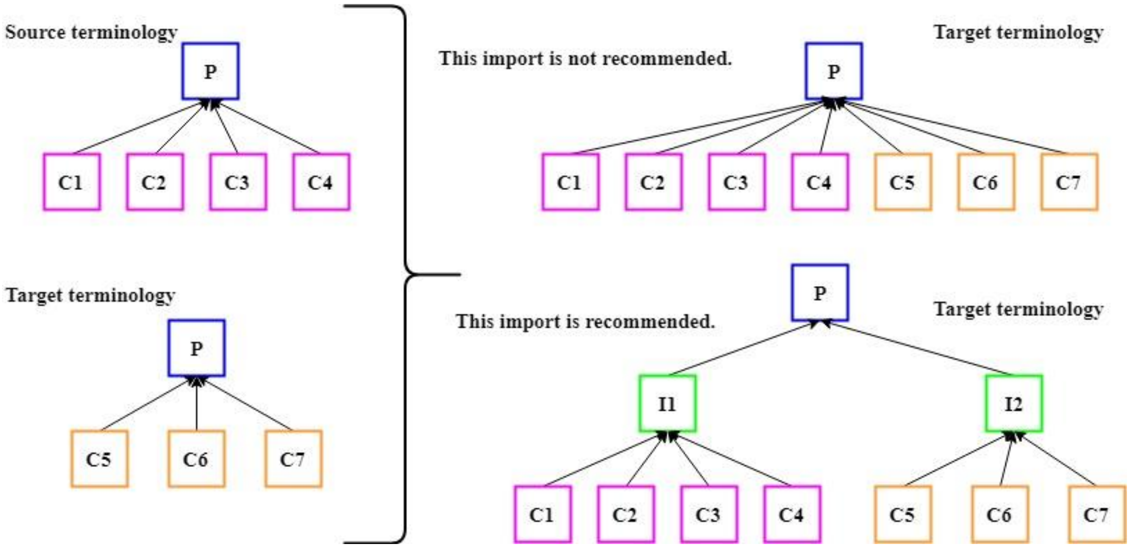


Figure 4.3 Proposed solution for categories 1 and 3 of alternative classification.

For category 2, all concepts can be imported from the source terminology into the target terminology and this merge would not create any loss of information. For category 4 the correct solution would be to clean up the child structure first, e.g., in the above example by moving "infected" to a more appropriate position and then allowing for direct import or for import while adding an additional level of concepts (Figure 4.3).

Sample 2: For Sample 2, out of the 50 concepts, the domain expert determined that 40 concepts should be imported. Thus, we have 80% (=40/50) of cases of import. Table 4.2 shows the results for this sample. Fisher's exact test (two-tailed) was performed on the control and experimental group for Sample 2. For Hypothesis 4.2, statistical significance was obtained with p-value<0.001 supporting the hypothesis. Thus, the child concepts suggested by the algorithm for import based on their EFI values' proximity to 1 could be distinguished from the non-children (cousins) with statistical significance.

Table 4.2 Results of each Group as Analyzed by the Domain Expert for Sample 2

Sample 2		
Groups	**Number of concepts agreed for import**	**Number of concepts disagreed for import**
Experimental	40	10
Control	6	44

Sample 3: For Sample 3, out of the 50 concepts, the domain expert determined that 27 concepts should be imported. This corresponds to 54% of cases for import, which is much less than the percentage of import for Sample 2. Table 4.3 shows the results for both the experimental and control groups. Fisher's exact test (two-tailed) was performed on the control and experimental group for Sample 3 and statistical significance was obtained with p-value<0.001 supporting Hypothesis 4.3. Thus, the child concepts suggested by the algorithm for import based on their EFI values' proximity to 0 could be distinguished from the cousins with statistical significance.

Table 4.3 Results of each Group as Analyzed by the Domain Expert for Sample 3

Sample 3				
Groups	**Number of concepts agreed for import**	**Number of concepts disagreed for import**	**Alternative classifications**	**Error**
Experimental	27	5	12	6
Control	6	29	1	14

4.3 Discussion

It was initially assumed that a pair of concepts appearing in two terminologies, designated in the UMLS to be identical, but having no common children, would sometimes indicate that the concepts are in fact homonyms and not identical. This study and the detailed analysis of Sample 1 did not bear this out.

Hypothesis 4.2 and Hypothesis 4.3 were separated, because the raw numbers seemed to indicate different "behavior" when EFI tends to 1, compared to when EFI tends to 0. Indeed, there are relatively more alternative classifications and errors for EFI close to 0. Nevertheless, for both polarities of EFI, the largest number of choices according to the domain expert was "import."

An important distinction in this study is between EFI = 0 and EFI ≠ 0. The case of EFI = 0 provides a fully automatic criterion for recognizing likely alternative classifications. To the best of our knowledge, no such criterion has been previously reported in the literature. Recognizing an alternative classification can then be followed by a human review to determine which approach should be taken when importing concepts into the target terminology. Ideally, the distinctions between different methods of import and different kinds of errors should be recognized by an algorithm with high reliability. While this goal is hard to reach, the current results establish a step in that direction.

CHAPTER 5

EXTENDING CONCEPT IMPORT TO THREE TERMINOLOGIES

The topological patterns discussed in Chapters 3 and 4 used a single source terminology with one target terminology for identifying potentially missing concepts in the target terminology. The study presented in this chapter extends the algorithmic detection of "candidate concepts for import" from one source terminology to two source terminologies used in tandem. The combination of two source terminologies relative to one target terminology leads to the discovery of candidate concepts for import that could not be found with the same "reliability" when comparing one source terminology alone to the target terminology. The analysis revealed a specific configuration of concepts, overlapping two source terminologies and one target terminology, for which the name "fire ladder" pattern was coined. The three terminologies in this pattern are tied together by a kind of "transitivity." In this chapter, a quantitative analysis of the discovered fire ladder patterns is provided with a report on the inter-rater agreement concerning the decision of importing candidate concepts from source terminologies into the target terminology.

5.1 Fire Ladder Pattern for Concept Import

Consider three terminologies A, B, and C shown in Figure 5.1. Concept A1 in terminology A has a child concept A3, the concept B1 in terminology B has a child B2, and the concept C2 in terminology C has a child C3. The concepts A1 and B1 are identical by means of having the same UMLS CUI. Similarly, the concepts B2 and C2 are identical, and so are A3 and C3. It should also be noted that the concept C3 (=A3) does not exist anywhere in terminology B, the concept B2 (=C2) does not exist anywhere in terminology A, and the concept A1 (=B1) does not exist anywhere in terminology C. Looking only at A1, B1, B2, C2, and ignoring that the connections between them are of two different kinds (IS-A versus identity) this identifies a kind of transitivity (Figure 5.1) [111].

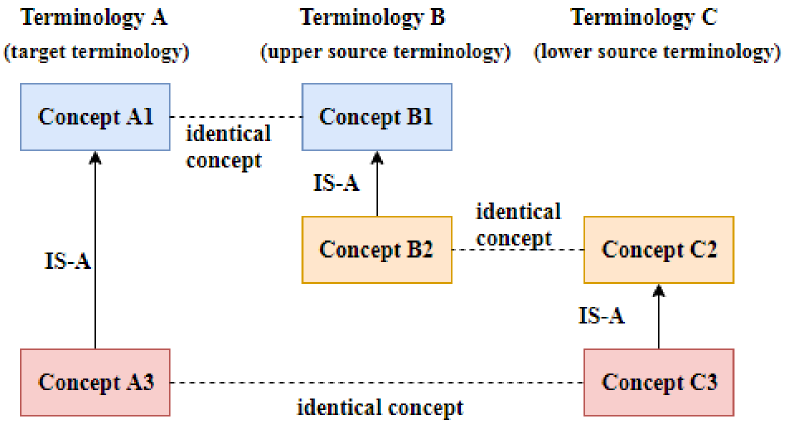


Figure 5.1 An abstract fire ladder pattern involving three terminologies.
Source: [32]

Figure 5.2 Image of a fire truck with an extensible fire ladder.
Source: [112]

As two vertical patterns are chained together to jointly achieve a "higher reach," the pattern resembles an extensible ladder carried by fire trucks (Figure 5.2). Thus, the pattern in Figure 5.1 is referred to as the *fire ladder pattern* in contrast to the diamond patterns that used vertical density differences [25]. Terminology A is denoted as the target terminology, Terminology B as the "upper source terminology," and Terminology C as the "lower source terminology." The primary questions that arise from Figure 5.1 are whether B2 (=C2) should be proposed for import into Terminology A and whether C3 should be recommended for import into terminology B.

In this study, a quantitative exploration of the fire ladder patterns formed by the concepts from 10 different terminologies in the UMLS Metathesaurus was conducted. An algorithm that identifies the fire ladder patterns and suggests concepts that could potentially be imported into the target terminology was developed. Two domain experts reviewed the suggestions made by the algorithm for deciding whether the concepts should be imported or not. An example involving the terminologies HPO, NCIt, and SNOMED CT is shown in Figure 5.3.

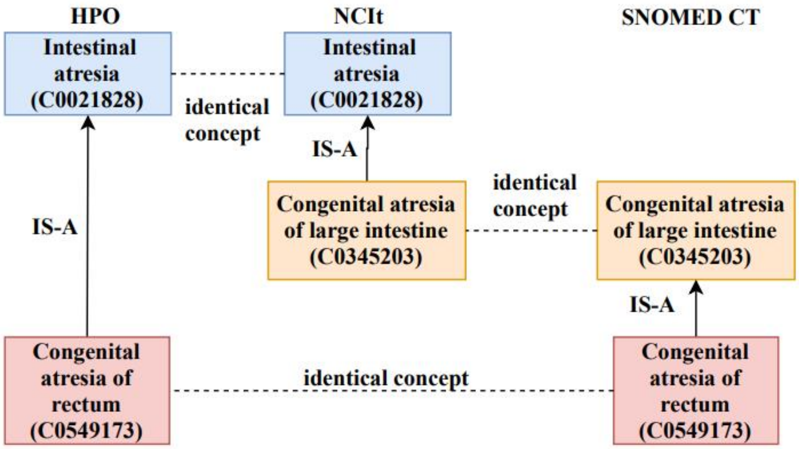

HPO
NCIt
SNOMED CT
Intestinal atresia (C0021828)
identical concept
Intestinal atresia (C0021828)
IS-A
IS-A
identical concept
Congenital atresia of large intestine (C0345203)
Congenital atresia of large intestine (C0345203)
IS-A
Congenital atresia of rectum (C0549173)
identical concept
Congenital atresia of rectum (C0549173)

Figure 5.3 An example of a fire ladder pattern with the terminologies HPO, NCIt, and SNOMED CT. The UMLS CUIs of the concepts are provided inside the parentheses.
Source: [32]

5.1.1 Algorithm

The fire ladder pattern is formed by concepts having a PAR relationship with an inverse_isa Relationship Attribute, which denotes in the UMLS the hierarchical IS-A relationship. As discussed previously in Sections 3.1.1 and 4.1.1 there are 10 such terminologies excluding SNOMEDCT_VET and UWDA (excluded as they are subsets of other terminologies). In the algorithm, these 10 terminologies are referred as T_1, $T_{2,\,\ldots}$, $T_{10.}$ For this study, the 2018 AB release of the UMLS was used.

The algorithm has two parts. FIRE_LADDER is the top-level algorithm. It generates the set PT of all distinct triples of terminologies taken from the set T = {T_1, $T_{2,}$ …, T_{10}}, i.e., PT = {<T_1, T_2, T_3>, <T_1, T_2, T_4>, …. <T_8, T_9, T_{10}>}. Because one of these three terminologies is designated the target terminology, the second is the "upper source" and the third is the "lower source," (Figure 5.1) <T_1, T_2, T_3> is distinct from <T_1, T_3, T_2>, etc. Thus, PT is the set of all permutations [113] of three terminologies taken from 10 terminologies. Therefore, there are 720 triples in PT, according to the formula,

$$P\ (n,k) = \frac{n!}{(n-\ k)!}, where\ k\ =\ 3 \text{ and } n\ =\ 10 \qquad (5.1)$$

The second part of the algorithm, named FIRE_LADDER_SUB, takes two inputs, namely ontDAG (described in detail in Section 3.1.3) and the set PT generated by FIRE_LADDER. The pseudocode of FIRE_LADDER_SUB is given below.

The algorithm outputs a file (F) with information about sets of concepts that form a fire ladder pattern and the three terminologies each concept set is derived from. The total time to execute the script corresponding to the above algorithms and to generate the output file was approximately 22 seconds on an Intel(R) Core i5 CPU with four cores and

~2.4GHz clock speed.

```
Procedure FIRE_LADDER_SUB(in: (ontDAG, PT) out: file F)
for each <$T_i$, $T_j$, $T_k$> in PT do
  $T_A \leftarrow T_i$
  $T_B \leftarrow T_j$
  $T_C \leftarrow T_k$
  for each concept in ontDAG{$T_A$} do
    if concept in ontDAG{$T_B$} then
      $A_1$, $B_1 \leftarrow$ concept
      $A_1$_children ← ontDAG{$T_A$}{$A_1$}{'children'}
      $B_1$_children ← ontDAG{$T_B$}{$B_1$}{'children'}
      for each $b_1$_child in $B_1$_children do
        if $b_1$_child not in ontDAG{$T_A$} then
          if $b_1$_child in ontDAG{$T_C$} and $B_1$ not in ontDAG{$T_C$} then
            $B_2$, $C_2 \leftarrow$ $b_1$_child
            $C_2$_children ← ontDAG{$T_C$}{$C_2$}{'children'}
            for each $c_2$_child in $C_2$_children do
              if $c_2$_child not in ontDAG{$T_B$}and $c_2$_child in $A_1$_children then
                $C_3$, $A_3 \leftarrow$ $c_2$_child
                Output: Write ($T_A$, $T_B$, $T_C$, $A_1$, $B_1$, $B_2$, $C_2$, $C_3$, $A_3$) to file F
              end if
            end for
          end if
        end if
      end for
    end if
  end for
end for
```

5.1.2 Sample Preparation and Evaluation

Two data sets (Data Sets 1 and 2) were created from the fire-ladder pattern obtained by the algorithm. The patterns were reviewed for import by two domain experts. Data Set 1 corresponds to the enrichment of Terminology A by importing B2. For this data set, the domain experts were provided with the names of the three terminologies (A, B, and C) and also the concepts A1 (=B1), B2 (=C2), and C3 (=A3) and asked for their judgement on whether the concept B2 should be imported into Terminology A as the child of A1 and parent of A3.

