Fall 2013

Using structural and semantic methodologies to enhance biomedical terminologies

Zhe He
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

USING STRUCTURAL AND SEMANTIC METHODOLOGIES TO ENHANCE BIOMEDICAL TERMINOLOGIES

by
Zhe He

Biomedical terminologies and ontologies underlie various Health Information Systems (HISs), Electronic Health Record (EHR) Systems, Health Information Exchanges (HIEs) and health administrative systems. Moreover, the proliferation of interdisciplinary research efforts in the biomedical field is fueling the need to overcome terminological barriers when integrating knowledge from different fields into a unified research project. Therefore well-developed and well-maintained terminologies are in high demand. Most of the biomedical terminologies are large and complex, which makes it impossible for human experts to manually detect and correct all errors and inconsistencies. Automated and semi-automated Quality Assurance methodologies that focus on areas that are more likely to contain errors and inconsistencies are therefore important.

In this dissertation, structural and semantic methodologies are used to enhance biomedical terminologies. The dissertation work is divided into three major parts. The first part consists of structural auditing techniques for the Semantic Network of the Unified Medical Language System (UMLS), which serves as a vocabulary knowledge base for biomedical research in various applications. Research techniques are presented on how to automatically identify and prevent erroneous semantic type assignments to concepts. The Web-based *adviseEditor* system is introduced to help UMLS editors to make correct multiple semantic type assignments to concepts. It is made available to the National Library of Medicine for future use in maintaining the UMLS.

The second part of this dissertation is on how to enhance the conceptual content of SNOMED CT by methods of semantic harmonization. By 2015, SNOMED will become the standard terminology for EHR encoding of diagnoses and problem lists. In order to enrich the semantics and coverage of SNOMED CT for clinical and research applications, the problem of semantic harmonization between SNOMED CT and six reference terminologies is approached by 1) comparing the *vertical* density of SNOMED CT with the reference terminologies to find potential concepts for export and import; and 2) categorizing the relationships between structurally congruent concepts from pairs of terminologies, with SNOMED CT being one terminology in the pair. Six kinds of configurations are observed, e.g., alternative classifications, and suggested synonyms. For each configuration, a corresponding solution is presented for enhancing one or both of the terminologies.

The third part applies Quality Assurance techniques based on "Abstraction Networks" to biomedical ontologies in BioPortal. The National Center for Biomedical Ontology provides BioPortal as a repository of over 350 biomedical ontologies covering a wide range of domains. It is extremely difficult to design a new Quality Assurance methodology for each ontology in BioPortal. Fortunately, groups of ontologies in BioPortal share common structural features. Thus, they can be grouped into families based on combinations of these features. A uniform Quality Assurance methodology design for each family will achieve improved efficiency, which is critical with the limited Quality Assurance resources available to most ontology curators. In this dissertation, a family-based framework covering 186 BioPortal ontologies and accompanying Quality Assurance methods based on abstraction networks are presented to tackle this problem.

USING STRUCTURAL AND SEMANTIC METHODOLOGIES TO ENHANCE BIOMEDICAL TERMINOLOGIES

by
Zhe He

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

January 2014

.

APPROVAL PAGE

USING STRUCTURAL AND SEMANTIC METHODOLOGIES TO ENHANCE BIOMEDICAL TERMINOLOGIES

Zhe He

Dr. James Geller, Dissertation Co-Advisor (Date)
Professor and Chair, Department of Computer Science, NJIT

Dr. Yehoshua Perl, Dissertation Co-Advisor (Date)
Professor, Department of Computer Science, NJIT

Dr. Mei Liu, Committee Member (Date)
Assistant Professor, Department of Computer Science, NJIT

Dr. Michael Halper, Committee Member (Date)
Professor and Program Director, Information Technology Program, NJIT

Dr. Gai Elhanan, Committee Member (Date)
Chief Medical Information Officer, Halfpenny Technologies, Inc.

Dr. Chunhua Weng, Committee Member (Date)
Assistant Professor, Department of Biomedical Informatics, Columbia University

BIOGRAPHICAL SKETCH

Author: Zhe He

Degree: Doctor of Philosophy

Date: January 2014

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2014

- Master of Science in Computer Science,
  Columbia University in the City of New York, New York, NY, 2009

- Bachelor of Engineering in Computer Science and Technology,
  Beijing University of Posts and Telecommunications, Beijing, P. R. China, 2007

Major: Computer Science

Publications:

Zhe He, Yehoshua Perl, Gai Elhanan and James Geller. "Auditing the Extents of Top Semantic Types of the UMLS Semantic Network," to be submitted for journal publication.

Zhe He, Charles Paul Morrey, Yehoshua Perl and James Geller. "Sculpting the UMLS Refined Semantic Network," to be submitted for journal publication.

Zhe He, Christopher Ochs, James Geller and Yehoshua Perl. "A Structural Meta-Ontology for Family-Based Quality Assurance Framework for Biomedical Ontologies in BioPortal," to be submitted for journal publication.

Siviram Arabandi, Christopher Ochs, Zhe He, Yehoshua Perl and James Geller. "Quality Assurance for the Sleep Domain Ontology Using Abstraction Networks," to be submitted for journal publication.

James Geller, Zhe He and Gai Elhanan. "Categorizing the Relationship between Structurally Congruent Concepts from Pairs of Terminologies," submitted for publication.

Zhe He, Christopher Ochs, Ankur Agrawal, Yehoshua Perl, Dimitris Zeginis, Konstantinos Tarabanis, Gai Elhanan, Michael Halper, Natasha Noy and James Geller. "A Family-based Framework for Supporting Quality Assurance of Biomedical Ontologies in BioPortal," Proceedings of AMIA 2013 Annual Symposium, November 16-20, Washington, D.C: 581-90.

Zhe He, Christopher Ochs, Larisa Soldatova, Yehoshua Perl, Sivaram Arabandi and James Geller. "Auditing Excess Reuse of a Top Level Ontology for the Drug Discovery Investigation Ontology," Proceedings of 2013 International Workshop on Vaccine and Drug Ontology Studies. July 7, 2013, Montreal, QC, Canada.

James Geller, Christopher Ochs, Zhe He and Yehoshua Perl. "A Structural Meta-ontology for the BioPortal Ontologies," Proceedings of Bio-ontologies 2013. July 20, 2013, Berlin, Germany.

Christopher Ochs, Zhe He, Yehoshua Perl, Sivaram Arabanadi and James Geller. "Refining the Granularity of Abstraction Networks for the Sleep Domain Ontology," Proceedings of the 4th International Conference on Biomedical Ontology. July 8-9, 2013, Montreal, QC, Canada: 84-9.

Ankur Agrawal, Zhe He, Duo Wei, Michael Halper, Yehoshua Perl and Gai Elhanan. "The Readiness of SNOMED Problem List Concepts for Meaningful Use of EHRs," Artificial Intelligence in Medicine, 2013. (58)2: 73-80.

James Geller, Zhe He, Yehoshua Perl, C. Paul Morrey and Julia Xu. "Rule-Based Support System for Multiple UMLS Semantic Type Assignments," Journal of Biomedical Informatics. 2013. (46)1: 97-110.

Zhe He, Michael Halper, Yehoshua Perl, and Gai Elhanan. "Clinical Clarity Versus Terminological Order – The Readiness of SNOMED CT Concept Descriptors for Primary Care," Proceedings of the Second International Workshop on Managing Interoperability and Complexity in Health Systems (MIX-HS`12) in conjunction with the 21st International Conference on Information and Knowledge Management (CIKM`12), Maui, HI. October 2012: 1-6.

Huanying Gu, Gai Elhanan, Michael Halper and Zhe He. "Questionable Relationship Triples in the UMLS," Proceedings of the 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI`12), Hong Kong, China, January 2-7, 2012: 713-716.

Presentations

Zhe He, "A Family-Based Framework for Supporting Quality Assurance of Biomedical Ontologies in BioPortal," AMIA 2013 Annual Symposium, Washington D.C. November 17, 2013

Zhe He, "Auditing the Unified Medical Language System based on Abstraction Network," Guest Lecture of Medical Informatics, Richard Stockton College of New Jersey, Galloway, NJ. November 7, 2013.

Zhe He, "Auditing OWL-Based Biomedical Ontologies in BioPortal Using Structural Methodologies," Guest Lecture of Methods in Biomedical Informatics, Columbia University, New York, NY. October 23, 2013.

Zhe He, "Auditing Excess Reuse of a Top Level Ontology for the Drug Discovery Investigation Ontology," the 4th International Conference on Biomedical Ontology (ICBO`12), Montreal, QC, Canada. July 7, 2013.

Zhe He, "Clinical Clarity versus terminological order – the readiness of SNOMED CT concept descriptors for primary care," the 21st International Conference on Information and Knowledge Management (CIKM`12), Maui, HI. October 29, 2012.

Zhe He, "Questionable relationship triples in the UMLS," the 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI`12), Shenzhen, China, January 6, 2012.

To my beloved parents and friends

ACKNOWLEDGMENTS

Foremost, I would like to express my deepest appreciation to my dissertation co-advisors, Dr. James Geller and Dr. Yehoshua Perl for their invaluable technical guidance, kindness, encouragement, and tremendous support in my PhD studies. Part of the research work in this dissertation conducted in the summer of 2009, 2010 and 2011 was supported by the United States National Library of Medicine under grants R-01-LM008445-01A2 and R-01-LM008912-01A1. Their mentorship has not only made me a computer scientist but also an independent thinker.