It should be noted that the fire ladder pattern supports another possible import resulting from the horizontal density difference between the terminologies B and C. Thus, the domain experts were also asked about their judgement on importing C3 (=A3) into Terminology B as a child of B2. Accordingly, for this Data Set 2, the domain experts were provided with the names of the terminologies (B and C) and the concepts B2 and C3. For this import, B would become the target terminology and C would simply be the source terminology without qualification as upper or lower. This kind of import would be similar to the study on horizontal density differences [30] in Chapter 2. However, a larger number of ontology combinations are investigated in this study.

The review of Data Set 1 was done in two phases. In the first phase, along with the decision on whether a concept should be imported or not, the domain experts were also asked to provide the reasons behind their judgment. Once the results of the first phase from both of our domain experts were received, another round of reviews was initiated limited to those patterns on which the domain experts disagreed with each other. In this phase, both of the domain experts were shown the reasons behind each other's decisions. This resulted in only one change to the data for Data Set 1, increasing the metric of agreement minimally. The inter-rater agreements based on Krippendorf's α and Cohen's Kappa were also calculated.

5.2 Results

Out of the 720 triples of terminologies possible according to Equation 5.1, 26 triples formed fire ladder patterns. For Data Set 1, 55 *distinct* B2 concepts were identified by the algorithm for import into Terminology A. These concepts were reviewed by the experts.

There were two cases (in addition to the 55 mentioned above) in which the same triple of concepts (A1, B2, C3) was formed by different permutations of terminologies. For example, A1: *Rhabdomyoma*, B2: *Cardiac rhabdomyoma*, C3: *Congenital rhabdomyoma of heart* was formed by the triple <SNOMED CT, NCIt, MEDCIN> and the triple <SNOMED CT, HPO, MEDCIN>. Since the target terminology is the same (SNOMED CT in this case), these two permutations were considered together for Data Set 1, yielding a total of 55 distinct B2 concepts for a total of 57 fire ladder patterns discovered.

Table 5.1 shows each triple of terminologies and the number of fire ladder patterns formed by the permutations of these terminologies. There were 18 instances formed by permutations of {SNOMED CT, MEDCIN, CPT} and another 17 instances by permutations of {SNOMED CT, NCIt, MEDCIN} accounting for more than half of the candidate concepts. It should be noted that columns one and three in Table 5.1 represent permutations of triples of terminologies and not a single triple. For example, the triples <HPO, SNOMED CT, NCIt> and <HPO, NCIt, SNOMED CT> contributed two fire ladder patterns each to get the four patterns listed in the third row of Table 5.1.

Table 5.1 Triples of Terminologies and the Number of Fire Ladder Patterns Formed by Permutations of each Triple

All permutations of a triple of terminologies	Number of fire ladder patterns	All permutations of a triple of terminologies	Number of fire ladder patterns
CPT, SNOMED CT, MEDCIN	18	UMD, SNOMED CT, NCIt	2
NCIt, SNOMED CT, MEDCIN,	17	MEDCIN, ATC, NCIt	2
HPO, SNOMED CT, NCIt	4	SNOMED CT, HPO, MEDCIN	2
CPT, NCIt, MEDCIN	3	CPM, SNOMED CT, NCIt	1
FMA, SNOMED CT, NCIt	3	HPO, SNOMED CT, MEDCIN	1
SNOMED CT, CPM, MEDCIN	3	SNOMED CT, GO, FMA	1

Out of the 55 concepts suggested for import by our algorithm for Data Set 1, one domain expert agreed on importing 42 concepts (76.3%) and the other agreed on 45 concepts (81.8%) (Table 5.2). The two domain experts agreed in their decisions regarding 39 out of 55 concepts (71%). The inter-rater agreements using Krippendorff's α score and Cohen's Kappa were 0.51 and 0.507, respectively. Examples of fire ladder patterns are shown in Table 5.3.

Table 5.2 Details of the Domain Experts' Decisions Regarding Importing the Concepts out of 55 Suggestions Made by the Algorithm

Domain Expert 1	Domain Expert 2	Two Domain Experts
Recommends import	Recommends import	Both recommend import
		39
42	45	Both recommend non-import
		7
Recommends non-import	Recommends non-import	One Expert for import one against
13	10	9

Table 5.3 Examples of Fire Ladder Patterns

Term. A	Term. B	Term. C	Concept A1	Concept B2	Concept A3
SNOMED CT	MEDCIN	CPT	Drug measurement	Therapeutic drug assays	Theophylline assay
NCIt	MEDCIN	SNOMED CT	Urologic surgical procedures	Operation on urethra	Urethrostomy
HPO	SNOMED CT	NCIt	Adrenal gland hypofunction	Adrenal cortical hypofunction	Secondary adrenal insufficiency

For Data Set 2, 105 distinct pairs of concepts (B2, C3) in terminologies B and C were identified. For one concept B2, there were several concepts in the position of C3. For instance, for the fire ladder pattern formed by A1: *Tract of spinal cord*, B2: *Descending spinal cord tract* two different C3s namely *Structure of medial reticulospinal tract* and

Structure of lateral reticulospinal tract were observed. While for Data Set 1 each algorithmic suggestion would potentially result in importing one concept into Terminology A, for Data Set 2 two potential imports into Terminology B were possible in this example. The domain expert agreed to the import of 98 concepts out of 105 concepts (93.33%). Examples are shown in Table 5.4.

Table 5.4 Examples from Data Set 2, for Concept C3 Agreed to be Imported into Terminology B as the Child of Concept B2

Term. B	Term. C	Concept B2	Concept C3
MEDCIN	NCIt	Vital signs measurements	Heart rate
HPO	MEDCIN	Cardiac rhabdomyoma	Congenital rhabdomyoma of heart
NCIt	SNOMED CT	Colon carcinoma	Carcinoma of descending colon
ATC	NCIt	Thyroid hormones	Levothyroxine sodium
GO	FMA	Region of chromosome	Short arm of chromosome

An error analysis was performed for cases in which the domain experts did not recommend algorithmically determined candidate concepts for import. One example from Data Set 1 consists of the fire ladder pattern formed by A1: *Metastatic Neoplasm*, B2: *Secondary Neoplasm*, and C3: *Metastasis to digestive organs*. According to the domain experts, A1 and B2 are sufficiently close to each other to be considered synonyms. For Data Set 2, the concept *anterior radial head dislocation* was not imported as the child of *Congenital dislocation of radial head*, because the former concept is not necessarily congenital.

5.3 Discussion

The study discussed two possible cases of import. One can think of a third possible case of import based on Figure 5.1, which is importing B1 (=A1) into terminology C as a parent of C2. However, this presents another question as to how to find a parent for the new C1, given that there should be a path from every concept to the root of its terminology, following design standards in the field of ontologies and terminologies.

The question arises whether transitive patterns can be constructed for four terminologies at a time. Preliminary research was performed on this question and no such patterns were identified with the 10 terminologies from the UMLS. Another question, to be explored in the future, is whether the import of a concept could lead to the subsequent discovery of new vertical density differences. Thus, after importing B2 into A (Figure 5.1), A1 and B2 together could form the right side of a new diamond pattern (Figure 3.1) with a fourth terminology.

There is one more approach to extend the set of density-based methods for discovering candidate concepts for import. Referring to Figure 5.1, B2 is a child of B1. However, it is possible that B1 and B2 together define a path with one or more intermediate concepts between them. Consider that there is exactly one such intermediate concept (say B1.5). In that case, the fire ladder pattern of Figure 5.1 would suggest the import of both B2 and B1.5 into the terminology A. This approach can also be extended for importing concepts from terminology C into terminology B, by extending the length of the path between C2 and C3 and adding intermediate concepts such as "C2.5" between them. Investigating this kind of pattern requires a more complicated algorithmic approach and is left for future work.

It is important to stress the contribution of using two source terminologies in tandem, which is a novel method reported for the first time in this study. In Data Set 1, the level of confidence that a suggested candidate for import is correct is high because it is constrained from above and below. While there have been cases [27] where candidates were constrained from above and below by a *single* source terminology, this was not possible for the 55 candidate concepts that there were discovered in this study. For Data Set 2, a candidate concept for import is only constrained from above, similar to the previous work [30], which is a weaker indication that an import is desirable.

The question of the right degree of pre-coordination has been discussed previously in the literature, e.g., [114]. On one hand, the difficult task of post-coordinating concepts should not be left to the users, who are likely not experienced and knowledgeable about ontologies. On the other hand, creating many pre-coordinated concepts increases both the effort of the curator and the search effort of the user, because these concepts are "cluttering up" the ontology. Finding the right balance between too much pre-coordination and too little pre-coordination is difficult.

CHAPTER 6

SUSTAINABILITY OF BIOMEDICAL ONTOLOGIES

Biomedical ontologies are developed by investing large number of resources including money and manpower. The study described in this chapter presents an attempt to understand how well the ontologies are maintained over the years after their initial development and thus whether developing maintenance methods is of wider interest. This study is motivated by the fact that the attempts at identifying missing concepts in ontologies and suggesting concepts for import become fruitless if the ontologies are not being updated and maintained by the ontology developers. The study is based on ontologies in BioPortal. BioPortal maintains information about various aspects of each ontology. The information from BioPortal collected on January 18th, 2018 was used to analyze various aspects of the ontologies. On that date, there were 684 ontologies listed on BioPortal.

One of the primary goals of this study is to identify the ontologies in BioPortal that are not regularly updated and try to understand the root causes of this situation. To this end, 83 ontologies were identified with at least 1000 distinct concepts that had not been updated since January 1, 2016, on the BioPortal site. Despite not being updated for quite a long time, some of these ontologies still had a substantial user base as evident by their recent usage patterns reported on BioPortal. However, the analysis shows that there are several reasons behind the sporadic update/maintenance of these ontologies.

6.1 Analysis of BioPortal Ontologies

BioPortal maintains a submission page for each ontology uploaded into it. This submission page contains information pertaining to that particular ontology, including the release date, upload date, ontology format, short description, URL of the homepage, version number, contact information of ontology curators, etc., for different submissions of the ontology on BioPortal. Out of 684 ontologies at the time of this study, 653 had detailed information on the submission page. Figure 6.1 shows the number of ontologies released each year, based on the first release date information in BioPortal.

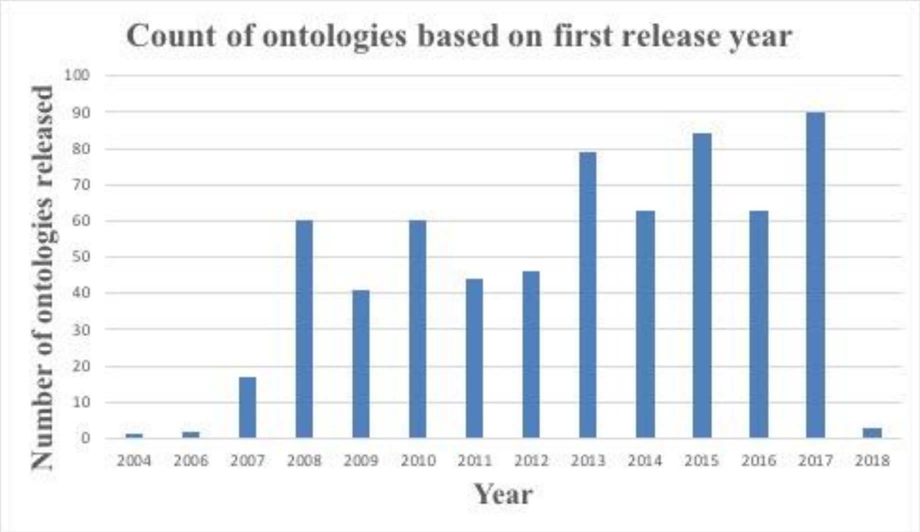


Figure 6.1 Number of ontologies released per year based on the release date in BioPortal for the first submission.
Source: [29]

The information provided in the metrics page of the ontologies in BioPortal contains data about 601 ontologies. According to this data, the 10 largest ontologies (based on the number of classes) include the National Center for Biotechnology Information (NCBI) Organismal Classification (NCBITAXON) [115], Gazetteer (GAZ) [116], the Drug Ontology (DRON) [58], SNOMED CT [117], the Protein Ontology (PR) [118], Robert Hoehndorf Version of MeSH (RH-MESH) [119], Medical Subject Headings (MeSH) [23], Logical Observation Identifier Names and Codes (LOINC) [120] and the

BioModels Ontology (BIOMODELS) [121].

It is not immediately obvious how to measure the popularity of an ontology. However, the NCBO Ontology Recommender service [122] uses the number of ontology accesses as a proxy for popularity. To this end, BioPortal provides a bar chart of the most visited ontologies of the current month on its home page. For example, in February 2018, the Current Procedural Terminology (CPT) [123] (69,417) tops the list, followed by the Medical Dictionary for Regulatory Activities (MEDDRA) [124] (14,437), RXNORM [60] (11,798), SNOMED CT (9,153) and the National Drug Data File (NDDF) [125] (7,511).

Additionally, the number of visits per month for each ontology is also presented on the corresponding pages. This information was exploited to generate a scatter plot of/between the number of days since an ontology was last updated in BioPortal and the number of times that ontology has been accessed since the last update, for all the ontologies in BioPortal at the time. Figure 6.2 depicts this scatter plot. The top-5 accessed ontologies mentioned above were excluded to avoid the scaling effect caused by them. It was observed that many of the ontologies, even though they were not updated for more than two years, are being actively used, which is evident from the number of accesses shown in the plot. To be specific, for the 83 ontologies that were not updated since January 2016, the average number of accesses since the last update date was 1548 at the time of the study.

Another proxy for the popularity of an ontology could be the number of projects that make use of it. According to BioPortal, the Gene Ontology (GO) [22] is used in the largest number of projects (61). The other ontologies that have many associated projects are the Basic Formal Ontology (BFO) [126], the Ontology for Biomedical Investigations (OBI) [127], SNOMED CT, the Foundational Model of Anatomy (FMA) [128], National

Cancer Institute Thesaurus (NCIt), Chemical Entities of Biological Interest Ontology (ChEBI) [55] and MeSH. Around 35% of the ontologies in BioPortal have project information associated with them.

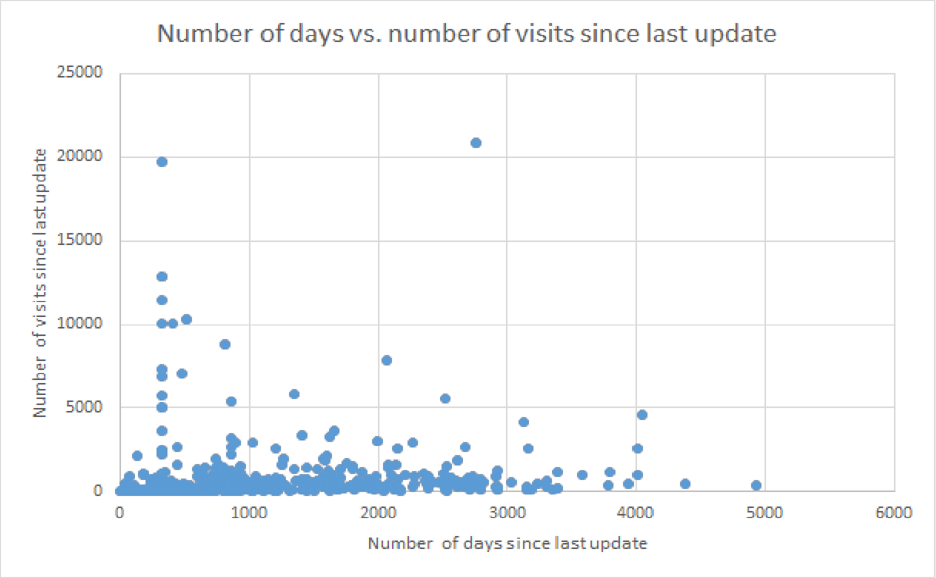


Figure 6.2 Number of days since the last update vs. number of visits since the last update.
Source: [29]

The frequencies with which the ontologies are updated in BioPortal also vary widely. For example, some ontologies are updated 3-4 times a week, while others are updated quarterly or twice a year. The most commonly updated ontology in BioPortal is the Gene Ontology (GO). Some other frequently updated ontologies include the Mosquito Insecticide Resistance Ontology (MIRO) [129], the Systems Biology Ontology (SBO) [130], the Human Disease Ontology (DOID) [131], and the Human Phenotype Ontology (HP) [56].

In this study, the primary interest is in ontologies that have not been updated for a long time, and that are of a substantial size. The goal is to understand why ontologies that were apparently built with a great investment of time, effort, and budget are not being maintained/sustained and how big a problem this constitutes. Another question of interest is whether the fact that an ontology is not actively maintained has an effect on the frequency of access to it.

6.1.1 Methods

To limit the scope of the study to a manageable size, only those ontologies that had not been updated in BioPortal since January 1, 2016, were chosen. There was a total of 317 such ontologies. In Figure 6.3 the number of ontologies based on their last update date in BioPortal is presented.

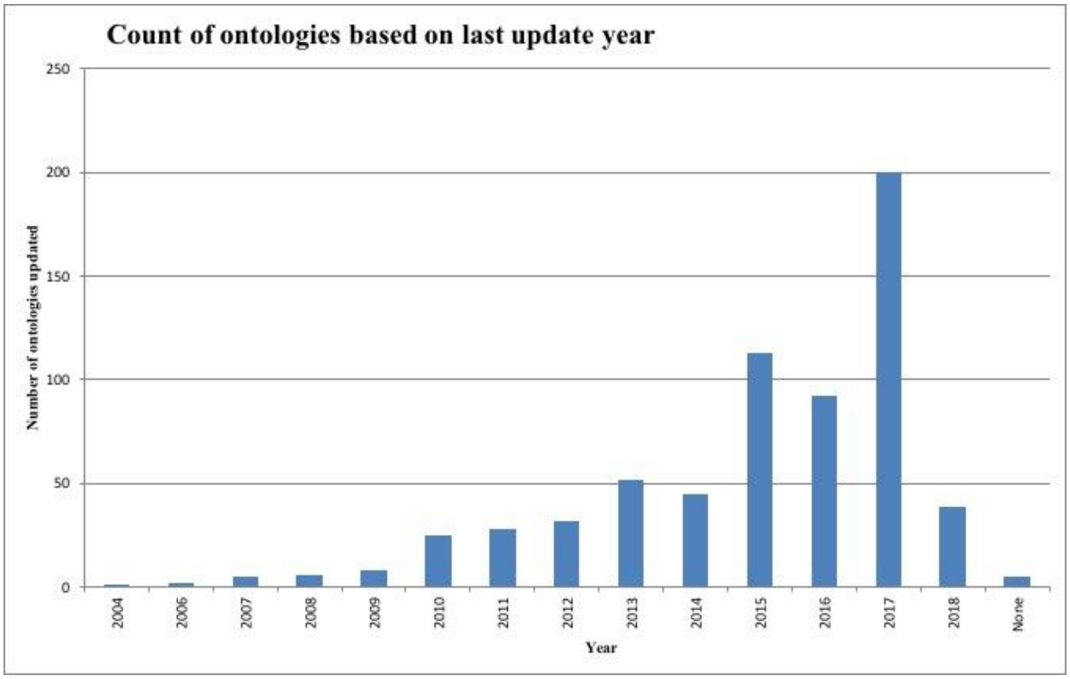


Figure 6.3 Number of ontologies based on the date of their last submission in BioPortal.
Source: [29]

Roughly 47% of the ontologies in BioPortal had not been updated since January 1, 2016. Small ontologies with few concepts may not be of great interest to a broader community. Hence, the study is limited to the subset of ontologies that have a substantial number of concepts. From the above 317, a subset of 83 ontologies with at least 1,000 distinct concepts was filtered out.

To identify the root causes why the ontologies are not regularly updated/maintained, personalized email messages were sent to the curators/owners on file for these 83 ontologies. The traditional route of questionnaires was avoided due to general questionnaire fatigue [132] and scientific evidence of continuously falling response rates [133]. The email was composed with three questions shown below:

I am working on a study on the life cycle of ontologies. In BioPortal you are listed as the contact for the xxx ontology. According to the information in BioPortal, the xxx ontology has not been updated for two years or longer.

1. *For my study, I am wondering what has motivated the developers of the xxx ontology to stop updating it on BioPortal.*

2. *Has the development in general stopped, or is it being made available at a different website?*

3. *If the development has stopped, I wonder if you could shed some light on the question why this happened.*

If you are the wrong person for these questions, I apologize, and could you please direct me to the Principal Investigator of this ontology project.

The responses received were studied in detail, to categorize them into groups that had not been preconceived to avoid any limiting biases. This *modus operandi* is loosely based on the phenomenological approach [134, 135] to qualitative data analysis. There were seven categories derived that accounted for almost all the email responses.

As a second source of information concerning whether an ontology is being maintained, the publications indexed by the name of the ontology and papers on which the curators of the ontology appear as authors were used. For this purpose, an extensive search of PubMed and Google Scholar was performed. The specific interest was to identify when the most recent publication about an ontology appeared. Allowing for delays in publishing a paper, a hypothesis was formulated that work on an ontology might be terminated after a concluding paper had been published by a journal or a major conference.

6.2 Results

For sending emails, the contact information of the curators of the 83 ontologies that were filtered for the study was collected from BioPortal. There was no contact information available for one ontology. Also, 14 emails bounced back with "address not found" messages returned to us. In the first five days after the initial email, 29 responses were received. After that, a reminder email was sent to the remaining ontology curators and then another 19 responses were received, giving a total of 48 responses. It should be noted that 48/83 = 0.5783 = 57.8% is much larger than the typical response rate for a survey for which no incentive is given, confirming the choice of using an individualized email instead of a questionnaire. According to Medway [136], a response rate of 10.9% was achieved in her study for a control group that was not incentivized. Below a few prototypical anonymized responses (not edited for spelling errors etc.) from ontology curators are presented.

Response Sample 1: *Ontology X1 has been developed in the frame of xxx (http://www.xxx.xxx) a xxx funded project in which out team in xxx were partners since the beginning. Roughly 3 years ago, in the last renewal of the project, our funding has been cut and we were obliged to drop active maintainance of xxx. Since we are firm supporters of the rules set by xxx we are still supporting xxx, replying to requests and even adding new terms and updating if we have such a request. During this period I haven't received* *any proposed term to be included in xxx, thus no updates are available.*

Response Sample 2: *The evolution of the X2 and X3 ontologies, that have been developed by our organization (xxx), and for which I have the responsibility, depends very heavily on the human and financial resources that we invest in it.*

The first phase (2008-2011) mobilized about 50 researchers who invested themselves in time, folowing the request of our organization without special funding, which led to an already very rich first version.

The following developments could only be achieved through major research programs. Thus, from 2011 to 2015, the enrichment of ontologies, in particular with xxx, could be carried out thanks to the financing provided by a major xxx". During this period, new versions of xxx and xxx were deposited on the bioportal site (we can also find downloadable versions on the xxx website: http://www. xxx /).

Until the end of 2016, there was no more funding, which explains the lack of new developments in the two ontologies, which are already very complete.

In 2017, we obtained funding from xxx to enrich xxx with xxx relating to the xxx. We are therefore working on this project which should allow to put online an upgraded version of xxx within 6 months.

I hope that these few explanations will allow you to better understand the evolution of the life cycle of our ontologies.

Response Sample 3: *Simple reason for this - funding stopped, project ended. If someone wants to pick up the development, I am happy help.*

The email responses were constructive and informative. The shortest response consisted of only 11 words. The most elaborate response was a full 484 words long. The average response was 90 words long, counting URLs as single words. Overall, the response emails contained 17 URLs, typically referring to new ontology resources. None of the email responses expressed annoyance about the questions, and in nine cases respondents were thankful for the interest in their work.

Table 6.1 Main Categories of Responses

Groups	Reason	Number of Ontologies	Ontologies
Group 1	Incorporated into other ontologies/systems	9	ICECI, DDO, PAE, TAO, CSEO, EHDA, MAT, AAO, EHDAA
Group 2	Lack of funding, time and manpower, interruption in funding	15	BIOMODELS, RH-MESH, TGMA, ONSTR, HFO, SDO, IDOMAL, CCONT, DERMO, APAONTO, ATOL, GALEN, GMO, MCCL, NPO
Group 3	Updating in large time gaps / slow development	7	GEXO, RETO, REXO, ICNP, IDOBRU, HINO, CHMO
Group 4	Organization / Project ended	6	OBI_BCGO, NIGO, GENE-CDS, TRAK, ONTOPNEUMO, DINTO
Group 5	Paused for publication / redesigning	3	ESSO, ROO, GMM
Group 6	Concepts valid / Serves the purpose	3	EHDAA2, ONTOLURGENCE, IMGT-ONTOLOGY
Group 7	Not updating on BioPortal	3	OMIT, PMA, COGAT
Unassigned		2	HUGO, NIFSUBCELL
Total		**48**	

The inquiry to the owners of the 83 ontologies resulted in a varied set of responses. The seven major categories that explain the reasons behind the sparse updates of these ontologies are shown in Table 6.1. Two responses were hard to assign to any specific category. One of the two responses is given below.

xxx are not an ontology project, we are a project naming xxx, I suggest you look at our website www. xxx to find out more. The "ontology" was not created by us, it was created from our publicly available data by the BioPortal project. We have not been asked to submit any new data to the portal - we assumed if they wanted to update the data they would have an automated update process as all of our data is freely available from our website. Hence no update. I hope this answers your questions.

The largest number of responses indicated a variation on the theme that there was a lack of funding or manpower. A total of 15 responses expressed this issue. The second largest group (9 responses) indicated that the ontology had been folded into another ontology or system that obviated the need for continued independent development. The third biggest group of responses claimed that development was ongoing, however, it continued at a slow pace or with unpredictable gaps. Seven responses were in this category.

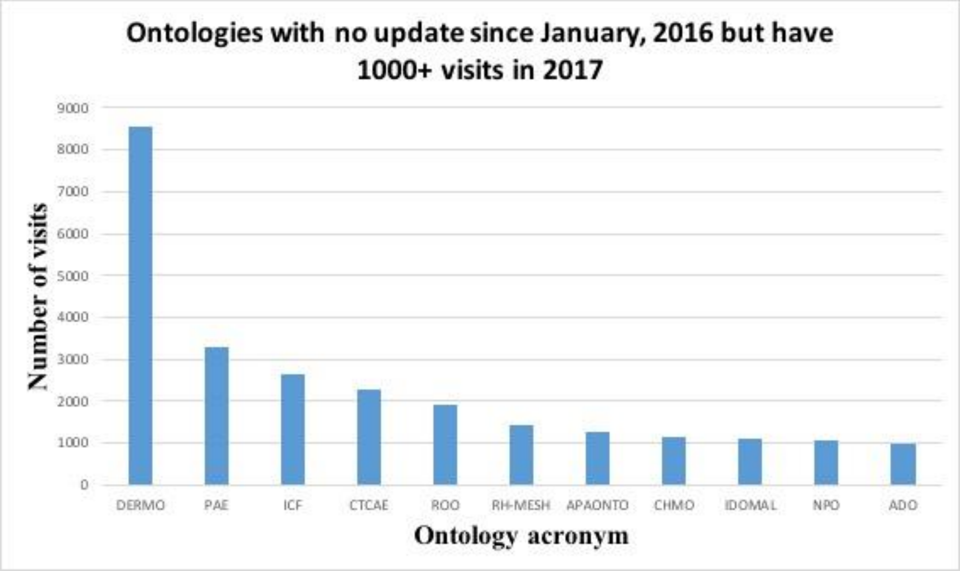


Figure 6.4 Total number of visits for the 11 most-visited ontologies not updated between January 1, 2016, and January 18, 2018.
Source: [29]

The usage pattern of these 83 ontologies was analyzed to understand whether the lack of updates has prevented users from accessing these ontologies. Figure 6.4 shows the number of visits for the 11 most-visited ontologies in this category for the year 2017. Development of these ontologies had been discontinued at least a whole year prior.

Secondary Analysis based on Publications: Out of the 83 ontologies under consideration, no corresponding publications were found in PubMed or Google Scholar for 26 ontologies. For seven ontologies their last update year coincided with their last publication year, while 12 ontologies were last updated one year before their last publication and 15 ontologies were last updated one year after the final publication. To summarize, for 60% (7+12+15=34) of the 57 ontologies that had at least one corresponding publication, updates in BioPortal stopped within a year before or after the last publication about this ontology. The remaining 23 (= 57–34) ontologies had their last update two or more years before or after the last date of publication.

6.3 Discussion

The choice of an email message with questions as opposed to a questionnaire is currently unusual but is compatible with the phenomenological approach. There are disadvantages of this modus operandi. As answers were given in free text, major and minor categories of responses had to be deduced from the totality of replies, as opposed to having predefined answer choices and categories. In a questionnaire, the participants are usually forced to make choices, possibly leaving her/him with lingering doubts that cannot be expressed in the given form. For example, if a Likert scale allows only five choices and a participant feel that the right answer is between the top two choices, this is not expressible. On the other hand, with the free format email, the evaluators have to assign answers to categories, ignoring any expressed doubts.

The most important category of responses (15) focused on funding and staff size. This indicates that some ontology development projects are apparently conceived without either (a) a plan for sustaining and maintaining the ontology after the initial development period or (b) an underestimate of the resources in staff and budget that might be required to maintain an ontology over a longer period. Achieving the necessary funding levels might be difficult. In fact, as pointed out by Baker [137], federal funding agencies such as NIH are directing their funds to research and innovation as opposed to managing and maintaining large resource-oriented databases. This issue has been further explored, and the need for funding agencies to develop better mechanisms for supporting infrastructure-related research has been pointed out in previous work [138]. One possibility would be to reincarnate funding programs such as "Data Ontologies for Biomedical Research" that were supported by NIH about a decade ago [139]. We are not aware of any other mechanism that can "do the job of ontologies," and the categorization of scientific classes connected by explicit relationships will remain important for the future of Medical Informatics.