Secondly, I am extremely grateful to Dr. Mei Liu, Dr. Michael Halper, Dr. Gai Elhanan and Dr. Chunhua Weng for serving on my dissertation committee and providing their insightful input to my research.

During my PhD studies, I am honored to have worked with many kind and smart fellow students. I would like to thank Christopher Ochs, and graduated students: Dr. Ankur Agrawal, Dr. Duo Wei, Dr. Paul Morrey, Dr. Helen Gu, and Dr. Chen Yan, without whom I wouldn't have achieved this far.

Having spent the past five years in the Department of Computer Science, I feel like I am part of a family. I owe a debt of gratitude to all the staff members of this department. I would like to specially thank Dr. David Nassimi, Dr. George Olsen, Ms. Casey Hennessey, and Ms. Angel Bell for providing me enormous help.

At the later stage of my PhD life, I am honored to be the President of NJIT Graduate Student Association. Being the first Chinese President of GSA since it was founded in 1984, I am extremely thankful and have dedicated most of spare time. I would

like to thank Dr. Sotirios Ziavras and Ms. Clarisa Lenahan-Gonzalez for their limitless advice and support to GSA in my term.

PhD study is a tough journey. But I'm lucky that I am not alone in this journey. My life wouldn't have been so fantastic without their company. Thank you for always being there, my friends: Huan Li, Weiping Huang, Yan Ji, Zeyu Li, Yuyang Zhang, Tian Tian, Peng Liu, Siyang Wen, Beibei Wu, Xiupeng Wang, Yixuan Li, Ying Xu, Yulin Huang, Lei Yang, Dan Liu, Haixu Wang, Qing Li, Jiawei Chen, Zhenqing Zheng, Li Zheng, Dewen Li, Xiang Ji, Cheng Niu, Weiwei Guo, Hao Dang and many more. I believe friendship lasts forever.

My girlfriend, Yue Li, is always supportive and considerate. Thanks for being with me finishing this journey.

I don't know how to express my gratitude to my parents, Hongbin He and Qiuhong Lu, who have brought me to this world, taught me how to walk, to speak and most importantly, to grow to be independent and strong. I believe my success and happiness are the best return for their love.

As Albert Einstein said, "There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle." I firmly believe that anything can be achieved with whole-hearted effort.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Objective

With an increasing role of health information systems in health care, more attention has been paid to improving computerized medical records and interoperable heterogeneous medical systems. The HITECH (Health Information Technology for Economic and Clinical Health) Act defined "Meaningful Use" of interoperable EHR (Electronic Health Record) adoption in the health care system as a critical national goal [1]. Under the HITECH Act, $25.9 billion are now being spent by the United States Department of Health and Human Services to promote the adoption of health information technology [2]. Standard biomedical terminologies are a foundation of EHR systems (e.g., eClinicalWorks [3], Allscripts [4], Epic [5], etc.), clinical decision support systems (CDSSs) (e.g., DXplain [6], DiagnosisPro [7], VisualDX [8], etc.), Health Information Exchanges (HIEs) (e.g., Harvard Pilgrim Health Care [9], Delaware HIN [10], Indiana HIE [11], etc.), healthcare billing systems (e.g. CareCloud [12], ADS [13], NueMD [14], etc.), and biomedical research. Use of standard biomedical terminologies in the collection, storage and reporting of medical information helps to ensure a consistent interpretation of data of different systems and data repositories by verifying their semantics.

Biomedical terminologies play an important role in today's clinical practice, biomedical research, and various healthcare applications [15]. However, the explosion of such resources over the last several decades was not accompanied by a successful drive toward standardization of their structure and content. Therefore, software systems that take

advantage of terminologies to achieve interoperability with various other systems may very well encounter difficulties [16].

The purpose of this dissertation research is to improve biomedical terminological systems to support biomedical research and EHR systems in health care. This purpose is approached from two directions: 1) Most standard terminologies are large and complex. Errors may occur when updating or adding terms to standard biomedical terminologies. Imprecise representations of patients' medical conditions and symptoms may cause problems [17]. Correcting errors in standard terminologies by automated or semi-automated auditing methods is likely to have a positive impact on various health information systems that use standard terminologies. 2) Structural methods based on hierarchical relationships will be presented to approach the semantic harmonization problem in order to enrich the semantic content of terminologies and facilitate the semantic interoperability between terminologies.

1.2 Biomedical Terminological Systems

1.2.1 The Unified Medical Language System and the Refined Semantic Network

1.2.1.1 Structure of the Unified Medical Language System. The Unified Medical Language System (UMLS) [18, 19], is derived from 173 source terminologies. In the 2013AA release of the Metathesaurus (META) [20, 21], there are over 9 million terms mapping to 2.9 million concepts. The UMLS Semantic Network [22-24] provides a compact abstraction network, consisting of 133 high-level, broad categories, called semantics types. One or more of the semantic types of the Semantic Network are assigned to each of the META concepts, describing the semantics of the concept by identifying its

broad category or categories. For example, the semantics of *Dental Fistula*[1] is described by its assigned semantic type **Anatomical Abnormality**[2]. The *extent* of a semantic type of the Semantic Network is the set of META's concepts that are assigned this semantic type. The Semantic Network supports the ongoing integration of new and revised source terminologies into the UMLS.

1.2.1.2 The Refined Semantic Network. In previous research at the Structural Analysis of Biomedical Ontologies Center (SABOC) at NJIT it was determined that the UMLS Semantic Network has many shortcomings [25, 26]. For example, it has a strict tree structure. To address this problem, the Refined Semantic Network (RSN) [27] was developed. It has two kinds of refined semantic types (RSTs) derived from the original semantic types of the Semantic Network and their assignments to the concepts of META.

One kind of RST, the Pure Semantic Type (PST) is assigned to concepts for which the corresponding semantic type is the only semantic type assigned. This kind of semantics is exclusive. Exclusive semantics refer to the fact that the concepts assigned this semantic type are not assigned any other semantic type. The semantics of a pure semantic type is the exclusive semantics of the corresponding original semantic type.

The other kind of refined semantic type is called Intersection Semantic Type (IST), which represents an existing combination of multiple semantic types. An IST is only created in the Refined Semantic Network, if there is at least one concept in the META that has exactly these semantic types assigned. Its compound semantics [27] is defined as the conjunction (AND) of the semantics of the various semantic types. For example, an IST

[1] Concepts are denoted by italics.
[2] Semantic types are denoted by bold typeset.

has to be assigned to the concepts which are assigned the two semantic types **Diseases or Syndrome** and **Anatomical Abnormality**. As shown in Figure 1.1, this IST is denoted by **Disease or Syndrome** ∩ **Anatomical Abnormality**, where ∩ is the mathematical intersection symbol. For example, the concept *Fistula of lip* is assigned this IST. An IST has semantic uniformity since all concepts of its extent share the same compound semantics.

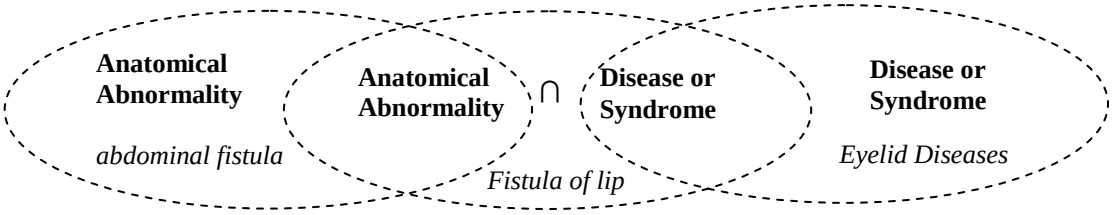

Figure 1.1 Example of a concept assigned two semantic types.

1.2.1.3 Relationships in the META. The META of the UMLS contains many relationships between different concepts. Most of the relationships come from its 168 source vocabularies while some of them were added by the National Library of Medicine (NLM) during the META construction. In the 2013AA release of the UMLS, there are 56,532,106 META relationships. There are two types of relationships: intra-source vocabulary relationships and inter-source relationships [28]. Most of intra-source relationship are asserted or implied by the source vocabularies and some of them are computed by the frequency with which concepts in specific vocabularies co-occur in records in a database. Most of inter-source relationships in the META are synonym relationships and their existence mainly contributes to the functionality of mapping of source vocabularies.

All META relationships carry a general label (REL) to describe their basic nature. There are 11 RELs, which are Child of (CHD), Parent of (PAR), Broader (RB), Narrower

(RN), Qualified by (QB), Sibling of (SIB), Synonym of (SY), Alike (RL), Related and possibly synonymous (RQ), Related unspecified (RU), and Other (RO). About 41.3% of the relationships in the META also carry a relationship attribute (RELA), which comes from a source vocabulary. The RELA describes the relationship nature in a more specific way. There are 643 different RELAs in the META. For example, RO is the REL from *Aluminum Hydroxide Gel, compounding powder* to *Aluminum Hydroxide*. The RELA of this REL is *has_ingredient*. In the UMLS, 'PAR' represents an explicit parent-child relationship in a source, and 'RB' indicates an implied one (as interpreted by the UMLS editorial team).