The results of the literature study supported the hypothesis that work on an ontology might be terminated after a concluding paper had been published in a journal or a major conference. Limitations of this part of the study include that not all publications appear in Google Scholar or PubMed. When the search was made for publications, the name of each ontology and the names of the contact persons listed on BioPortal were used as search keys.

It is entirely possible that a different author, not listed on BioPortal, has published a paper about a listed ontology and it was not discovered.

CHAPTER 7

CONCEPT MINING FOR INTERFACE TERMINOLOGY FOR ANNOTATING EHRs

Recent years have witnessed a major transformation in the field of health care, with the federal government taking the lead to encourage the use of Electronic Health Records (EHRs) [140]. With the wide use of EHRs, large volumes of discharge summaries, lab reports, progress notes, etc. have become available to the healthcare community. Such clinical notes, specifically progress notes, contain the most up-to-date relevant information per patient. While extremely valuable in describing the clinical conditions of a specific patient, the information is mostly recorded in unstructured text with highly specialized clinical phrases. To enable and augment interoperability and enhance healthcare quality by facilitating *post hoc* research studies, these records need to be annotated with concepts from a standard terminology. Without annotations, the text is often vague, ambiguous, and inadequate for automated processing. One of the main goals for converting paper records to EHRs was to support interoperability and research. This requires annotation of EHR notes with concepts from reference terminologies. In today’s EHRs, coded data entry is limited to specific segments, such as problem lists and quality measures.

The vision of "Meaningful Use" (MU) [141] required the use of standardized ontologies. SNOMED CT, which was selected for clinical recording, is rarely utilized. Two major reasons are that currently there are no satisfactory off-the-shelf tools to enable clinical note annotation; and physicians record clinical notes with medical phrases, many of which are not contained in standard medical reference terminologies used for annotation. To provide an example from the cardiology domain, consider the cardiology concept *supraventricular tachycardia* (SVT), which is a potentially dangerous fast heart rhythm arising from the upper part of the heart. Two important types of SVT are *atrioventricular nodal reentrant tachycardia* (AVNRT) and *atrioventricular reentrant tachycardia* (AVRT). SNOMED CT is the most comprehensive clinical reference terminology, nevertheless, it cannot represent this example. While SNOMED CT provides a code for SVT and AVRT, the latter is not classified as a child of SVT. SNOMED CT does not provide a code for AVNRT at all. This example illustrates that SNOMED CT lacks many fine granularity concepts. In fact, it lacks many such concepts.

To provide a recent example of how important annotations are for biomedical research, consider the COVID-19 pandemic. COVID-19 has turned into the greatest healthcare challenge since the Spanish flu pandemic, causing millions of infections and over 2.5 million deaths [142]. At the early stages of this pandemic, doctors have been describing signs and symptoms in various organ systems, e.g., "COVID toes" and Multisystem Inflammatory Syndrome in Children (MIS-C). However, most of these terms could not be coded and were only recorded as free text, inhibiting interoperability, and the use of EHR notes for research on the disease. How can the research on "COVID toes" and other related COVID-19 rashes (for example) be supported, if such findings are not coded in the EHRs to make them easily discoverable, and doctors and clinical software are forced to search for them as free text instead of as concepts?

To overcome these issues, two studies are described in this chapter that present how natural language processing-based concept mining techniques can be used to mine fine granularity concepts from text that are missing in reference ontologies/terminologies (discussed in Section 2.5). These concepts are added to an interface terminology (discussed in Section 2.5) that can facilitate annotation of unstructured text in EHRs. In some cases, an initial reference ontology exists. If this is not the case, an initial ontology can be constructed from SNOMED CT concepts. Both these situations are demonstrated in two separate case studies. One study below presents the creation of a Cardiology Description Interface Terminology that can facilitate the annotation of EHRs of cardiology patients. The second study presents the creation of a COVID Interface Terminology that can be used to annotate the EHRs of COVID-19 patients.

7.1 Creating Initial Interface Terminologies

An interface terminology [92] is different from reference terminology. The former is designed to maximize utilization by end-users and is serving specific applications, while the latter provides a formal representation of concepts acting as a comprehensive reference resource for aggregating data about the entire healthcare enterprise. An example application facilitated by an interface terminology is to support clinicians' entry of patient information into an EHR. Since interface terminologies are designed with user-specific applications in mind, they usually contain colloquial usages and common clinical phrases constituting a richer synonym content compared to reference terminologies. One of the recommendations for developing an interface terminology is to construct it from an existing ontology [93]. This section presents details on how this recommendation is followed to construct initial versions of both a Cardiology and a COVID Interface Terminology by starting from existing ontologies.

7.1.1 Initial Cardiology Description Interface Terminology

To create an Initial Cardiology Description Interface Terminology (ICDIT), the cardiology-related concepts from SNOMED CT were used. *SNOMED CT Concept* is the root concept in SNOMED CT. Under this concept, all the 19 SNOMED CT hierarchies are arranged. Several hierarchies of SNOMED CT were identified that contain subhierarchies of cardiology-related concepts, e.g., *Procedure on cardiovascular system* and *Cardiovascular finding*. A program was developed that takes as input the SNOMED CT ID of a concept and extracts its entire subhierarchy. Along with the Fully Specified Names (FSNs) of the concepts, all the synonyms of all the extracted concepts were also collected since synonyms are critical for annotation of the clinical notes. This process was repeated for all the roots of the subhierarchies containing cardiology concepts. All these concepts together constituted the ICDIT.

7.1.2 Initial COVID Interface Terminology

The Coronavirus Infectious Disease Ontology (CIDO) (discussed in Section 2.1.4) was used as the backbone for creating the Initial COVID Interface Terminology (ICIT). The concepts from five COVID-related ontologies in BioPortal were included into CIDO to create ICIT. A brief description of these ontologies is provided below.

The COVID-19 ontology [143] (2268 concepts) predominantly covers concepts related to cell types, genes, and proteins involved in virus-host-interactions, as well as medical and epidemiological concepts relevant to COVID-19. This ontology is similar to CIDO, but includes more concepts related to diseases affecting various systems of the human body. The COVID-19 Infectious Disease Ontology (IDO-COVID-19) [144] (486 concepts) extends the Infectious Disease Ontology (IDO) [145] and the Virus Infectious

Disease Ontology (VIDO) [146] to solely represent concepts related to the virus and diseases associated with COVID-19. The World Health Organization's (WHO) COVID-19 Rapid Version CRF semantic data model (COVIDCRFRAPID) [147] (398 concepts) aims at capturing the semantic references to the questions and answers in the case report form. Apart from this, there are two small ontologies in BioPortal - the COviD-19 Ontology for Cases and Patient information (CODO) [148], and the COVID-19 Surveillance Ontology (COVID19) [149], both with 52 concepts and mainly dealing with concepts related to the surveillance, geography, treatment facilities and tracking of patients. The ACT COVID Ontology v3.0 is available on GitHub [150] as SQL files that can be loaded into a database and was created to support cohort identification and related research by incorporating terms related to diagnosis, procedure, and medication codes from ICD [151], LOINC [9], CPT [7] and NDC [152].

About 2,446 concepts were extracted from the available files in this research and included in ICIT. In addition to these, UMLS, SNOMED CT [153], and LOINC have published lists of concepts related to COVID-19 on their respective websites that were also included. All the concepts were thoroughly examined, and duplicates were removed before adding them into ICIT.

7.2 Extracting Auxiliary Concepts for Extending Initial Interface Terminology

Apart from specific disease-related information and medications, EHRs also contain the anamnesis of a patient, which plays an important role in deciding the course of treatment for a patient. For example, the sentence "*74 year old male with Phmx of nephrolithiasis, prostate ca s/p XRT presented to ED…*" extracted from a clinical note describes the prior history of kidney stones and prostate cancer in a COVID-19 patient. Similarly, concepts such as *hypertension*, cholesterol, and *diabetes mellitus* are frequently recognized and annotated in cardiology notes. Such concepts are not present in ICIT or ICDIT. However, they are essential for the annotation of EHRs of COVID-19 patients and cardiology patients. Thus, it is important to add such concepts to the initial interface terminologies under the appropriate auxiliary hierarchies. SNOMED CT is a good source for providing such concepts. This section describes the process of extracting auxiliary concepts for extending both ICDIT and ICIT.

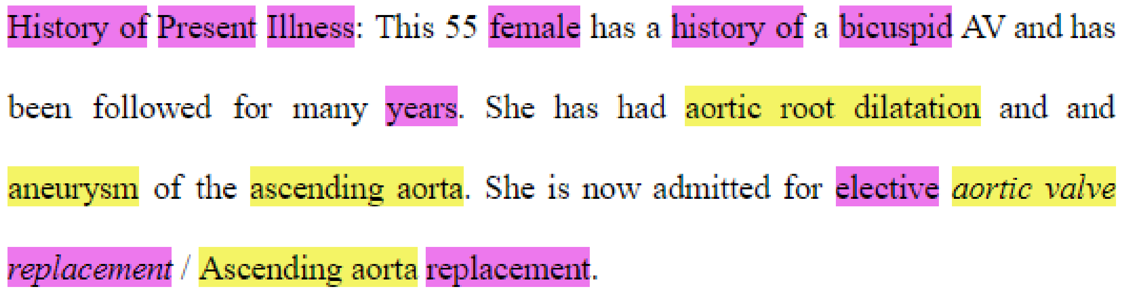
7.2.1 Auxiliary Concepts for Cardiology Description Interface Terminology

The clinical notes from the MIMIC-III database (discussed in Section 2.3.1) were used to extract auxiliary concepts for Cardiology Description Interface Terminology (CDIT). To retrieve clinical notes of cardiology patients, the patients who stayed in the two ICUs associated with cardiology - the Coronary Care Unit (CCU) and the Cardiac Surgery Recovery Unit (CSRU) were identified. The *Icustays* table in MIMIC-III defines every ICU stay for every patient. From this table, the subject ids (unique to every patient) along with the unique hospital admission ids for every patient who was admitted either to the CCU or the CSRU were selected. Then the clinical notes (available in the *Noteevents* table) for the above filtered (*subject_id*, *hospital_admission_id*) pairs were extracted. The extraction was limited to the category 'discharge summary.'

The text in the discharge summaries is organized under various section headers, such as the *history of present illness, family history, discharge medications,* etc. A review of the notes by a domain expert revealed that most of the cardiology-related concepts were found under the section headers related to the *history of present illness, past medical history,* *brief hospital course*, and *major surgical or invasive procedures*. Hence, only the textual information under the above mentioned four section headers (and their variations) was extracted and retained. Out of the 18,901 discharge summaries of patients who stayed in a CCU or CSRU, 500 discharge summaries were randomly chosen for this study. These 500 discharge summaries together constitute the dataset for mining concepts for a Cardiology Description Interface Terminology.

In the next step, the NCBO Annotator [98] was used to annotate cardiology EHR notes with relevant terminology concepts. The rationales for the use of the NCBO Annotator are the following: 1) Cardiology-specific annotated datasets are not publicly available; 2) The NCBO Annotator allows to annotate text with concepts from any of the many ontologies/terminologies available in BioPortal, and 3) Terminologies can be uploaded to BioPortal and then its concepts can be used for annotation by NCBO Annotator. This property is important for annotating clinical notes, as the interface terminology is expanded in different stages by adding new concepts at each stage.

To obtain the auxiliary concepts, the dataset is first annotated with concepts from SNOMED CT and then with concepts in ICDIT using the NCBO Annotator. A program that outputs only concepts annotated by SNOMED CT but not by ICDIT (i.e., it displays the difference in annotations) called DIFF was developed. An excerpt from one of the cardiology notes is shown in Figure 7.1, with concepts from ICDIT in yellow and concepts missing in ICDIT but present in SNOMED CT in pink. *History of*, *bicuspid*, *female*, etc. are examples of auxiliary concepts.



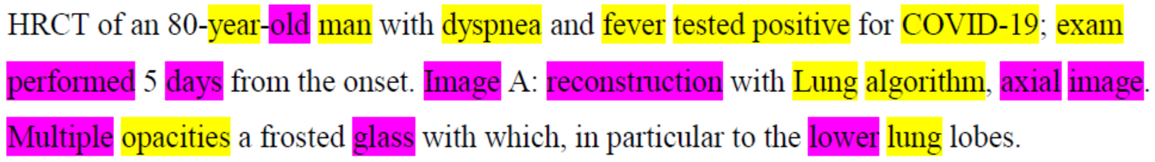
History of Present Illness: This 55 female has a history of a bicuspid AV and has been followed for many years. She has had aortic root dilatation and and aneurysm of the ascending aorta. She is now admitted for elective *aortic valve replacement* / Ascending aorta replacement.

Figure 7.1 An excerpt from a cardiology EHR, annotated by ICDIT (in yellow) and the DIFF (in pink: Concepts in SNOMED CT, missing in the ICDIT).

DIFF is applied to all cardiology notes in the dataset and the concepts obtained are added to separate auxiliary hierarchies and integrated with ICDIT to form the Cardiology Description Interface Terminology version 0 (CDIT_v0).

7.2.2 Auxiliary Concepts for COVID Interface Terminology

The construction of the COVID Interface Terminology version 0 (CIT_v0) follows similar steps. For extracting auxiliary concepts for the COVID Interface Terminology (CIT), we use the COVID-19 radiology case studies (discussed in Section 2.3.2). As previously discussed, DIFF is applied to the collection of 115 radiology case studies in the dataset to collect all auxiliary concepts and integrate them into ICIT to form a new interface terminology, the COVID Interface Terminology version 0 (CIT_v0). In this case, DIFF identifies all the concepts annotated with SNOMED CT that are not present in ICIT. Fig. 7.2 shows an excerpt from a radiology case study of a COVID-19 patient, annotated with ICIT (in yellow) and the DIFF (in pink). The concepts such as *old*, *reconstruction*, and *lower* are examples of auxiliary concepts that are needed even though they are not COVID specific.



HRCT of an 80-year-old man with dyspnea and fever tested positive for COVID-19; exam performed 5 days from the onset. Image A: reconstruction with Lung algorithm, axial image. Multiple opacities a frosted glass with which, in particular to the lower lung lobes.

Figure 7.2 A snippet from a radiology case study of a COVID-19 patient, annotated with ICIT (in yellow) and the DIFF (in pink) based on SNOMED CT.

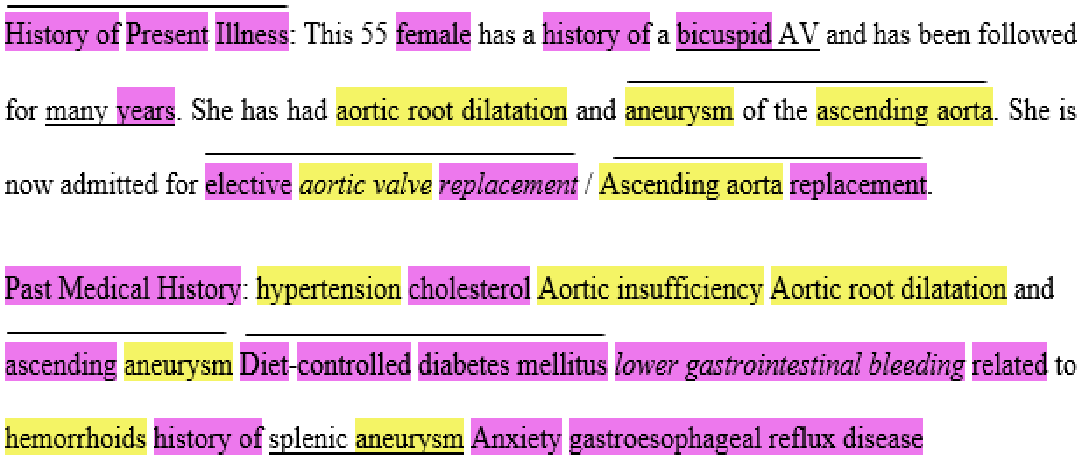
7.3 Concept Mining Techniques

Reference ontologies and terminologies often do not contain many of the fine granularity phrases that appear in EHR notes. Because of this, some critical information is lost during the annotation process. This issue can be addressed by mining concepts from the EHR itself. Thus, extracting high granularity concepts from EHR notes is one of the challenges to overcome for enriching interface terminologies with such essential concepts.

For addressing this challenge, the techniques of concatenation and anchoring of existing concepts are used. **Concatenation** involves combining two or more existing concepts that appear next to each other into a fine granularity phrase. Stop words are allowed in between existing concepts. **Anchoring** extracts phrases by adding one or two words to the left, right, or both sides of an existing concept, and stop words are allowed to intervene. For example, consider w1, w2 as two words, sw as a stop word, and we define * to mean 0, 1, or more occurrences [Kleene Star, in the theory of Algorithms], then the candidate anchoring phrases can be represented using the following three rules. The "+" stands for string concatenation.

1. w1 + sw* + [existing concept]
2. [existing concept] + sw* + w1
3. w1 + sw* + [existing concept] + sw* + w2

Both these techniques are illustrated in Figures 7.3 and 7.4 for cardiology and COVID-19, respectively.

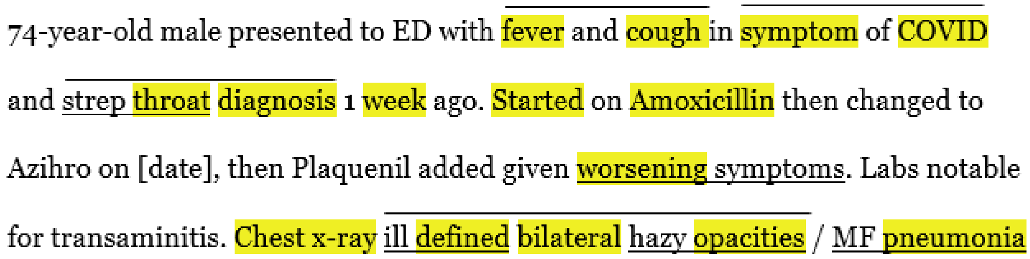


History of Present Illness: This 55 female has a history of a bicuspid AV and has been followed for many years. She has had aortic root dilatation and aneurysm of the ascending aorta. She is now admitted for elective *aortic valve replacement* / Ascending aorta replacement.

Past Medical History: hypertension cholesterol Aortic insufficiency Aortic root dilatation and ascending aneurysm Diet-controlled diabetes mellitus *lower gastrointestinal bleeding* related to hemorrhoids history of splenic aneurysm Anxiety gastroesophageal reflux disease

Figure 7.3 An excerpt from a cardiology note visualizing concatenation (overbars) and anchoring (underlines).

Figure 7.3 shows examples from a cardiology note, with anchoring marked by underlines and concatenation indicated by bars above the component concepts. *Aneurysm of the ascending aorta*, *aortic valve replacement*, and *elective aortic valve replacement* are examples of concatenation. *Aneurysm of the ascending aorta* is obtained by concatenating existing concepts *Aneurysm* and *ascending aorta* allowing the stop words "of" and "the" in between. The phrase *elective aortic valve replacement* is formed by the concatenation of three existing concepts. The phrase *splenic aneurysm* is obtained by Rule 1 of anchoring as the word splenic is added to the left of an existing concept *aneurysm*, and *bicuspid AV* is obtained by Rule 2.



74-year-old male presented to ED with fever and cough in symptom of COVID and strep throat diagnosis 1 week ago. Started on Amoxicillin then changed to Azihro on [date], then Plaquenil added given worsening symptoms. Labs notable for transaminitis. Chest x-ray ill defined bilateral hazy opacities / MF pneumonia

Figure 7.4 An excerpt from a synthetic EHR illustrating some example phrases obtained by concatenation and anchoring procedures. Overbars represent concatenation and underlines represent anchoring.

An excerpt from a synthetic note of a COVID-19 patient annotated with CIT_v0 is shown in Figure 7.4. Concatenation is marked by overbars and anchoring is marked by underlines. For example, the existing concepts *symptom* and *COVID* can be concatenated to form the phrase *symptom of COVID*. The concept *strep throat* is obtained by anchoring "strep" to the existing concept *throat*. In the next step, this new concept can be concatenated with *diagnosis* providing the phrase *strep throat diagnosis*. Similarly, the phrase *ill defined bilateral hazy opacities*, is obtained by first applying anchoring to get *ill defined* and *hazy opacities* and then by concatenating these phrases with *bilateral*.

Concept mining proceeds by applying the concatenation and anchoring procedures alternatingly on the dataset, annotating it with version 0 of the interface terminology (CDIT_v0 or CIT_v0). To be explicit, the dataset is first annotated with CIT_v0 (or CDIT_v0) concepts. Then concatenation is applied. Those phrases that are accepted by a human expert are then added to CIT_v0 (or CDIT_v0) to obtain CIT_v1.1 (or CDIT_v1.1). Next, the dataset is annotated with CIT_v1.1 (or CDIT_v1.1), and anchoring is applied to obtain more candidate phrases. The phrases accepted by the expert are added to CIT_v1.1 (or CDIT_v1.1) to obtain CIT_v1.2 (or CDIT_v1.2). This process is continued, alternating between concatenation and anchoring. The advantage of alternating the concatenation and anchoring steps is that the phrases obtained by concatenation can participate in anchoring in the next step and vice versa. For example, *strep throat* mentioned above is obtained as a concept in CIT_v1.2 by anchoring and is used in the subsequent concatenation phase with the concept *diagnosis* to obtain the phrase *strep throat diagnosis* as a concept in CIT_v2.1. Since concatenation and anchoring are brute-force techniques, human review is necessary. This review process and the metrics used for evaluation are discussed in the following section.

7.4 Review Process and Evaluation Metrics

After each application of concatenation and anchoring, the extracted phrases were reviewed in a two-step process. Concepts were first prescreened by the core team and then the accepted candidates were reviewed by a medical expert. Prescreening was possible, because most of the automatically generated phrases were parts of a larger phrase or spanned two partial phrases. For example, the obtained phrase "thickening of pulmonary" was a part of *thickening of pulmonary interstitium,* and "MRSA and port-a-cath" spanned two phrases, *sepsis from MRSA* and *port-a-cath infection*. All the phrases that passed both review steps were integrated into the version of the interface terminology that was current at that time.

It should be noted that all the phrases that were rejected at any review step were automatically excluded from the candidate phrases list and never appeared again in the subsequent processing steps. Hence, each rejected phrase is reviewed only once, saving review time. Similarly, the accepted phrases were integrated into the interface terminology as concepts and used for annotation in the next iteration. Thus, they cannot appear again as candidate phrases. Therefore, the number of extracted phrases decreases significantly after each application of concatenation and anchoring. After a few iterations, when the number of new phrases falls below a threshold, the processing is terminated.

After the domain expert review, a synonym check was performed on the accepted phrases. Phrases that were synonyms were combined under a single concept ID. One phrase was chosen as the concept name and the other phrases as synonyms. This is exemplified by the two phrases *history positive for contact with COVID-19 patient* and *positive history of contact with COVID-19 patient*.

As a safeguard against false negatives at the first review step (by the core team), a sample of 200 phrases was created, selected randomly from the rejected phrases in all the iterations. This sample was then reviewed by the domain expert to check for cases of false negatives, i.e., acceptable phrases that were rejected.

Evaluation metrics: The performance of the techniques was evaluated using two metrics – Coverage and Breadth. **Coverage** is the percentage of words being annotated. **Breadth** is the average number of words per annotated concept. As interface terminologies have more fine granularity phrases compared to concepts in reference terminologies the breadth increases.

$$Coverage = \frac{Number\ of\ words\ in\ all\ annotated\ concepts}{Total\ number\ of\ words} * 100 \tag{7.1}$$

$$Breadth\ = \frac{Number\ of\ words\ in\ all\ annotated\ concepts}{Number\ of\ annotated\ concepts} \tag{7.2}$$

It should be noted that metrics such as precision, recall, etc. are not applicable for this study, due to the lack of gold standard annotations. One of the main goals of the study was to find an alternative way to reduce the expensive and time-consuming manual annotation process. A detailed discussion regarding this is provided in Section 7.5. As the interface terminologies will have more concepts and richer synonym content compared to reference ontologies/terminologies, using these interface terminologies as the basis for the annotation process would yield better efficacy.

7.5 Results

In the study mining concepts for CDIT, only a single iteration of concatenation and anchoring was performed. For mining concepts related to COVID-19, multiple iterations of the two techniques were performed, until convergence in terms of the number of phrases accepted was obtained, meaning that additional iterations increased the number of terms only minimally. This section describes the results for both CDIT and CIT.

7.5.1 Results for Cardiology Description Interface Terminology

To create ICDIT, root concepts of eleven subhierarchies in SNOMED CT with cardiology-related concepts were identified. Each such root is listed below with the SNOMED CT hierarchy it belongs to (within parenthesis) and the number of concepts in the subhierarchy: *Cardiovascular agent* (*substance*) (533), *Cardiovascular equipment* (*physical object*) (343), *Cardiovascular event* (*event*) (1), *Cardiovascular finding* (*finding*) (7723), *Cardiovascular material* (*substance*) (30), *Cardiovascular observable* (*observable entity*) (536), *Cardiovascular sample* (*specimen*) (39), *Procedure on cardiovascular system* (*procedure*) (6208), *Structure of cardiovascular system* (*body structure*) (4418), *Cardiovascular implant* (*physical object*) (135), and *Finding present on electrocardiogram* (*finding*) (269). In addition to these concepts, 3346 concepts were obtained by DIFF with SNOMED CT and placed into several auxiliary hierarchies. Thus, there are 23,517 concepts in CDIT_v0.

Table 7.1 Examples of High-Frequency Phrases Extracted by Concatenation (Annotated Concepts are Separated by '|') and by Anchoring

	Frequency	**Concatenation of phrases**
Cardiology Concepts	112	Cardiac Care | Unit
	84	Ejection | Fraction of
	66	Chest | X-ray
	53	Cardioverter | Defibrillator
	51	Aortic valve | replacement
General Concepts	455	History of | Present | illness
	98	Postoperative | day
	76	Outside | hospital
	72	Stable | condition
	30	Emergency | medical service
	Frequency	**Anchoring at phrases**
Cardiology Concepts	57	implantable **cardioverter defibrillator**
	42	**Beta** blocker
	41	**Stress** test
	28	**computerized tomography** scan
	22	**Heparin** drip
General Concepts	493	**Hospital** course
	62	**Emergency** room
	58	**Years** ago
	41	**Emergency** department
	40	**Neurologically** intact

To mine fine granularity concepts, concatenation was applied to concepts in CDIT_v0 and 6042 fine granularity phrases were obtained. Among these, 1044 phrases occurring multiple times were selected and reviewed by the domain experts. They accepted 770 phrases out of which 116 phrases were synonyms of existing concepts. Thus, 654 new concepts and 116 synonyms were added to the CDIT_v0 to obtain CDIT_v1.1 Next, anchoring was performed with this new version of CDIT containing 24,203 concepts. This yielded a total of 21,083 new phrases. Among these, 3013 phrases with multiple occurrences were selected. The experts accepted 1409 phrases, 95 of which were synonyms of existing concepts. Thus, the CDIT-1.2, the version of CDIT after the first iteration of concatenation and anchoring contained 25,485 concepts. Table 7.1 shows examples of high-frequency phrases obtained by concatenation and anchoring. In Table 7.2, examples of accepted and rejected concepts obtained by concatenation and anchoring are shown.

Table 7.2 Examples of Extracted Phrases Obtained by Concatenation and Anchoring Reviewed by the Experts

	Frequency	**Accepted phrases**
Concepts created by concatenation	26	aortic stenosis
	24	ST elevation myocardial infarction
	22	history of coronary artery disease
	17	mitral valve repair
	15	peripherally inserted central catheter line
Concepts created by anchoring	14	diastolic congestive heart failure
	13	epicardial pacing wires
	5	macular degeneration
	2	antitachycardia pacing
	2	coronary revascularization
	Frequency	**Rejected phrases**
Concepts created by concatenation	228	Major surgical
	14	left internal
	9	low dose beta
	8	left lower
	6	bilateral pleural
Concepts created by anchoring	25	hypertension hypercholesteremia
	13	saphenous vein graft to obtuse
	3	hemorrhoids Meckel
	8	wave myocardial infarction
	3	left with a thoracoacromial

Figures 7.5 and 7.6 show the distribution of the phrases obtained by frequency. The bar chart in Figure 7.5 shows the number of extracted phrases (by concatenation), divided

into accepted and rejected, for different frequencies of occurrence. The average acceptance rate across all frequencies is about 74%.