1.2.1.4 Problems with Multiple Semantic Type Assignments. When there are two semantic types assigned to the same concept, a number of problems may occur. In some cases, one semantic type assignment may be redundant, because the other semantic type expresses the meaning of the concept in a more specific way. As illustrated in Figure 1.2, an assignment to a concept *X* of a semantic type **A** is redundant if *X* is also assigned another semantic type **B**, such that **B** IS-A **A**. In other cases, one semantic type assignment may outright contradict another one, indicating an inconsistency in the UMLS semantic type assignments.

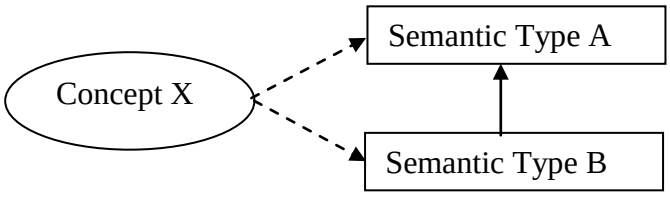

Figure 1.2 Configuration of a redundant semantic type assignment.

For example, in the documentation of the Semantic Network, the usage note of **Finding** prevents it from being double-typed with either **Pathologic Function** or **Anatomical Abnormality.** These problems notwithstanding, multiple assignments are important to express fine shades of semantics. For some cases, e.g. for chemical concepts, multiple assignments are explicitly encouraged in the documentation of the UMLS Semantic Network. There is no public repository that expresses all the different legitimate ways of interplay between the 133 semantic types. Neither is there a complete list of prohibited combinations of semantic types.

Concepts assigned multiple semantic types are complex, due to their compound semantics of being simultaneously "this and that." It was shown that concepts with rare combinations of semantic types [26, 27, 29-31], i.e., there are only a few Metathesaurus concepts assigned exactly this combination, have a high likelihood of erroneous semantic type assignments. In other words, ISTs with small extents are more likely to have wrong semantic type assignments [25]. Furthermore, some semantic type assignments stand in contradiction to the explicit documentation of the Semantic Network. This situation suggests that UMLS editors would benefit from a rule-based support system, informing them regarding the permissibility of assigning a specific combination of semantic types to a concept.

1.2.2 Systematized Nomenclature of Medicine – Clinical Terms

1.2.2.1 Structure of Systematized Nomenclature of Medicine – Clinical Terms.

The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) was formed through the merger of SNOMED RT (Reference Terminology) and CTV3 (Clinical Terms Version 3) [32]. It is a description-logic-based [33] medical terminology, which

covers a wide scope of clinical concepts, including diseases, procedures, specimens, findings, substances, etc. It has about 298,000 active concepts organized in 19 top-level, singly-rooted hierarchies, which are 1) Clinical finding, 2) Procedure, 3) Observable entity, 4) Body structure, 5) Organism, 6) Substance, 7) Pharmacologic / biologic product, 8) Specimen, 9) Special concept, 10) Linkage concept, 11) Physical force, 12) Event, 13) Environment or geographical location, 14) Social context, 15) Situation with explicit context, 16) Staging and scales, 17) Physical object, 18) Qualifier value, and 19) Record artifact.

Each SNOMED CT concept has three types of descriptors, Fully Specified Name (FSN), Preferred Term (PT) and Synonyms. The FSN uniquely describes a concept and clarifies its meaning. Each FSN term ends with a "semantic tag" in parentheses, which indicates the semantic category to which the concept belongs. For example, *chronic obstructive lung disease (disorder)* is the FSN of the concept that represents the clinical diagnosis that a clinician makes when a person has a "chronic obstructive lung disease." The semantic tag is "disorder." Many semantic tags, but not all, are identical to the roots of the hierarchies the terms are in. The Preferred Term of a concept is a commonly used word or phrase used by clinicians for this concept. A synonym represents a term other than the FSN or the Preferred Term that can also be used to represent the concept.

Concepts in SNOMED CT may be assigned relationships. Each SNOMED CT concept, except for the 19 root concepts, has at least one IS-A relationship to a supertype concept. A second kind of relationship is called *attribute relationship*. An attribute relationship is an association between two concepts that specifies a characteristic of the

source of the relationship. For example, the concept *myocardial infarction* has an attribute relationship *finding site* to the concept *myocardium structure*.

1.2.2.2 Significance of SNOMED CT. SNOMED CT [34-37] is considered to be of increasing importance in the Medical Informatics community. One reason for this importance is related to government mandates of using Electronic Medical Record systems, meaningful use and incentive payments to physicians. By 2015, SNOMED CT will become the standard terminology for EHR encoding of diagnoses and problem lists [1]. SNOMED CT is to be used to "enable a user to electronically record, modify, and retrieve a patient's problem list for longitudinal care (i.e., over multiple office visits)." To accelerate the adoption and meaningful use of EHRs by providers, incentives and penalties were defined [1, 38].

1.2.2.3 Problems with Conceptual Content of SNOMED CT. In a recent survey [39], missing concepts and missing synonyms were reported as the top two deficiencies in SNOMED CT mentioned by 23% and 17% of responders, respectively. More than half of the SNOMED CT users responding indicated that expanding synonym coverage is important to them.

Making conceptual content adequacy more critical is the fact that the HITECH regulations [40, 41] and the "meaningful use" initiative portend nearly exponential growth in the adoption of Electronic Health Record (EHR) systems in the near future [1, 41]. As SNOMED CT is slated to become the exclusive encoding system for problem lists by 2015 [1], a much wider range of users is expected to interact with SNOMED CT-based content in clinical applications. Such users will expect correct and appropriate synonyms to allow for ease of differentiation between similarly worded concepts in order to efficiently select

the clinical concepts that best apply to their patients. Errors in synonyms, lack of synonyms, or insufficient concept information to decipher the exact meaning of concepts' descriptors may prove detrimental to widespread clinical adoption.

1.2.3 Biomedical Ontologies in BioPortal

BioPortal is a repository and uniform development and visualization system for biomedical ontologies provided by the National Center for Biomedical Ontology (NCBO) [42]. BioPortal contains over 330 biomedical ontologies developed in the Web Ontology Language (OWL) [43], Resource Description Framework (RDF) [44], Open Biological and Biomedical Ontologies (OBO) [45] format, Protégé frames, and Rich Release Format of the UMLS. It also provides tools for browsing, developing, editing, and visualizing ontologies to support research in the biomedical sciences.

1.3 Structure of the Dissertation

The objective of this dissertation is to improve the quality of biomedical terminological systems using abstraction networks and other structural methodologies. The remainder of this dissertation is organized as follows.

Chapter 2 provides background information about abstraction networks and Quality Assurance of biomedical terminological systems.

Chapter 3 presents a longitudinal study of the process of improving the UMLS as a result of auditing its semantic type assignments. The chapter first examines previously collected data and then segues into a study of the UMLS evolution between 2010 and 2013.

Chapter 4 describes a rule-based algorithm, for helping a human editor with overcoming the problems caused by inconsistent multiple semantic type assignments.

Chapter 5 presents the Web-based software tool AdviseEditor that implements the algorithm introduced in Chapter 4.

Chapter 6 shows a study in a simulated clinical scenario to assess whether SNOMED CT's concept descriptors provide sufficient differentiation to enable possible concept selection between similar terms.

Chapter 7 presents a study that categorizes the relationships between structurally congruent concepts from two terminologies, one of which is assumed to be SNOMED CT.

Chapter 8 extends the approach of Chapter 7 from configurations with one intermediate concept in each terminology to configurations with n ($n > 1$) intermediate concepts in one or both of the two terminologies.

Chapter 9 presents a family-based framework for supporting Quality Assurance (QA) of biomedical ontologies in BioPortal. This new paradigm will achieve high efficiency of ontology QA, which is critical due to the limited availability of QA resources.

Major parts of this dissertation work have been published in the Journal of Biomedical Informatics (Chapter 4 and Chapter 5) [46], the International Workshop on Managing Interoperability and Complexity in Health Systems (Chapter 6) [47], and the American Medical Informatics Association 2013 Annual Symposium (Chapter 9) [48]. Preliminary work for Chapter 3 has been published [49], and new results for the 2013AA release of the UMLS with suggestions to UMLS editors about possible corrections will be submitted for peer review. The work of Chapters 7 has been submitted for peer review [50]. The work of Chapter 8 is under preparation and will be submitted for peer review [51].

CHAPTER 2

BACKGROUND

2.1 Abstraction Networks

2.1.1 General Characteristics of Abstraction Networks

Most ontologies are complex, heavily connected, and lack a natural linear order. Thus, diagrammatic representations of ontologies have long been preferred over textual representations. Such representations typically take the shape of "node/box and link/arrow" pictures. However, when ontologies become too large, the advantages of diagrammatic representations disappear, and neither they nor text representations can easily support orientation and QA efforts. Thus, an alternative compact network called *abstraction network*, which summarizes the structure and content of an ontology, is utilized to make such an ontology more comprehensible.

Figure 2.1 demonstrates the general process of deriving an abstraction network from a small ontology (of 25 classes, shown as small ovals on the left side). As can be seen on the left, six groups (large ovals) are identified and each is subsequently mapped to and represented by one node (blue rectangle) on the right side. The exact nature of the mapping of subsets of the ontology's classes to the abstraction network's nodes is defined as part of the derivation methodology for a specific type of abstraction network. By its nature, an abstraction network provides a high-level compact view of the original ontology and can serve as a good entry point for the exploration of its structure and content.

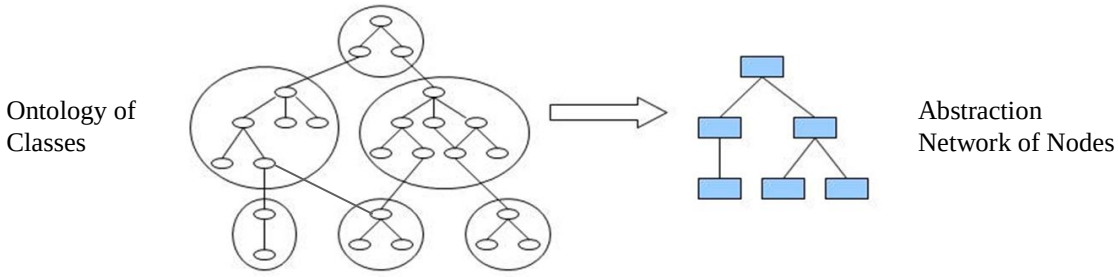

Figure 2.1 General process of deriving an abstraction network from an ontology.

2.1.2 Auditing the UMLS Using the Refined Semantic Network

In [27], an alternative abstraction network for the UMLS, the Refined Semantic Network (RSN) was introduced. This introduction was motivated by two deficiencies of the Semantic Network, one implying the other. For the Semantic Network, the extents of the types are not disjoint. For example, there are 989 concepts which are assigned both **Disease or Symptom** and **Anatomical Abnormality**. An abstraction network supports orientation into a repository of concepts by categorizing the concepts into broad categories. However an abstraction network is less effective in providing such orientation if the extents of its categories are not disjoint, since it does not provide knowledge on the proportion of the overlaps of the extents of various categories. The implied deficiency of the Semantic Network is that the extent of a semantic type does not necessarily exhibit semantic uniformity. For example, as shown in Figure 1.1, the concept *abdominal fistula* is just categorized as **Anatomical Abnormality** while the concept *Fistula of lip* is assigned both **Anatomical Abnormality** and **Disease or Syndrome**. Hence, the extent of the semantic type **Anatomical Abnormality** is not semantically uniform, since some of its concepts are categorized only as **Anatomical Abnormality**, while others are categorized by two different categories namely **Anatomical Abnormality** and **Disease or Syndrome**. An

abstraction network is more effective in its support for orientation if each category represents a semantically uniform set of concepts. The extent of any refined semantic type is semantically uniform and the extents of all refined semantic types of the RSN are disjoint. Thus, the RSN is an abstraction network which provides better orientation into the content of META. For example, the RSN of the 2013AA release shows that there are 2543 concepts which are anatomical abnormalities, 90691 concepts which describe diseases or syndromes and 989 concepts which are both anatomical abnormalities and diseases or syndromes. The Semantic Network does not provide this kind of sharp distinction.

The utility of the RSN for auditing the UMLS was manifested in enabling several auditing methodologies. In [25, 26, 52, 53] the utility of ISTs of small extents to expose erroneous semantic type assignments was demonstrated. Group auditing techniques for large extents of refined semantic types were described in [31, 54]. Improved modeling for conjugate and complex chemicals is explored in [55].

Gu et al. conjectured that many of the ISTs of small extents are erroneous and should not exist in the RSN [27]. For example, a review of 100 out of 422 ISTs assigned only a single concept found 89 erroneous assignments. Furthermore, 77 of the 1163 ISTs represented cases of redundant semantic type assignments. An assignment of a semantic type **A** to a concept is redundant if it is also assigned another semantic type **B**, such that **B** IS-A **A**. Redundant assignments are forbidden in the UMLS [22].

The plan at the time of the creation of the RSN was that by an effort of removing all redundant semantic type assignments and other erroneous combinations of semantic types from the UMLS only ISTs which stand for legitimate combinations of semantic types would remain, making the RSN considerably smaller.

Definition: An IST is considered *illegitimate* if its combination of semantic types satisfies any of the following:

(1) The combination of semantic type assignments to a concept is forbidden by the definitions or usage notes of the semantic types of the Semantic Network. For example, the combination of **Anatomical Abnormality** and **Neoplastic Process** is forbidden.

(2) The combination of semantic type assignments to a concept implies a redundant semantic type assignment. For example, if a concept is assigned both **Finding** and **Sign or Symptom**, the assignment of **Finding** is redundant, since **Sign or Symptom** is a child of **Finding** in the Semantic Network.

(3) The semantic types of the IST are mutually exclusive, e.g. for sibling semantic types in the subhierarchy of **Organism**.

(4) The semantic types of the IST do not refer to the same concept but to two concepts with different real world semantics.

Examples for the last category are the concept *Video Recording* and its child *Videotape recording*, which (in the 2008AB release of the UMLS) were assigned both **Manufactured Object** and **Human-caused Phenomenon or Process** [29]. This is a semantically impossible combination since an object cannot be a process. In the analysis of Geller et al. [29] it was realized that the **Manufactured Object** semantics referred to the product of the recording while the **Human-caused Phenomenon or Process** semantics referred to the recording process involved in producing this product. Indeed, in the current version of the UMLS, both these concepts are assigned only **Manufactured Object**,

Similar to the 2008 categorization of *Video Recording*'s other two children *Videodisk recording* and *Videotape/Videodisc*.

Definition: An IST is considered *legitimate* if it is not illegitimate.

The legitimate ISTs deserve to be elevated to be first class citizens in the RSN. The assumption was that not too many legitimate ISTs will remain in the RSN after all the illegitimate ISTs will have been removed. The legitimate ISTs occur mostly for chemical concepts where both a structurally viewed chemical semantic type and at least one functionality viewed chemical semantic type are expected, according to the definition of the **Chemical** semantic type [56].

As mentioned before, there were 77 ISTs in the 1998 UMLS release where one of the semantic types was an ancestor of the other, violating the rule of the UMLS forbidding redundant assignments of semantic types [22]. An algorithm for the detection of all concepts with redundant semantic type assignments was designed by the SABOC Center in 2002 [57]. In 1998 there were 8622 such concepts reported to the NLM, the curator organization of the UMLS. From that time, the UMLS has been periodically monitored by SABOC members for redundant semantic type assignments, and the findings were systematically reported to the NLM. Apparently, influenced by the published algorithm [57] and repeated reports to NLM staff, the NLM eventually implemented an automatic procedure that removes redundant semantic type assignments before each release of the UMLS [58].

2.1.3 Auditing Biomedical Terminologies Using Area and Partial Area Taxonomies

The area taxonomy and the partial area taxonomy are abstraction networks developed for auditing description-logic-based terminologies, e.g., the National Cancer Institute

Thesaurus (NCIt) [59] and SNOMED CT [60]. An *area* is defined as the set of all classes that are explicitly defined or inferred as being in exactly the domains of a given set of relationships. It is a collection of all concepts with the exact same structure in terms of relationships. The list of names of the relationships is used to name the area. A *root* of an area is defined as a class that has no parents in the same area. An area may have more than one root. A root of an area defines a *partial area*: a set of classes that includes the root and all its descendants in the area. *Partial areas* are connected by *child-of* links derived from the underlying IS-A relationships.

In this dissertation, the terms "concept" and "class" are used interchangeably. Figure 2.2 shows a terminology fragment with four concepts, *A* through *D*, that are introducing new relationships, and some other unlabeled concepts that do not introduce any new relationships. The bold arrows represent the IS-A relationships between pairs of concepts. The thin arrows represent lateral relationships between pairs of concepts. For example, the arrow from *A* to *D*, labeled r_2, means that *A* has a relationship r_2 to *D*. The children and grandchildren of *A* all exhibit the relationship r_2 due to inheritance. Thus, all these concepts are grouped into an area called *A*. Specifically, the partial area *A* is *child-of* the partial area *B* if a parent of *A*'s root resides in *B*.

Figure 2.3 shows a multi-rooted area with roots *A* and *E*. Even though *A* and *E* introduce the same kind of relationship r_2, they each represent a unique semantics given that the targets of the relationships reside in different areas. Thus, *A* and its descendants can be seen as a unique semantic grouping. This same is true for *E* and its descendants in this area. Such a grouping is defined as a *partial area*. This multi-rooted area is named after its relationship. Each partial area is named after its root. Note that while the partial areas form

a semantic division of an area, they do not constitute a partition of the area, in other words they are not necessarily disjoint.