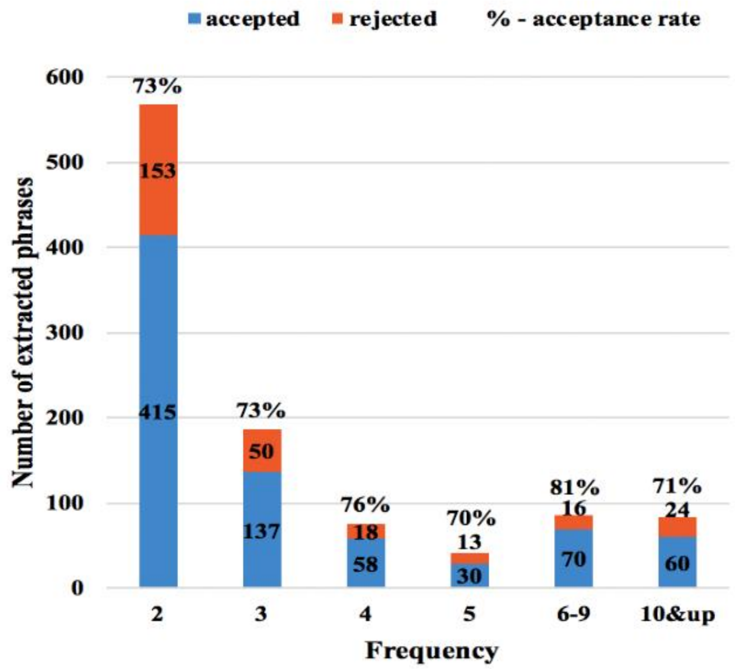


Figure 7.5 Number of phrases extracted for the reviewed frequency (or range), divided into accepted and rejected phrases for concatenation.

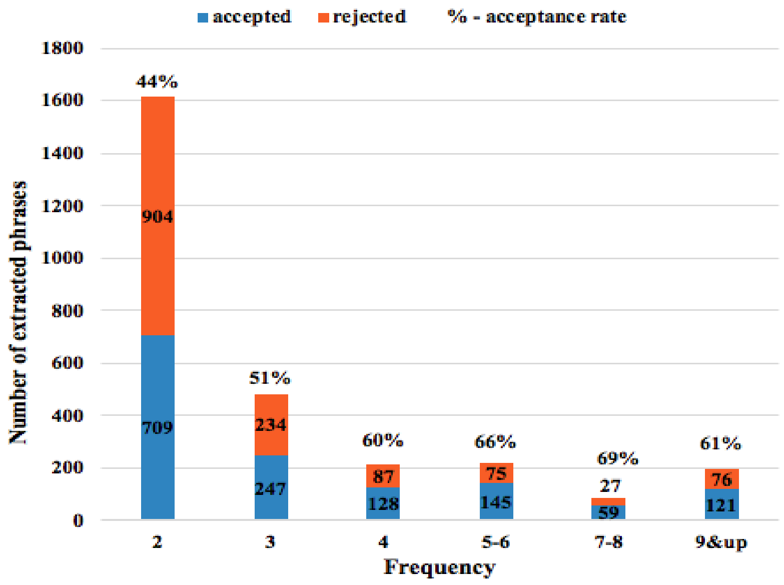


Figure 7.6 Number of phrases extracted for the reviewed frequency (or range), divided into accepted and rejected phrases for anchoring.

The bar chart in Figure 7.6 shows the same data for anchoring. It shows an increase in the acceptance rate with the increase of frequency leveling off at about 65% for concepts occurring more than five times. The coverage and breadth were computed for SNOMED CT and CDIT-1.2. For the dataset of 500 cardiology notes, the total number of words was 171,777. The coverage was 39.35% for SNOMED CT and 43.04% for CDIT-1.2. The breadth was 1.38 for SNOMED CT and 1.71 for CDIT-1.2.

7.5.2 Results for COVID Interface Terminology

To create the ICIT, concepts were integrated from other COVID ontologies into CIDO. After removing duplicates, 1780 concepts from the COVID-19 ontology were added into CIDO, as well as 352 concepts from IDO-COVID-19, 272 concepts from COVIDCRFRAPID, 46 concepts from CODO, and 50 concepts from the COVID-19 Surveillance Ontology. From the ACT COVID ontology, a total of 2445 concepts were included. In addition to this, 113, 74, and 2 concepts from SNOMED, LOINC, and UMLS, respectively were incorporated. After identifying and accounting for synonyms, the total number of concepts in ICIT at this stage was 10,024. For creating the CIT_v0, auxiliary concepts from SNOMED CT were integrated into ICIT. Integrating 904 auxiliary concepts, the total number of concepts in CIT_v0 increased to 10,928.

To mine new phrases for the CIT, the dataset was annotated with the NCBO Annotator and concatenation and anchoring were applied alternatingly. The numbers of extracted phrases, the numbers of phrases retained after the core team (1st) reviews and after the expert (2nd) reviews, and the corresponding percentages are shown in Table 7.3 for all the versions of CIT created. The last column of the table shows the percentages of phrases that were retained by the expert with respect to the percentages from the core team review.

Table 7.3 Statistics of Extracted Phrases for all Versions of CIT

Version of CIT	**Procedure**	**Total number of phrases**	**Number of phrases after 1st review**	**Percentage phrases retained after 1st review**	**Number of phrases after 2nd review**	**Percentage phrases after 2nd review**	**Percentage retained by expert with respect to 1st review**
v1.1	Concatenation	1893	873	46.12%	781	41.25%	89.5%
v1.2	Anchoring	3923	1590	40.53%	1351	34.44%	84.97%
v2.1	Concatenation	1002	439	43.81%	389	38.82%	88.6%
v2.2	Anchoring	969	295	30.44%	268	27.66%	90.86%
v3.1	Concatenation	314	92	29.30%	83	26.43%	90.20%
v3.2	Anchoring	185	34	18.37%	30	16.21%	88.24%
v4.1	Concatenation	66	6	9.09%	6	9.09%	100%
v4.2	Anchoring	69	6	8.69%	6	8.69%	100%

In Table 7.4, some examples of phrases obtained as a result of concatenation during the creation of CIT_v1.1 are shown. The existing concepts that were combined to form the fine granularity phrases are shown between two '|' symbols. The phrase *tested positive for COVID-19* was obtained by combining two existing concepts *tested positive* and *COVID-19,* allowing for the stop word "for." Another example, *history of contact with Covid-19 patient* is a combination of four existing concepts, as shown in the third row in Table 7.4. Some example phrases obtained by applying anchoring that were accepted for inclusion in CIT_v1.2 are also shown. The existing concept that was used as an anchor is marked in bold. The phrase *subpleural distribution* is an example of the first rule of anchoring, where a left word was added to the existing concept *distribution*. The phrase, *mediastinal lymphadenomegalies* is the result of applying the second rule of anchoring, which adds a right word; *ground glass areas* demonstrate the third rule of combining both left and right words with an existing concept.

Table 7.4 Examples of Accepted Phrases from CIT_v1

Version	Accepted Phrases
CIT_v1.1 (concatenation)	|tested positive| for |COVID-19|
	|history of| |contact with| |Covid-19| |patient|
	|peri|-|bronchial| |thickening|
	|limited| |lymphadenopathy|
	|spider| |web| |sign|
CIT_v1.2 (anchoring)	subpleural **distribution**
	mediastinal lymphadenomegalies
	parenchymal **thickening**
	interstitial-alveolar **pneumonia**
	ground **glass** areas

Examples of rejected phrases are shown in Table 7.5 for both concatenation and anchoring. As in Table 7.4, for concatenation, the existing concepts are between two '|' symbols, and for anchoring, they are marked in bold. As discussed in Section 7.3, there are phrases that are part of longer phrases (e.g., partial pleurogenic) or spanning two phrases (e.g., axis with Fogarty catheter).

Table 7.5 Examples of Rejected Phrases

Procedure	Rejected Phrases
Concatenation	|hyperpyrexia| |refractory|
	|axis| with |Fogarty catheter|
	|thrombocytopenia| and |need for|
	|chest x-ray| with |multiple|
	|pneumonia| with |radiographic|
Anchoring	**Multiple opacities** a frosted
	increased density with a ground
	segment of the **upper**
	Partial pleurogenic
	reticular and interstitial

Examples of longer phrases that were added to CIT_v2 and CIT_v3 are provided in Table 7.6. For example, the phrase *extensive areas with crazy-paving patterns* was obtained for addition to CIT_v2.1 as a result of concatenating two phrases - *extensive areas,* and *crazy-paving patterns* that were already in CIT_v1.1. Similarly, the phrase *subpleural distribution* present in CIT_v1.2 was used as an anchor to extract *predominantly subpleural distribution* for inclusion in CIT_v2.2.

Table 7.6 Examples of Accepted Longer Phrases

Accepted Longer Phrases
inter- and intra-lobular septal thickening
extensive areas with crazy-paving patterns
parenchymal consolidation area with subpleural distribution
bilateral subpleural ground glass opacities
widespread fibrotic-like reticular bands
parenchymal consolidations in both upper lobes
predominantly subpleural distribution
centrolobular and subpleural paraseptal emphysema

To check for false negatives among the phrases that were rejected by the core team, a random sample of 200 rejected phrases was created from all the iterations. This sample was reviewed by the domain expert. The domain expert found only eight out of the 200 phrases to be false negatives.

Table 7.7 Coverage and Breadth for Different Versions of CIT

Version	**Number of concepts**	**Coverage**	**Breadth**
CIT_v1.1	11,644	41.30%	1.55
CIT_v1.2	12,984	53.66%	2.16
CIT_v2.1	13,364	53.97%	2.47
CIT_v2.2	13,628	58.09%	2.65
CIT_v3.1	13,711	58.19%	2.73
CIT_v3.2	13,741	58.41%	2.74
CIT_v4.1	13,747	58.42%	2.74
CIT_v4.2	13,753	58.46%	2.74

The dataset was annotated with CIDO, ICIT, and CIT_v0, obtaining coverages of 6%, 13.55%, and 40.84%, respectively. The breadths were 1.18, 1.22, and 1.21, respectively. The coverage and breadth for different versions of CIT are shown in Table 7.7. The number of concepts in each version is also shown.

7.6 Discussion

By demonstrating the iterative terminology construction approach for two domains it may be expected that this approach would work for any medical subspecialty. The approach presented in both studies is based on several assumptions. The first assumption is that by mining concepts from clinical notes of cardiology and COVID-19 patients and including them in the interface terminology, many concepts that correspond to fine granularity concepts recorded by healthcare professionals can be obtained. Numerous such concepts do not appear in reference terminologies. However, most of the multi-word names of such concepts tend to contain shorter concepts that do appear in the reference terminologies. Thus, the operations of anchoring and concatenation of existing concepts in the interface terminology can be used to obtain fine granularity concepts.

The only manual step that requires a domain expert is reviewing which of the generated phrases obtained by concatenation or anchoring are valid for inclusion in the interface terminology. As was previously discussed, each phrase is reviewed only once during the life cycle of creating the interface terminology. Hence, despite the manual review required, the creation of the interface terminology is efficient. Once the interface terminology has been created, it can be used for annotation of an unlimited number of clinical notes of cardiology and COVID-19 patients.

The second assumption is that the described process will converge. That means that on average later steps will find fewer and fewer new phrases than those preceding them.

Although individual doctors write free text in individualized ways, the number of possible concepts used in a specific discipline is fundamentally limited and is growing slower as the field matures. Different healthcare professionals will use similar, but not identical, phrases to express the same concept. One of those will be designated the name of the concept and all the others will be synonyms of the concept in the interface terminology. When annotating text with the interface terminology, the synonyms are also identified and annotated. A second reason for assuming a limited number of phrases in each discipline is that the length of the phrases is limited to a few words. Thus, exponential growth in the number of new terms is less likely than a polynomial increase, which is considered as manageable growth in the theory of computing.

One argument by the ontology curators for not including new concepts is that they have no known use case, are too fine-grained, would "clutter up the ontology," and make maintenance more difficult. This may be true in the case of some reference ontologies, which represent knowledge about an entire biomedical domain. But the "damage" for a medical user not finding a desired concept in an interface terminology is bigger than for another user having to ignore an additional concept.

As mentioned before, the results of concatenation and anchoring need to be reviewed by a human expert. The time of domain experts is a limited and expensive resource. To minimize the use of this resource, a preliminary review was performed by members of the SABOC research team [154], after having gone through training based on samples that were previously reviewed by domain experts. The main purpose of this preliminary review was to exclude phrases that obviously should not be part of the CIT. Only the phrases accepted by the core team in the preliminary review were passed on to the domain expert for validation. According to Table 7.3, for the first six versions of CIT, on average about 88% of the phrases accepted by the core team were also validated by the domain expert. Hence it is safe to say that an initial review by the core team helped to eliminate a large number of phrases that were not corresponding to concepts, thereby minimizing the time and effort expended by the domain expert.

The concatenation and anchoring operations have a different impact on the evaluation metrics. Concatenation provides a minimum contribution to the coverage of the dataset. This is because concatenation combines already existing annotated concepts and hence does not add new words to the annotated word list, except for the stop words that bridge the gaps between existing concepts. There is only a 0.31% change in coverage obtained by concatenation in CIT_v2.1 compared to the coverage of the previous version CIT_v1.2. Similarly, the change in coverage is only 0.10% when moving from version CIT_v2.2 to CIT_v3.1. Concatenation favors the breadth metric, since breadth increases with the length of the phrases, and combining existing phrases creates longer phrases.

Anchoring tends to increase both coverage and breadth. During anchoring "unannotated" words are added to the left, right, or on both sides of a concept, and hence accepted phrases resulting from anchoring capture words that were previously not annotated and contribute to increasing the number of annotated words, thereby increasing coverage. Anchoring also increases breadth, as the newly added words increase the length of the phrases annotated by specific concepts. The increases in coverage obtained by the first three anchoring iterations are 12.36%, 4.12%, and 0.22%. The corresponding increases in breadth are 0.61, 0.18, and 0.01.

As noted above, the methodology described for designing an interface terminology to support annotation of EHR notes is applicable to other medical specialties, allowing the design of a dedicated interface terminology for each individual specialty, as there will be some SNOMED CT terms for every specialty (as in the cardiology case study). If a dedicated reference terminology exists as in the COVID-19 case study the task becomes easier.

CHAPTER 8

FUTURE WORK

In the studies based on density differences between identical concepts in pairs or triples of terminologies, the Concept Unique Identifier (CUI) of the concepts in the UMLS Metathesaurus was used to determine whether the concepts are identical. This has a limitation that only vocabularies in the UMLS could be used as source and target terminologies for the study. Several ontologies and terminologies that have a hierarchical IS-A backbone exist that are not included in the UMLS.

In future, the density studies can be extended to include those ontologies and terminologies. The challenge for studies based on horizontal density differences would be to identify identical parent concepts in two different terminologies. Techniques based on lexical and semantic similarity need to be explored to identify overlapping concepts in different ontologies.

The fire ladder pattern has scope for further expansion. In the pattern explored in the study described in Chapter 5 in Figure 5.1, Concept B2 is an immediate child of Concept B1. But there always exists a possibility that there can be one or more intermediate concepts between B1 and B2. This opens up the possibility of importing more concepts into Terminology A. Also, since Concept B1 is not present in Terminology C, theoretically B1 can be imported into Terminology C as a parent of Concept C2. But further extensive studies need to be performed regarding the existing parent of C2 and how this concept relates to B1.