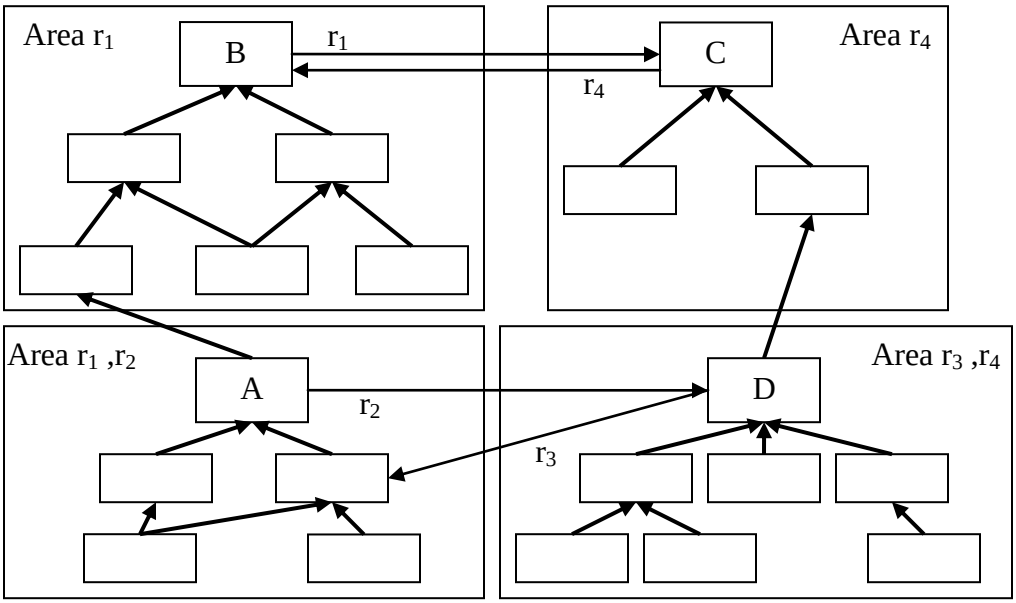

Figure 2.2 Four concepts introducing relationships and associated areas.

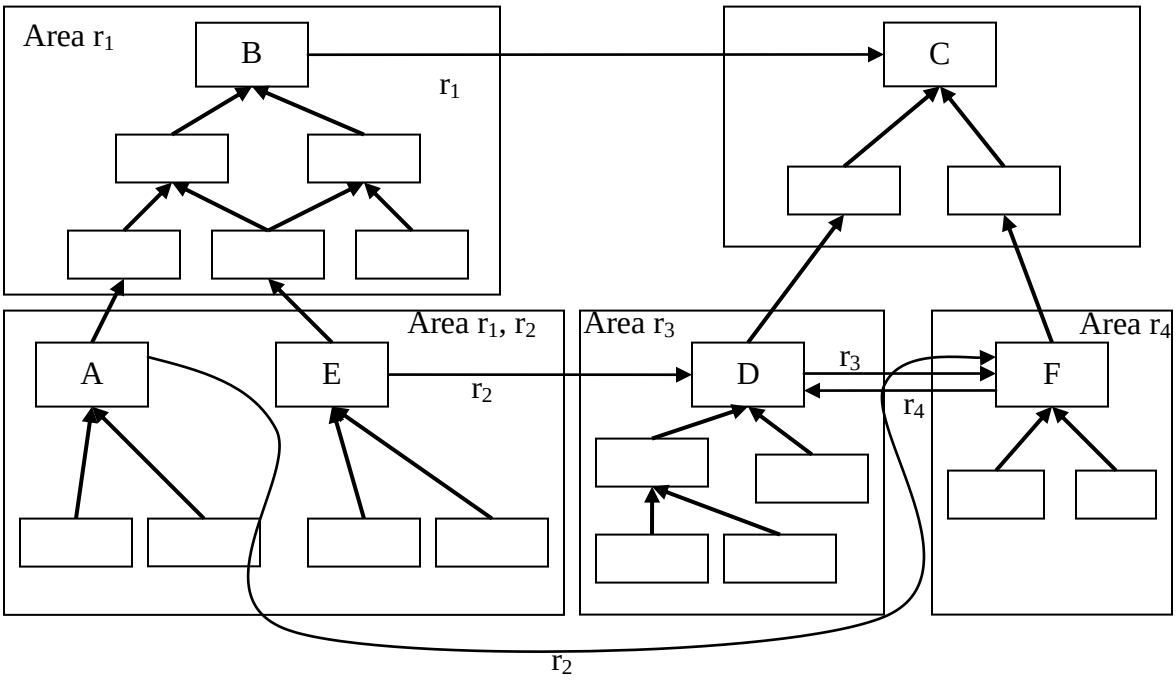

Figure 2.3 A multi-rooted area (with roots A and E).

Previously, area and partial area taxonomies were derived for SNOMED CT [60], NCIt [59], Ontology of Clinical Research (OCRe) [61], and Sleep Domain Ontology (SDO) [62]. These abstraction networks were shown to support semi-automated qualitiy assurance of the underlying ontologies by algorithmically identifying sets of classes (or concepts) that are more likely to be erroneous than the general class population. In particular, an abstraction network supported the exposure of errors and inconsistencies missed by a Description Logic classifier [63]. Examples of such sets of concepts in SNOMED CT include small partial areas, and sets of overlapping concepts (concepts belonging to two partial areas in the same area) [64] corresponding to nodes in a specific kind of abstraction network called the *disjoint partial area taxonomy* [65].

2.2 Relevant Work on Semantic Harmonization and Granularity

Semantic interoperability is one of the big challenges in biomedical informatics. In order to enrich the semantics and coverage of a terminology and facilitate translational biomedical informatics to be utilized in clinical and research applications, semantic harmonization efforts have recently been extended for various terminologies, e.g. SNOMED CT [66]. However, structural methodologies for semantic harmonization of terminologies have not been studied sufficiently. Weng et al. [67] presented a conceptual design of a collaborative system for semantic harmonization. Three key design principles were defined: (1) reuse, (2) collaboration, (3) harmonization as modeling. In [68], the BRIDG model was presented as a user-centric semantic harmonization framework. The harmonization in the BRIDG model is based on the concepts and their definitions. Tao et al. have discussed the importance of ontology harmonization before using ontologies to annotate clinical data

[69]. Bodenreider performed a study of redundant relations and similarity across families of terminologies and discussed the relationship between redundancy and semantic consistency [70].

Previously, granularities of medical terminologies have been analyzed based on hierarchical relationships to facilitate terminology integration and semantic harmonization. Many of these methodologies involved comparing the granularities of terminologies. Kumar et al. [71] lay out a comprehensive theory of granularity in the context of medical terminologies, based on prior work of Bittner and Smith [72] in the area of Geographic Information Science. They identify shortcomings of SNOMED CT with respect to granularity but do not quantify granularity differences.

Sun and Zhang [73] compare the granularities of the Adult Mouse Anatomical Dictionary [74] and the anatomy subset of the NCI Thesaurus [75]. They provide numerical results for three types of subclass configurations. Sun and Zhang accept the fact that their two terminologies are from different domains (mouse anatomy versus human anatomy). Thus, they consider differences such as "mice have thirteen ribs, humans only twelve" as legitimate. In this dissertation, this would be interpreted as a domain difference, and is therefore irrelevant to the study, which assumes terminologies (or overlapping sub-terminologies) in the same narrowly defined domain.

Schulz et al. identify granularity-related problems with "cross-granularity integration" in the biomedical domain [76]. Rector et al.'s analysis provides logical formulations of important distinctions, but does not contain an attempt to quantify granularity differences, as their notion of granularity is domain-oriented, while the differences in this dissertation are concept-oriented [77].

2.3 Relevant Work on Biomedical Ontologies in BioPortal

The NCBO BioPortal has been used in various research projects on biomedical ontologies. Mortensen et al. [78] encoded the Ontology Design Pattern (ODP) from several BioPortal ontologies to facilitate ontology development. Bail et al. [79] examined the justifications from an independently motivated corpus of actively used BioPortal ontologies and exhibited the structural features represented in description logic (DL). In [80], Quesada-Martínez et al. used all the ontologies available in BioPortal as external resources and examined their labels for supporting the axiomatic enrichment of existing biomedical ontologies. Ghazvinian et al. [81] analyzed BioPortal ontologies to create 4 million mappings between concepts in the ontologies based on lexical similarity of concept names and synonyms and discussed how the mappings may help in the process of ontology design and evaluation. Ghazvinian et al. [82] analyzed 53 BioPortal ontologies, identified OBO Foundry candidates and examined their level of term reuse and overlapping. Vescovo et al. [83] analyzed various aspects of partitioned BioPortal ontologies using "atomic decomposition" and presented an algorithm for extracting modules from decomposed ontologies, which makes it possible to quickly identify atoms for logically complete reasoning.

CHAPTER 3

SCULPTING THE UMLS REFINED SEMANTIC NETWORK

3.1 Introduction

The Refined Semantic Network (RSN), as originally created by the SABOC Center from the UMLS Semantic Network, has a major deficiency as an abstraction network. An abstraction network needs to be compact to be effective. However, for the 1998 release of the UMLS, the RSN had 1163 ISTs and thus it was an order of magnitude bigger than the UMLS Semantic Network with its 132 semantic types in the 1998 release. This deficiency made the RSN a less attractive alternative for the Semantic Network as a UMLS abstraction network.

A long range effort has been under way to achieve the goal of eliminating illegitimate ISTs from the UMLS, in the expectation to obtain a compact RSN. This chapter is dedicated to describing the process and techniques used to "sculpt" a compact RSN out of its initial version and the results obtained. The term "sculpting" is used metaphorically, because a sculpture is created by removing the excess material from a shapeless block of raw material. In the same way, the "correct" RSN with only legitimate ISTs should emerge. As will be reported in this chapter, this goal of obtaining a compact RSN was achieved to a substantial degree, but it required a multiyear process. This process has been slowed down by the phenomenon of ISTs that had been removed from the RSN being reintroduced by the NLM due to new erroneous semantic type assignments in new UMLS releases. The AdviseEditor system, which can help the UMLS team with

preventing the reintroduction of erroneous ISTs in new UMLS releases, will be described in Chapter 4 (theory) and Chapter 5 (implementation). It is also described in [46].

The purpose of this chapter is not to introduce new methods for auditing the UMLS, but to describe various techniques previously employed to transform the RSN into a compact abstraction network. These techniques were at the time published for their own sake, but are reviewed here for their role in sculpting the RSN (and not for their own virtue.)

3.2 Methods

This chapter describes the methods which enable reshaping of the RSN into a compact abstraction network, materializing the vision defined more than a decade ago. In more than 15 years of research in Quality Assurance (QA) of terminologies of the SABOC Center, two recurring themes regarding concentration of errors in medical terminologies [84] were identified. Errors typically appear in complex concepts or in unusual concepts. The following rational is offered. Modeling of complex concepts is more difficult than modeling of other concepts, and thus they have more likelihood for errors. For "unusual" concepts, the reason for the different modeling may be the unique nature of these concepts, but also a high likelihood that the modeling is wrong, and this is why these concepts are unusual.

The interpretation of "complex" or "unusual" varies from one terminology to another according to the different natures of various terminologies. Wang et al. has shown that complex concepts in overlapping partial areas [65] have a high likelihood of errors in SNOMED CT [64, 85]. If a partial area is small, i.e., contains few concepts, these concepts

can be labeled as being unusual. It has been shown that small partial areas contain relatively more errors in SNOMED CT and NCIt [59, 84]. An IST consisting of multiple semantic types is more complex than a single semantic type, because of compound semantics. Chen has found many errors in the extents of ISTs, e.g. **Experimental Model of Disease** ∩ **Neoplastic Process** [31, 54, 86]. A small IST extent naturally contains unusual concepts, since out of 2.9 million concepts in the META, only a few concepts are assigned its semantic type combination. Thus, it is hypothesized that ISTs assigned to few concepts are more likely to have concepts with erroneous semantic type assignments, since the concepts assigned such ISTs are both complex and unusual.

In [25] a study was conducted by the SABOC Center, auditing concepts of ISTs with small extents. The finding was that for ISTs with up to six concepts there is a high likelihood of wrong semantic type assignments compared to concepts assigned an IST with a larger extent. If all the concepts assigned a specific IST with small extent (in short "small IST") have an erroneous semantic type assignment, then, after corrections are made, this IST disappears from the RSN. Over the years, several studies were conducted in the SABOC Center, e.g. [26, 52, 53], where a team of domain experts audited a sample of small ISTs. The consensus reached by the auditors was forwarded to the UMLS editors for review. In some cases the UMLS editors chose an alternative correction than the one suggested by the auditors, but the "erroneous" ISTs still disappeared from the RSN, whenever no concept was left with the IST's combination of semantic types.

This action of eliminating erroneous ISTs from the RSN is called *sculpting*, since it raises the mental image of an artist removing excess material from a block of raw material to obtain the desired sculpture. The sculpting of the RSN is continued by extending some

IST extents [31, 54], which is done after detecting concepts missing appropriate semantic type assignments. That is, the sculpting does not only involve only removing erroneous ISTs, but also obtaining the correct sets of concepts which should be assigned an IST. In other words, sometimes concepts are missing a necessary second semantic type, and correcting their assignments may increase the size of an IST extent that was not small to begin with. This phenomenon was demonstrated for the IST **Experimental Model of Disease** ∩ **Neoplastic Process** which was enlarged from 33 to 948 concepts by work of Chen et al. [54], and was further expanded to 1397 concepts using another technique by Chen et al. [86]. Similarly, the IST **Governmental or Regulatory Activity** ∩ **Intellectual Product** was expanded from 22 to 32 concepts [54]. The extent of the IST **Environmental Effect of Humans** ∩ **Hazardous or Poisonous Substance** was enlarged from three to nine concepts, i.e., it was no longer a small IST [31].

This chapter presents a newly performed audit of all ISTs with small extents (1 – 6 concepts) in the 2013AA UMLS release, removing erroneous semantic type assignments. The resulting RSN, with a smaller number of ISTs, is an outcome of this dissertation research.

3.3 Results

First, the progress of sculpting the RSN over multiple releases of the UMLS is reported. Table 3.1 presents the information monitored, including the number of concepts, number of semantic types and ISTs, number of concepts with redundant assignments and their ISTs, as well as the number of small ISTs with their extent sizes, the combined number of ISTs with extent sizes 1-6, and finally their numbers of concepts, for different UMLS

releases. The first part of Table 3.1 contains data collected previously by Morrey [49]. The data in Table 3.1 marked in yellow was collected for this dissertation research and consists of original results.

Information was regularly collected starting with UMLS version 2006AC. During 2006-2007, reports of redundant and wrong semantic type assignments for small ISTs were submitted to the NLM. For example, for the 2006AC version, 42 erroneous, small extent IST assignments were submitted, 39 of which have one concept and three have two concepts each. The NLM implemented most of the corrections, causing many small ISTs to disappear. Note that feedback from the NLM regarding the error reports was never received, but by reviewing the changes in the next UMLS release, corrections can be tracked.

Of these 42 small extent ISTs, 38 disappeared by the 2007AA version. One of these ISTs was **Mammal** ∩ **Experimental Model of Disease** assigned to the concept *Knock-in Mouse*, with erroneous compound semantics; of course a mammal cannot be a disease. Another IST that disappeared, **Congenital Abnormality** ∩ **Neoplastic Process**, which was assigned to *Port-Wine Stain*, was a forbidden combination of semantic types according to the UMLS usage note of the semantic type **Neoplastic Process** [56]. No change was made only for one IST, **Gene or Genome** ∩ **Enzyme**.

In three cases, the concept assignments were changed, but the IST remained in the RSN, because a new concept was assigned simultaneously to the same IST by the UMLS editors. In other words, in some cases new errors were introduced while old errors were being corrected.

As noted before, the NLM did not always follow the submitted suggestions. However, the changes they made in the semantic type assignments still frequently resulted in the deletion of small ISTs. Nevertheless, the total number of ISTs between 2006AC and 2007AA was only reduced from 559 to 555.

As hinted above, while some wrong small ISTs disappeared, others were created due to the assignment of multiple semantic types to new concepts coming from new sources added to the UMLS or from new releases of existing UMLS sources. A systematic decrease in the number of ISTs is evident in Table 3.1 from 2007AC on. The number of ISTs went down from 532 in 2007AC to 397 in 2008AB, a reduction of 135 ISTs, 110 of which were small ISTs with a total of 235 concepts, and in particular 78 ISTs with one or two concepts each. The removal of such ISTs from the RSN is consistent with the finding of Gu et al. [25] that concepts assigned ISTs with extents of up to six concepts have a higher likelihood of erroneous semantic type assignments. Many erroneous assignments have been removed either due to the SABOC reports (e.g.,[26]) or independently by the UMLS team. Furthermore, as mentioned in the previous section, the NLM implemented an automatic procedure for detecting all redundant assignments in the UMLS, which is applied before any new release starting in 2008 [58]. However, in the UMLS 2011AA release, **Finding** and **Sign or Symptom** are assigned to the concept C2711130 *Subungual swelling*. **Finding** is the parent of **Sign or Symptom,** thus the assignment of **Finding** is redundant and unexpected.

As mentioned above, data for Table 3.1 starting with UMLS version 2010AA (highlighted in yellow) were collected for this dissertation. During 2009 – 2013 it was observed that a plateau was reached, with about 400 ISTs, of which about 170 are small

ISTs, containing a total of about 410-420 concepts. One may think that the RSN had reached a stable state during these years. However, the impression created by the numbers of ISTs and small ISTs is misleading.

During the period from 2009 to 2013, two ongoing phenomena may be observed that have contradictory effects on the numbers of ISTs. From one side, erroneous semantic type assignments were detected by the UMLS team and as a result 69 erroneous ISTs of typically small extents disappeared (see Table 3.2). From the other side, new UMLS concepts were assigned semantic types, and for 78 of them, new combinations of semantic types were created (see Table 3.2), leading to the addition of new ISTs of typically small extents. Many times those newly created ISTs are the same ones that were removed from the RSN in earlier releases while erroneous assignments of such ISTs were corrected.

According to Table 3.2, there are 35 such ISTs over the five releases 2011AA – 2013AA. Furthermore, 11 of these ISTs were added and deleted more than once during this period. These "oscillations" were not detected, since the NLM did not adapt the RSN as an additional abstraction network for monitoring the UMLS, in spite of many publications about the RSN and the presentations about the RSN in the NLM-sponsored workshop on "Future Directions of the Semantic Network" [87]. A recommendation how to avoid such "oscillations" appears in Section 3.4.