All studies discussed in Chapters 3-5 utilize the parent-child hierarchical relationship. Even though the most studied relationships in terminologies are IS-A relationships, several lateral relationships exist. IS-A relationships are hierarchical. Another hierarchical relationship is the *part-of* (part-whole) relationship. It would be interesting to explore the structural differences in this relationship between terminologies in similar domains, especially in anatomy.

The creation of a cardiology description interface terminology was performed to demonstrate how to start the process based only on SNOMED CT. Hence, only one iteration was performed. The study on the COVID interface terminology had a relatively small dataset due to the non-availability of large public COVID-related EHR datasets. In the future, an extensive study will be conducted that will test the assumption that after a few iterations the process will converge even for larger datasets. Only a few new concepts will be added in later iterations. For this, the dataset will be randomly divided into two parts, the "training" set, and the test set (80-20 split). The interface terminology will be created based on the training set and then it will be used to annotate the test set.

The hypothesis is that the coverage and breadth values for the test dataset will be marginally smaller than that for the training set, but almost on par, which would provide one data point to demonstrate the generalizability of the approach.

For the study with a larger dataset, only phrases that appear multiple times in the dataset will be reviewed. These phrases are more likely to appear in the test dataset than those that appear only a few times. Another hypothesis to be tested is that the phrases with higher frequency in the dataset will have higher rates of acceptance by the expert. To test this hypothesis, a study of the correlation between frequency and acceptance rate will be performed in future work.

CHAPTER 9

CONCLUSIONS

With the exponential growth of knowledge in the life sciences and the technological advancements in all aspects of the biomedical sciences, biomedical ontologies and terminologies have grown rapidly to facilitate the systematic storage and retrieval of this knowledge. Even though the benefits of extending biomedical ontologies appear evident, it has been argued by some in the biomedical ontology community that bigger is not necessarily better. However, many major ontologies and terminologies have been growing monotonically for the past several years. That means that every release in recent years has contained more concepts than the previous release. This has been the case for the SNOMED CT, with more than 50,000 concepts added in the past five years [155]. Similarly, more than 40,000 concepts have been added to NCIt [156].

If ontologies are demonstrably extended "anyway," they should be extended in a systematic process that leads to more harmonization between major, widely used ontologies in the field. The studies described in Chapters 3-5 of this dissertation focused on methodologies that identify missing concepts in an ontology in a way that leads to better harmonization among ontologies. This was done by identifying structural and domain similarities among ontologies and creating topological patterns that leverage these structural properties for identical concepts (based on the UMLS Concept Unique Identifiers) in pairs of ontologies or triples of ontologies.

The second part of this dissertation (Chapter 7), extended ontologies by mining fine granularity phrases from unstructured text in EHRs assigning them to concepts or synonyms and creating an interface terminology that improves annotation of EHRs in a medical specialty. This approach has been particularly successful for the COVID Interface Terminology identifying several important concepts that were neither present in a reference ontology nor in several COVID-specific ontologies.

Chapter 3 presented an algorithm to identify missing child concepts in SNOMED CT and NCIt, based on a set of other sources in the UMLS. A quantitative analysis of the identified concepts was performed, and the results were verified by a domain expert. The hypothesis that algorithmically proposed new children are distinguishable from cousins was supported with statistical significance for both SNOMED CT and NCIt.

In Chapter 4, a detailed analysis of alternative classifications and the *EFI* (Evidence for Import) metric to indicate when concepts with a common parent should be imported into a target terminology was presented. When *EFI* becomes zero, it provides good evidence that the occurrences of the same concept in two terminologies define an alternative classification. In 72% of the cases, the domain expert agreed that the child concepts in the source and target terminologies are alternative classifications of the same parent concept. In contrast, for *EFI* values in the range of 0.10 – 0.35, it was found that direct import was the most likely choice, with statistical significance, when compared to a control sample. As the domain expert agreed with 80% of the algorithmically recommended imports, the proposed metric outperformed the approach in Chapter 3 where only 56% of imports were recommended by the domain expert.

Chapter 5 demonstrated a novel topological pattern called *fire ladder* and an algorithm to discover such patterns in triples of terminologies to help identify potentially missing concepts in 10 UMLS terminologies. This pattern consists of two source terminologies used in tandem and one target terminology. A total of 55 instances of fire ladder patterns were identified, out of which two experts agreed on 39 instances of concept imports. For 48 (=39+9; 87%) instances at least one expert agreed that the algorithm reported a viable import. Furthermore, the import of 98 additional concepts out of 105 algorithmically discovered candidate concepts was recommended, based on only one source terminology and one target terminology.

A supplementary study performed to identify reasons for not maintaining ontologies and ensuring that the ontologies and terminologies analyzed in this study do not fall under this category are presented in Chapter 6. Out of 83 ontologies of a minimum size with a defined period of no updates, a response to an email inquiry was received from 48 curators. Of these, 46 could be assigned to seven categories that were created a posteriori. Half (24) of these responses fit into two categories: Either there was a lack of funding or manpower that precluded continuation of maintenance (15), or the ontology had been folded into another ontology or system (9) and was not developed by itself anymore. With 15 out of 48 projects (31.25%) being discontinued due to budget or manpower problems, the limited availability of federal funding for sustaining these projects might have been a decisive issue.

Chapter 7 presented two studies describing an alternative semiautomatic approach for creating two interface terminologies (CDIT and CIT) for annotation of cardiology EHR notes and EHRs of COVID-19 patients. Repeated alternating applications of the two operations, concatenation, and anchoring applied to a sample of 500 notes from MIMIC III was used to enrich the initial CDIT extracted from SNOMED CT. A preliminary study for CDIT achieved in the first iteration about 10% and 24% higher annotation coverage and breadth, respectively, relative to annotation with SNOMED CT. By choosing a larger sample and iterating over the process, substantial improvements can be expected in subsequent iterations. CIT was initialized with concepts from several COVID ontologies and with existing general-purpose concepts (non-COVID concepts) from SNOMED CT encountered in the dataset. Its content was significantly extended by mining fine granularity concepts from the radiological studies in the available dataset. Version 4.2 of the CIT (CIT_v4.2) achieved a 43% increase in coverage compared to CIT_v0.

REFERENCES

[1] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of Medical Informatics.2008:67-79.

[2] Bodenreider O, Mitchell JA, McCray AT. Biomedical ontologies. Pacific Symposium on Biocomputing. 2005:76-8.

[3] Epic: with the patient at the heart. https://www.epic.com/, (accessed 22 Feb 2021).

[4] Cerner: Hospitals and health systems. https://www.cerner.com/solutions/health-systems, (accessed 22 Feb 2021).

[5] Allscripts. https://www.allscripts.com/ehrs/, (accessed 22 Feb 2021).

[6] Finnegan R. ICD-9-CM coding for physician billing. Journal of the American Medical Record Association. 1989;60:22-3.

[7] Hirsch JA, Leslie-Mazwi TM, Nicola GN, Barr RM, Bello JA, Donovan WD, et al. Current procedural terminology; a primer. Journal of Neurointerventional Surgery. 2015;7:309-12.

[8] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in Health Technology and Informatics. 2006;121:279-90.

[9] McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical Chemistry. 2003;49:624-33.

[10] SNOMED CT and COVID-19. https://www.snomed.org/snomed-ct/covid-19, (accessed 22 Feb 2021).

[11] SARS-CoV-2 and COVID-19 related LOINC terms. https://loinc.org/sars-cov-2-and-covid-19/, (accessed 22 Feb 2021).

[12] Grabar N, Hamon T, Bodenreider O. Ontologies and terminologies: continuum or dichotomy? Journal of Applied Ontology. 2012;7:375-86.

[13] Zemmouchi-Ghomari L, Ghomari AR. Ontology versus terminology, from the perspective of ontologists. International Journal of Web Science. 2012;1:315-31.

[14] Guarino N. Understanding, building and using ontologies. International Journal of Human-Computer Studies. 1997;46:293-310.

[15] Whetzel P, Shah N, Noy N, Dai B, Dorf M, Griffith N, et al. BioPortal: Ontologies and integrated data resources at the click of a mouse. Nature Precedings. 2009.

[16] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research. 2011;39:W541-W5.

[17] Salvadores M, Alexander PR, Musen MA, Noy NF. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. Semantic Web. 2013;4:277-84.

[18] NCBO BioPortal. BioPortal. https://bioportal.bioontology.org/, (accessed 22 Feb 2021).

[19] Bodenreider O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research. 2004;32:D267-D70.

[20] Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html, (accessed 3 Apr 2020).

[21] Sioutos N, Coronado Sd, Haber MW, Hartel FW, Shaiu W-L, Wright LW. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics. 2007;40:30-43.

[22] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics. 2000;25:25-9.

[23] BioPortal. Medical Subject Headings. https://bioportal.bioontology.org/ontologies/MESH, 2018 (accessed 3 Dec 2018).

[24] Goltra PS. MEDCIN : a new nomenclature for clinical medicine. New York, NY: Springer-Verlag; 1997.

[25] He Z, Keloth VK, Chen Y, Geller J. Extended analysis of topological-pattern-based ontology enrichment. IEEE International Conference on Bioinformatics and Biomedicine. Madrid, Spain 2018. p. 1641-8.

[26] He Z, Geller J, Elhanan G. Categorizing the relationships between structurally congruent concepts from pairs of terminologies for semantic harmonization. AMIA Summits on Translational Science: American Medical Informatics Association; 2014. p. 48-53.

[27] He Z, Geller J, Chen Y. A comparative analysis of the density of the SNOMED CT conceptual content for semantic harmonization. Artificial Intelligence in Medicine. 2015;64:29-40.

[28] Rector A, Rogers J, Bittner T. Granularity, scale and collectivity: when size does and does not matter. Journal of Biomedical Informatics. 2006;39:333-49.

[29] Geller J, Keloth VK, Musen MA. How sustainable are biomedical ontologies? AMIA Annual Symposium. 2018;2018:470-9.

[30] Keloth VK, He Z, Chen Y, Geller J. Leveraging horizontal density differences between ontologies to identify missing child concepts: A proof of concept. AMIA Annual Symposium 2018. p. 644-53.

[31] Keloth VK, He Z, Elhanan G, Geller J. Alternative classification of identical concepts in different terminologies: different ways to view the world. Journal of Biomedical Informatics. 2019;94:103193.

[32] Keloth VK, Geller J, Chen Y, Xu J. Extending import detection algorithms for concept import from two to three biomedical terminologies. BMC Medical Informatics and Decision Making. 2020;20:1-11.

[33] Keloth VK, Zhou S, Lindemann L, Elhanan G, Einstein AJ, Geller J, et al. Mining concepts for a COVID interface terminology for annotation of EHRs. IEEE International Conference on Big Data (Big Data): IEEE; 2020. p. 3753-60.

[34] Keloth VK, Zhou S, Einstein AJ, Elhanan G, Chen Y, Geller J, et al. Generating training data for concept-mining for an 'interface terminology' annotating cardiology EHRs. IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2020. p. 1728-35.

[35] Shen F, Lee Y. Knowledge discovery from biomedical ontologies in cross domains. PloS One. 2016;11:e0160005.

[36] Liu H, Dou D, Jin R, LePendu P, Shah N. Mining biomedical ontologies and data using RDF hypergraphs. 12th International Conference on Machine Learning and Applications: IEEE; 2013. p. 141-6.

[37] Estival D, Nowak C, Zschorn A. Towards ontology-based natural language processing. Workshop on NLP and XML (NLPXML-2004): RDF/RDFS and OWL in Language Technology2004. p. 59-66.

[38] Zaihrayeu I, Sun L, Giunchiglia F, Pan W, Ju Q, Chi M, et al. From web directories to ontologies: natural language processing challenges. Semantic Web. 2007:623-36.

[39] Gai K, Qiu M, Chen L-C, Liu M. Electronic health record error prevention approach using ontology in big data. IEEE 17th International Conference on High Performance Computing and Communications, IEEE 7th International Symposium on Cyberspace Safety and Security, and IEEE 12th International Conference on Embedded Software and Systems: IEEE; 2015. p. 752-7.

[40] Martínez-Costa C, Schulz S. Validating EHR clinical models using ontology patterns. Journal of Biomedical Informatics. 2017;76:124-37.

[41] Blobel B, Kalra D, Koehn M, Lunn K, Pharow P, Ruotsalainen P, et al. The role of ontologies for sustainable, semantically interoperable and trustworthy EHR solutions. Studies in Health Technology and Informatics. 2009;150:953.

[42] U.S. National Library of Medicine. SNOMED CT. https://www.nlm.nih.gov/healthit/snomedct/, (accessed 3 Dec 2018).

[43] SNOMED International. SNOMED CT Managed Service - US Edition Release Notes - March 2021. https://confluence.ihtsdotools.org/display/RMT/SNOMED+CT+Managed+Service+-+US+Edition+Release+Notes+-+March+2021, (accessed 3 Mar 2021).

[44] U.S. National Library of Medicine. The U.S. SNOMED CT content request system (USCRS). https://uscrs.nlm.nih.gov/, (accessed 22 Feb 2018).

[45] National Cancer Institute. Enterprise vocabulary service – term suggestion. https://ncitermform.nci.nih.gov/ncitermform/?version=cdisc, (accessed 3 Dec 2018).

[46] The National Cancer Institute. NCI thesaurus. https://ncit.nci.nih.gov/ncitbrowser/, (accessed 20 Feb 2020).

[47] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics. 2005;38:114-29.

[48] National Cancer Institute. NCI term browser. https://nciterms.nci.nih.gov/ncitbrowser/pages/, (accessed 25 Mar 2019).

[49] Medicomp Systems Inc. Medicomp systems. http://www.medicomp.com/, (accessed 25 Mar 2019).

[50] He Y, Yu H, Ong E, Wang Y, Liu Y, Huffman A, et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific Data. 2020;7:1-5.

[51] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;25:1251-5.

[52] Coronavirus infectious disease ontology. https://github.com/CIDO-ontology/cido, (accessed 28 Feb 2021).

[53] Coronavirus infectious disease ontology on BioPortal. https://bioportal.bioontology.org/ontologies/CIDO, (accessed 28 Feb 2021).

[54] Coronavirus infectious disease ontology on Ontobee. http://www.ontobee.org/ontology/CIDO, (accessed 28 Feb 2021).

[55] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. Nucleic Acids Research. 2016;44:D1214-9.

[56] Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, et al. The Human Phenotype Ontology in 2017. Nucleic Acids Research. 2017;45:D865-D76.

[57] Brown SH, Elkin PL, Rosenbloom ST, Husser CS, Bauer BA, Lincoln MJ, et al. VA National Drug File Reference Terminology: a cross-institutional content coverage study. MEDINFO. 2004;11:477-81.