When the new ISTs in the 2013AA and 2012AA releases of the UMLS were reviewed, it was found that most of them are illegitimate. For example, in Table 3.3, showing 11 new ISTs in the 2013AA release, the IST **Mental or Behavioral Dysfunction** ∩ **Steroid** ∩ **Pharmacologic Substance** is illegitimate, because a dysfunction cannot be a chemical**.**

Table 3.1 Progress of RSN Over Time

UMLS Release	#cpts	#ST	#IST	#cpts w/ redun-dant STs	#ISTs w/ redun-dant assign	#ISTs w/ 1 cpt	#ISTs w/ 2 cpts	#ISTs w/ 3 cpts	#ISTs w/ 4 cpts	#ISTs w/ 5 cpts	#ISTs w/ 6 cpts	#ISTs w ≤ 6 cpts	# of concepts in IST w ≤ 6 cpts
1998	476K	132	1163	8622	77	422	n/a	n/a	n/a	n/a	n/a	n/a	n/a
2001	800K	134	874	12161	40	322	113	64	35	28	25	587	1170
2006AC	1.4M	135	559	91	7	124	68	37	32	26	18	305	737
2007AA	1.4M	135	555	598	11	111	65	40	33	23	17	289	710
2007AC	1.5M	135	532	0	0	116	56	35	34	20	15	276	659
2008AA	1.6M	135	464	3	2	105	44	25	25	15	14	228	499
2008AB	1.9M	135	397	0	0	64	30	29	14	17	12	166	424
2009AA	2.1M	135	381	0	0	59	32	24	13	16	11	155	393
2009AB	2.2M	135	385	0	0	61	30	25	15	14	13	158	404
2010AA	2.2M	133	384	0	0	58	32	24	15	16	9	154	388
2010AB	2.4M	133	392	0	0	66	35	19	16	16	8	160	385
2011AA	2.4M	133	409	1	1	75	38	24	16	17	6	176	408
2011AB	2.6M	133	406	0	0	72	34	25	16	19	8	174	422
2012AA	2.6M	133	407	0	0	73	33	26	16	17	7	172	408
2012AB	2.8M	133	402	0	0	61	37	26	14	18	9	165	413
2013AA	2.9M	133	401	0	0	63	33	27	18	16	11	168	428
2013 Audit	2.9M	133	336	0	0	48	28	10	3	8	6	103	222

Amino Acid, Peptide, or Protein ∩ Pharmacologic Substance ∩ Indicator, Reagent, or Diagnostic Aid ∩ Element, Ion, or Isotope is assigned one concept *Fluciclatide F18,* which is used as radioactive probe in PET imaging according to the definition of this concept. However, the UMLS usage note of '**Indicator, Reagent, or Diagnostic Aid**' [56] states: "Radioactive imaging agents should be assigned to this type and not to the type 'Pharmacologic Substance' unless they are also being used therapeutically." Thus the assignment of **Pharmacologic Substance** is deemed wrong.

In 2012AA, **Carbohydrate ∩ Chemical Viewed Functionally** is assigned to the concept *viridaphin A(1) glucoside* (see Table 3.4). It is surprising that a general semantic type such as **Chemical Viewed Functionally** is assigned to this concept. According to the rules of the UMLS [22], each concept should be assigned the most specific applicable semantic type. The UMLS auditor used in this study, proposed to change this semantic type assignment to a grandchild of **Chemical Viewed Functionally**, namely **Antibiotic.**

Table 3.2 Progress of IST Removal in the Past Five Releases

	2011AA	**2011AB**	**2012AA**	**2012AB**	**2013AA**	**Total**
Number of ISTs	409	406	407	402	401	n/a
Number of Small ISTs	176	174	172	165	168	n/a
Number of New ISTs	23	17	13	14	11	78
Appeared Before	12	6	4	6	7	35
Repeated Previously	3	1	3	1	3	11
Number of Deleted ISTs	6	20	12	19	12	69

Table 3.3 New ISTs in UMLS Release 2013AA

New ISTs in 2013AA	Extent	Appeared							
Bacterium + Pharmacologic Substance	1	2012AA	2011AB	2011AA	2010AB	2010AA	2009AB	2008AA	2007AC
Congenital Abnormality + Finding	1	2011AA	2007AC	2007AB					
Laboratory or Test Result + Laboratory Procedure	1	2008AA	2007AC	2007AB	2007AA				
Pathologic Function + Anatomical Abnormality	1	2007AC	2007AB	2007AA					
Mental or Behavioral Dysfunction + Steroid + Pharmacologic Substance	1								
Medical Device + Indicator, Reagent, or Diagnostic Aid	4	2012AA	2008AA	2007AC	2007AB	2007AA			
Amino Acid, Peptide, or Protein + Pharmacologic Substance + Indicator, Reagent, or Diagnostic Aid + Element, Ion, or Isotope	1								
Carbohydrate + Pharmacologic Substance + Food	2								
Lipid + Pharmacologic Substance + Food	5								
Biomedical or Dental Material + Food	2	2008AA							
Biomedical or Dental Material + Element, Ion, or Isotope	1	2007AA							

Legend	
	ISTs removed once
	ISTs removed twice
	IST appeared the first time
	IST appeared the second time

Table 3.4 New ISTs in UMLS Release 2012AA

New ISTs in 2012AA	Extent	Appeared		
Bacterium + Eukaryote	1			
Therapeutic or Preventive Procedure + Biomedical or Dental Material	4			
Natural Phenomenon or Process + Indicator, Reagent, or Diagnostic Aid	1			
Medical Device + Indicator, Reagent, or Diagnostic Aid	1	2008AA	2007AB	2007AA
Medical Device + Clinical Drug	1	2010AB		
Qualitative Concept + Clinical Attribute	1			
Amino Acid, Peptide, or Protein + Biomedical or Dental Material + Inorganic Chemical	1			
Carbohydrate + Chemical Viewed Functionally	1			
Chemical Viewed Functionally + Inorganic Chemical	1			
Pharmacologic Substance + Vitamin + Indicator, Reagent, or Diagnostic Aid	2			
Pharmacologic Substance + Vitamin + Inorganic Chemical	2	2008AA	2007AB	2007AA
Pharmacologic Substance + Food	1	2008AA	2007AB	2007AA
Vitamin + Element, Ion, or Isotope	1			
Legend				
	ISTs removed once			
	ISTs removed twice			
	IST appeared the first time			
	IST appeared the second time			

Finally, the results of an audit of the 428 concepts of the small ISTs of the 2013AA version are reported. They were divided into two sets, 98 non-chemical concepts and 330 chemical concepts. The first set was reviewed by two domain experts, an MD, trained in medical terminologies (Gai Elhanan) and a PhD who specialized in techniques for auditing medical terminologies with training in Sports Medicine (Yan Chen). The second set was

audited by a Chemistry Professor (Ling Chen), experienced in auditing chemical concepts. All three auditors were using the Neighborhood Auditing Tool (NAT) [52] designed at NJIT and have previously audited UMLS concepts' semantic type assignments.

Table 3.5 summarizes the results of auditing 29 small non-chemical ISTs from 2013AA release. If all audit results were implemented in the 2013AA release, 16 out of 29 small non-chemical ISTs would disappear and two new non-chemical ISTs would be added. For example, the IST **Congenital Abnormality** ∩ **Finding** is only assigned to *Congenital abnormality of systemic artery*. However, the UMLS usage note of **Finding** [56] states that "Only in rare circumstances will findings be double-typed with either 'Pathologic Function' or 'Anatomical Abnormality'." **Congenital Abnormality** has IS-A relationship to **Anatomical Abnormality**. Thus the assignment of **Finding** should be removed. Consequently, this IST should disappear from the RSN.

Table 3.6 summarizes the results of auditing 139 small chemical ISTs from 2013AA. As can be seen, 30 (= 139 – 109) small chemical ISTs were found correct and remained in the RSN. Also 58 new chemical ISTs were created in the auditing process, leaving a balance of 88 small chemical ISTs.

In some cases, an audit resulted in a semantic type combination that added a concept to the extent of an existing IST, which may have been large or small. For example, the concept *TrioMatrix* is the only concept assigned **Amino Acid, Peptide or Protein** ∩ **Biomedical or Dental Material** ∩ **Inorganic Chemical**. This is an implantable orthopedic device, namely, a surgical bone implant composed of living or natural materials. Because **Amino Acid, Peptide, or Protein** is an **Organic Chemical**, it should not be assigned together with **Inorganic Chemical**. With the assignment of **Inorganic Chemical**

Table 3.5 Auditing Impact on 2013AA Non-Chemical ISTs of the Sculpted RSN

Extent size of IST	Starting # of Non-Chemical ISTs 2013AA	# of Non-Chemical ISTs deleted by audit	Percentage of such ISTs deleted	# of Non-Chemical ISTs added by audit	Percentage of Non ISTs added	# of Non Chemical ISTs after audit	Net reduction
1	7	5	71.4%	1	14.3%	3	57.1%
2	3	2	66.7%	0	0%	1	66.7%
3	5	3	60%	1	33.3%	3	60%
4	6	4	66.7%	0	0%	2	33.3%
5	2	1	50%	0	0%	1	50%
6	6	1	16.7%	0	0%	5	16.7%
Total	29	16	55.2%	2	6.9%	15	48.3%

Table 3.6 Auditing Impact on 2013AA Chemical ISTs of the Sculpted RSN

Extent size of IST	Starting # of Chemical ISTs 2013AA	# of Chemical ISTs deleted by audit	Percentage of ISTs deleted	# of Chemical ISTs added by audit	Percentage of ISTs added	# of Chemical ISTs after audit	Net reduction
1	56	44	78.5%	33	58.9%	45	19.6%
2	30	19	63.3%	16	53.3%	27	10%
3	22	21	95.5%	6	27.3%	7	68.2%
4	12	11	91.7%	0	0%	1	91.7%
5	14	10	71.4%	3	21.4%	7	50%
6	5	4	80%	0	0%	1	80%
Total	139	109	78.4%	58	41.7%	88	36.7%

removed, this concept is reassigned the very large IST **Amino Acid, Peptide or Protein ∩ Biomedical or Dental Material**, while the previous IST disappears.