[58] Hanna J, Joseph E, Brochhausen M, Hogan WR. Building a drug ontology based on RxNorm and other sources. Journal of biomedical semantics. 2013;4:44-.

[59] Lindberg DA, Humphreys BF, McCray AT. The Unified Medical Language System. Methods of Information in Medicine. 1993;32:281-91.

[60] W. Ma RM, V. Ganesan, S. Nelson and S. Liu. RxNorm: Prescription for electronic drug information exchange. IT Professional. 2005;7:17-23.

[61] Bechhofer S, Van Harmelen F, Hendler J, Horrocks I, McGuinness DL, Patel-Schneider PF, et al. OWL web ontology language reference. W3C recommendation. 2004;10.

[62] Miles A, Bechhofer S. SKOS simple knowledge organization system reference. W3C recommendation. 2009.

[63] BioPortal SPARQL. http://sparql.bioontology.org, (accessed 28 Feb 2020).

[64] Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3:160035.

[65] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101:E215-20.

[66] Annas GJ. HIPAA regulations—a new era of medical-record privacy? New England Journal of Medicine. 2003;348:1486-90.

[67] SIRM. COVID-19 database. https://www.sirm.org/category/senza-categoria/covid-19/, (accessed 5 Jun 2020).

[68] Cunningham JA, Van Speybroeck M, Kalra D, Verbeeck R. Nine principles of semantic harmonization. AMIA Annual Symposium: American Medical Informatics Association; 2017. p. 451-9.

[69] Weng C, Fridsma DB. A call for collaborative semantic harmonization. AMIA Annual Symposium. 2006:1142-.

[70] Euzenat J, Shvaiko P. Ontology matching. 2 ed. Berlin, Heidelberg: Springer-Verlag 2007.

[71] Doan A, Madhavan J, Domingos P, Halevy A. Ontology matching: a machine learning approach. Handbook on Ontologies. Berlin, Heidelberg: Springer 2004. p. 385-403.

[72] Kalfoglou Y, Schorlemmer M. Ontology mapping: the state of the art. The Knowledge Engineering Review. 2003;18:1-31.

[73] Jérôme Euzenat, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Santos CTd. Ontology alignment evaluation initiative: six years of experience. Journal on Data Semantics. 2011; XV (6720):158-92.

[74] Lu Z, Michelle C, Adila K, Pascal H. A complex alignment benchmark: Geolink dataset. International Semantic Web Conference: Springer; 2018.

[75] Noy NF, Musen MA. PROMPT: algorithm and tool for automated ontology merging and alignment. Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence: AAAI Press; 2000. p. 450-5.

[76] Oliveira D, Pesquita C. Improving the interoperability of biomedical ontologies with compound alignments. Journal of Biomedical Semantics. 2018;9:1-13.

[77] Euzenat J. An API for ontology alignment. 3rd International Conference on Semantic Web. Hiroshima, Japan: Springer-Verlag; 2004. p. 698-712.

[78] Gouveia A, Silva N, Rocha J, Martins P. Debugging multi-property correspondences in ontology alignment scenarios. 2012.

[79] Vargas-Vera M, Nagy M. State of the art on ontology alignment. International Journal of Knowledge Society Research (IJKSR). 2015;6:17-42.

[80] Otero-Cerdeira L, Rodríguez-Martínez FJ, Gómez-Rodríguez A. Ontology matching: a literature review. Expert Systems with Applications. 2015;42:949-71.

[81] EDOAL. Expressive and declarative ontology alignment language. http://alignapi.gforge.inria.fr/edoal.html, (accessed 2 Apr 2019).

[82] Stoilos G, Geleta D, Shamdasani J, Khodadadi M. A novel approach and practical algorithms for ontology integration. International Semantic Web Conference 2018. p. 458-76.

[83] Sun P, Zhang S. Identifying granularity differences between large biomedical ontologies through rules. AMIA Annual Symposium. 2010;2010:927-31.

[84] Hayamizu TF, Mangan M Fau - Corradi JP, Corradi Jp Fau - Kadin JA, Kadin Ja Fau - Ringwald M, Ringwald M. The adult mouse anatomical dictionary: a tool for annotating and integrating data. Genome Biology. 2005;6:R29.

[85] Sun P, Zhang S. Using rules to investigate the differences in partonomy between biomedical ontologies. IEEE International Conference on Bioinformatics and Biomedicine. 2011:623-6.

[86] Luo L, Tong L, Zhou X, Mejino JLV, Ouyang C, Liu Y. Evaluating the granularity balance of hierarchical relationships within large biomedical terminologies towards quality improvement. Journal of Biomedical Informatics. 2017;75:129-37.

[87] He Z, Chen Y, de Coronado S, Piskorski K, Geller J. Topological-pattern-based recommendation of UMLS concepts for National Cancer Institute thesaurus. AMIA Annual Symposium. 2016;2016:618-27.

[88] He Z, Chen Y, Geller J. Perceiving the usefulness of the National Cancer Institute Metathesaurus for enriching NCIt with topological patterns. Studies in Health Technology and Informatics. 2017;245:863-7.

[89] He Z, Geller J. Preliminary analysis of difficulty of importing pattern-based concepts into the National Cancer Institute thesaurus. Studies in Health Technology and Informatics. 2016;228:389-93.

[90] Burgun A. Desiderata for domain reference ontologies in biomedicine. Journal of Biomedical Informatics. 2006;39:307-13.

[91] Freitas F, Schulz S, Moraes E. Survey of current terminologies and ontologies in biology and medicine. RECIIS-Electronic Journal in Communication, Information and Innovation in Health. 2009;3:7-18.

[92] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. Journal of the American Medical Informatics Association. 2006;13:277-88.

[93] Rosenbloom ST, Brown SH, Froehling D, Bauer BA, Wahner-Roedler DL, Gregg WM, et al. Using SNOMED CT to represent two interface terminologies. Journal of the American Medical Informatics Association. 2009;16:81-8.

[94] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;17:507-13.

[95] Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010;17:229-36.

[96] Soldaini L, Goharian N. Quickumls: a fast, unsupervised approach for medical concept extraction. MedIR workshop, sigir2016. p. 1-4.

[97] Jonquet C, Shah N, Youn C, Callendar C, Storey M-A, Musen M. NCBO annotator: semantic annotation of biomedical data. International Semantic Web Conference2009.

[98] Jonquet C, Shah NH, Musen MA. The open biomedical annotator. AMIA Summit on Translational Bioinformatics.2009:56-60.

[99] Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Research. 2019;47:W587-W93.

[100] Kim D, Lee J, So CH, Jeon H, Jeong M, Choi Y, et al. A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access. 2019;7:73729-40.

[101] Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2018;25:331-6.

[102] Nahler G. Anatomical therapeutic chemical classification system (ATC). Dictionary of Pharmaceutical Medicine: Springer; 2009. p. 8-.

[103] Cimino JJ. From data to knowledge through concept-oriented terminologies: experience with the medical entities dictionary. Journal of the American Medical Informatics Association. 2000;7:288-97.

[104] Rosse C, Mejino JL. The foundational model of anatomy ontology. Anatomy Ontologies for Bioinformatics: Springer; 2008. p. 59-117.

[105] Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics. 2008;83:610-5.

[106] The veterinary terminology services laboratory. The veterinary extension to SNOMED CT. http://vtsl.vetmed.vt.edu/extension/, (accessed 22 Feb 2018).

[107] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. AMIA Annual Symposium. 2001:57-61.

[108] Halper M, Morrey CP, Chen Y, Elhanan G, Hripcsak G, Perl Y. Auditing hierarchical cycles to locate other inconsistencies in the UMLS. AMIA Annual Symposium. 2011;2011:529-36.

[109] Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naïve vs. formal. AMIA Annual Symposium. 2005;2005:550-4.

[110] Howell DC. Statistical methods for psychology. Cengage Learning; 2009. p. 92-4.

[111] Wikipedia. Transitive relation. https://en.wikipedia.org/wiki/Transitive_relation, (accessed 5 Nov 2019).

[112] Fire ladder https://favpng.com/download/QBxztWD9, (accessed 15 Mar 2021).

[113] Wikipedia. Permutation. https://en.wikipedia.org/wiki/Permutation, (accessed 28 Oct 2019).

[114] Zhang G-Q, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annual Symposium. 2010;2010:922-6.

[115] Federhen S. The NCBI taxonomy database. Nucleic Acids Research. 2012;40:D136-D43.

[116] BioPortal. Gazetteer. https://bioportal.bioontology.org/ontologies/GAZ, (accessed 3 Mar 2018).

[117] Lee D, Cornet R, Lau F, de Keizer N. A survey of SNOMED CT implementations. Journal of Biomedical Informatics. 2013;46:87-96.

[118] Natale DA, Arighi CN, Blake JAA-OhooX, Bona J, Chen C, Chen SC, et al. Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. Nucleic Acids Research. 2017;45:D339-D46.

[119] BioPortal. Robert Hoehndorf Version of MeSH. https://bioportal.bioontology.org/ontologies/RH-MESH, (accessed 3 Mar 2018).

[120] Frazier P, Rossi-Mori A Fau - Dolin RH, Dolin Rh Fau - Alschuler L, Alschuler L Fau - Huff SM, Huff SM. The creation of an ontology of clinical document names. Studies in Health Technology and Informatics. 2012;84:94-8.

[121] Chelliah V, Juty N, Ajmera I, Ali R, Dumousseau M, Glont M, et al. BioModels: ten-year anniversary. Nucleic Acids Research. 2015;43:D542-D8.

[122] NCBO BioPortal. Ontology recommender. https://bioportal.bioontology.org/recommender, (accessed 3 Mar 2018).

[123] BioPortal. Current procedural terminology. https://bioportal.bioontology.org/ontologies/CPT, (accessed 3 Mar 2018).

[124] MedRA. Medical dictionary for regulatory activities. https://www.meddra.org/, (accessed 3 Mar 2018).

[125] First DataBank. NDDF (FDB MedKnowledge). http://www.fdbhealth.com/fdb-medknowledge/, (accessed 3 Mar 2018).

[126] Arp R, Smith B, Spear AD. Building ontologies with basic formal ontology: The MIT Press; 2015.

[127] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The ontology for biomedical investigations. PloS One. 2016;11:e0154556.

[128] Rosse C. MJLV. The foundational model of anatomy ontology. Anatomy Ontologies for Bioinformatics. London, UK: Springer; 2008. p. 59-60.

[129] Dialynas E, Topalis P, Vontas J, Louis C. MIRO and IRbase: IT tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors. PLoS Neglected Tropical Diseases. 2009;3:e465.

[130] Courtot M, Juty N, Knüpfer C, Waltemath D, Zhukova A, Dräger A, et al. Controlled vocabularies and semantics in systems biology. Molecular Systems Biology. 2011;7.

[131] Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Research. 2012;40:D940-D6.

[132] Schlesinger V. 5 Ways of preventing questionnaire fatigue. https://blog.smartsurvey.co.uk/5-ways-of-preventing-questionnaire-fatigue/, (accessed 3 Mar 2018).

[133] Battaglia MP, Khare, M., Frankel, M. R., Murray, M. C., Buckley, P. and Peritz, S. Response rates: how have they changed and where are they headed? Advances in Telephone Survey Methodology. Hoboken, NJ: John Wiley & Sons, Inc.; 2007.

[134] Moustakas C. Phenomenological Research Methods: SAGE Publications; 1994.

[135] Bowden C, Galindo-Gonzalez S. Interviewing when you're not face-to-face: the use of email interviews in a phenomenological study. International Journal of Doctoral Studies. 2015;10:79-92.

[136] Medway R. Beyond response rates: The effect of prepaid incentives on measurement error: University of Maryland, College Park; 2012.

[137] Baker M. Databases fight funding cuts: Online tools are becoming ever more important to biology, but financial support is unstable. Nature. 2012;489:19.

[138] Musen MA, the Protégé T. The Protégé project: a look back and a look forward. AI matters. 2015;1:4-12.

[139] National Institutes of Health. Data ontologies for biomedical research. https://grants.nih.gov/grants/guide/pa-files/PAR-07-425.html, (accessed 28 Jun 2018).

[140] Blumenthal D. Stimulating the adoption of health information technology. West Virginia Medical Journal. 2009;105:28-30.

[141] EHR incentive programs: 2015 through 2017 (modified Stage 2) overview. https://www.cdc.gov/ehrmeaningfuluse/docs/CMS_Stage_3_MU_Overview_2015_2017.pdf, (accessed 24 Aug 2020).

[142] Coronavirus resource center. https://coronavirus.jhu.edu/, (accessed 3 Mar 2020).

[143] COVID-19 ontology. http://bioportal.bioontology.org/ontologies/COVID-19, (accessed 30 Sep 2020).

[144] Babcock S, Cowell LG, Beverley J, Smith B. The infectious disease ontology in the age of COVID-19. 2020.

[145] Infectious Disease Ontology. https://bioportal.bioontology.org/ontologies/IDO, (accessed 30 Sep 2020).

[146] Virus infectious disease ontology. https://bioportal.bioontology.org/ontologies/VIDO, (accessed 30 Sep 2020).

[147] WHO COVID-19 rapid version CRF semantic data model. https://bioportal.bioontology.org/ontologies/COVIDCRFRAPID, (accessed 30 Sep 2020).

[148] Dutta B, DeBellis M. CODO: an ontology for collection and analysis of Covid-19 data. arXiv preprint arXiv:2009.01210. 2020.

[149] de Lusignan S, Bernal JL, Zambon M, Akinyemi O, Amirthalingam G, Andrews N, et al. Emergence of a novel coronavirus (COVID-19): protocol for extending surveillance used by the Royal College of general practitioners research and surveillance centre and public health England. JMIR public health and surveillance. 2020;6:e18606.

[150] ACT COVID ontology v3.0. https://github.com/shyamvis/ACT-COVID-Ontology/tree/master/ontology, (accessed 30 Sep 2020).

[151] WHO. International classification of diseases. http://www.who.int/classifications/icd/en/, (accessed 30 Sep 2020).

[152] National drug code database background information. https://www.fda.gov/drugs/development-approval-process-drugs/national-drug-code-database-background-information, (accessed 30 Sep 2020).

[153] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in Health Technology and Informatics. 2006;121:279-90.

[154] Structural Analysis of Biomedical Ontologies Center (SABOC). https://saboc.njit.edu/, (accessed 17 Mar 2021).

[155] UMLS. SNOMEDCT_US (US Edition of SNOMED CT) - Statistics. https://wayback.archive-it.org/4253/20190401044310/https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT_US/stats.html (accessed 5 Nov 2019).

[156] UMLS. NCI (NCI Thesaurus) - Statistics. https://wayback.archive-it.org/4253/20190401043652/https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NCI/stats.html, (accessed 5 Nov 2019).