The results of the audit of version 2013AA appear in Table 3.1. The shaded row in Table 3.1 shows the impact of this audit on the size of the RSN. Only 15 small non-chemical ISTs and 88 small chemical ISTs are left in the RSN. The total number of ISTs (small and large) decreases to 336 (fourth column, Table 3.1).

The audit reports of both samples were submitted to the NLM for review. Based on past experience, the recommendations are expected to be at least partially incorporated into the UMLS and have a positive impact on the size of the RSN.

3.4 Discussion

In the paper of McCray and Hole [24], which introduces the UMLS Semantic Network, the authors say

> "The current scope of the network is quite broad, yet the depth is fairly shallow. We expect to make future refinements and enhancements to the network based on actual use and experimentation."

This "future plan" for further development of the Semantic Network was never executed, in spite of the obvious need. For example, describing the integration of the Gene Ontology (GO) [88] into the UMLS, Lomax and McCray [89] point to deficiencies of the Semantic Network in covering the Genomics field. While the UMLS grew to be about 96 fold larger than in its first release [28], the Semantic Network changed very little, with a

few semantic types being added or deleted over the years (See, for example, the third column in Table 3.1).

In the specific field of Genomics, research proposing extensions of Genomics coverage by the Semantic Network [90, 91] was never implemented. One may consider the RSN as a step towards fulfilling the above original vision of the designers of the UMLS Semantic Network, since it adds to the network depth by adding the more refined IST categories. Another important observation is that the RSN is derived from the Semantic Network and the semantic type assignments of the META concepts in an intrinsic way, without using any knowledge sources that are external to the UMLS. The extensions provided by the RSN are thus in line with the vision for the UMLS at the time of its founding.

The RSN helps identifying ISTs with proper compound semantics and treating them as legitimate first order citizens, while removing all the semantically invalid semantic type combinations. For example, in the 2013AA release of the UMLS, 85 ISTs are assigned to at least 100 concepts, 36 ISTs are assigned to at least 500 concepts and 21 of these ISTs are assigned to at least 1000 concepts, demonstrating their validity as legitimate broad categories for META concepts.

Only 29 small non-chemical ISTs exist in 2013AA. According to the hypothesis of Gu et al. [25], concepts assigned such small ISTs have a high likelihood of wrong semantic type assignments. Indeed, many such ISTs have already disappeared in past releases. The efforts of the NLM editorial and QA teams should be applauded for achieving the current situation, by preventing redundant semantic type assignments and eliminating many erroneous small ISTs. Furthermore, even for the current (2013AA) small non-chemical

ISTs, the hypothesis of Gu et al. [25] was found true in the audit report presented here (see Table 3.5), according to which only 15 (about half) of the small non-chemical ISTs are legitimate, i.e., have proper compound semantics.

The situation of auditing small chemical ISTs is different. As mentioned earlier, ISTs are expected for chemical concepts, due to their multiple structural and functional views. As a result there are 28 ISTs which represent combinations of four chemical semantic types. For example, 118 concepts are assigned **Amino Acid, Peptide, or Protein** ∩ **Pharmacologic Substance** ∩ **Immunologic Factor** ∩ **Indicator, Reagent, or Diagnostic Aid**. While many of the small chemical ISTs are legitimate, Table 3.6 indicates that a large portion of them, (109/139) = 78%, are erroneous. However, many (58) small chemical ISTs were added during the audit, when the concepts of the deleted ISTs were reassigned. As a result, 88 small chemical ISTs were left in the RSN after the audit (see Table 3.6). The concepts of the other 51 (109 – 58) small chemical ISTs were typically reassigned existing ISTs with larger extents, as shown in the example above. The contrast between the 88 chemical and the 15 non-chemical small ISTs, reflects the frequency of categorizing chemical concepts by both structural and functional chemical semantic types, as documented in the usage note for the **Chemical** semantic type of the UMLS [56].

Interestingly, once all erroneous ISTs will have been eliminated from the RSN, the hypothesis of Gu et al. [25], which states that concepts assigned small ISTs have a high likelihood of wrong semantic type assignments, will not be true anymore. This is based on the expectation of preventing the current practice of reassigning erroneous ISTs to new UMLS concepts, which was demonstrated in the Section 3.3. This practice has turned the effort of sculpting the RSN into a Sisyphean task, since once an erroneous IST has been

eliminated by correcting the erroneous semantic type assignments, this IST often reappears in a future release, due to new erroneous semantic type assignments.

The question is what can be done to stop this phenomenon of reassigning erroneous semantic type combinations to new concepts without hurting the efficiency of the UMLS team. This issue will be the subject of Chapter 4 and Chapter 5 of this dissertation and was published by Geller et al. [46].

3.5 Conclusions

A longitudinal study of the process of improving the UMLS as a result of auditing its semantic type assignments was reported on. The main instrument used in this process is the auditing of small ISTs with high likelihood of erroneous semantic type assignments. Numerous audit reports were submitted to and reviewed by the NLM. The staff of the NLM also performed independent audits and adopted automatic testing for redundant semantic type assignments before a new UMLS version is released. Furthermore, a comprehensive group audit of all 168 small ISTs in the 2013AA version was conducted as part of this dissertation research. As a result, after the audit in this chapter is used to eliminate small ISTs, the RSN becomes compact abstraction network with a size of the same order of magnitude as the UMLS Semantic Network, providing better comprehension support for the content of the META.

The auditing data collected from 1998 to 2009 and the analysis of ISTs with small extents in the 2009AB version of the UMLS have been published [49]. The data collected from 2010AA to 2013AA and the new analysis of the ISTs with small extents in the 2013AA version of the UMLS will be submitted for peer review in the near future.

CHAPTER 4

RULE-BASED SUPPORT SYSTEM FOR MULTIPLE UMLS SEMANTIC TYPE ASSIGNMENTS

4.1 Introduction

This chapter presents a system, *adviseEditor*, that will inform an editor as to whether a specific tuple (pair, triple, quadruple, quintuple) of semantic types is permitted or prohibited. There is a need for such a system, because UMLS editors have introduced prohibited combinations of semantic types and even reintroduced them after the UMLS was corrected by eliminating those prohibited combinations. (Examples of such reintroduced combinations appear in Section 4.7.) Eight rule-categories that govern the possible interactions of pairs of semantic types are defined. Examples where concepts in the Metathesaurus violate the identified rules will be presented. If the *adviseEditor* system would have been in place when those concepts were originally introduced into the UMLS and assigned semantic types, these errors could have been prevented. Counts of semantic type pairs belonging to different rule-categories, as determined by the *adviseEditor* system will also be provided.

4.2 Background

An important conceptual tool for terminology integration into the Metathesaurus is the UMLS Semantic Network. Every concept in the Metathesaurus is assigned one or more semantic types of the Semantic Network at the time of integration [25, 92]. These assignments were performed by many UMLS editors at the National Library of Medicine over a long period of time, and thus are not necessarily done in a consistent manner.

The UMLS Semantic Network is structured as two separate trees, rooted in the semantic types **Entity** and **Event,** respectively**.** The 133 semantic types of the Semantic Network constitute its nodes and are connected by IS-A links. They are furthermore connected by 53 lateral relationship kinds. Inheritance of lateral relationships along IS-A links is by default a defined operation, except for a few cases where it is explicitly blocked.

As in previous chapters, the set of all concepts assigned a specific semantic type **T** is called the *extent of* **T,** which will be abbreviated as E(**T**)**.**

Whenever a concept is assigned two semantic types, then it is contained in the extents of both semantic types at the same time. Mathematically this means that the concept is in the *set intersection* of the two extents. As before, the mathematical symbol ∩**,** expressing intersection, will occasionally be used when describing sets of concepts that are assigned two semantic types.

In [25, 27, 29] auditing of the UMLS for inconsistencies was carried out, based on intersections of extents of semantic types. It is hypothesized [27] that concepts in small intersections have a high likelihood of wrong semantic type assignments. In a sample of 100 intersections, each containing only a single concept, analyzed by JJ Cimino [27], only 11 concepts were found to have correct semantic type assignments.

Gu et al. showed [25] that concepts assigned pairs of semantic types, such that the intersections of their extents are small, were more likely to have erroneous semantic type assignments than other concepts. In this chapter, this observation is used for developing an algorithm for classifying pairs of semantic types according to rule-categories.

This research also builds on an algorithm [57] for identifying all redundant semantic type assignments, namely assignments in which a concept is assigned the

semantic types **X** and **Y** such that **X** is a child or descendant of **Y**. Such redundant assignments are prohibited by the rules of the Semantic Network [22], and only **X** should be assigned. Assigning the respective pairs of semantic types is not legal, and they should never be assigned to the same concept. However, in the 1998 release of the UMLS, 8622 concepts were found with redundant semantic type assignments in 77 prohibited intersections [27].

To help both editors and users of the UMLS, the National Library of Medicine provides a definition for each semantic type in the Semantic Network source data. Usage notes (UNs) are provided for some, but by far not all, semantic types. Note that in the balance of this chapter, when a semantic type definition is mentioned, any usage notes attached to this definition is also used. Some usage notes include rules concerning the combination of two semantic types. These rules describe situations in which a concept assigned one semantic type may not, may, or should be assigned a specific second semantic type.

4.3 Methods

4.3.1 Text-Based Instructions

Studying the documentation of the Semantic Network, one can distinguish between two kinds of instructions, *inclusion instructions* and *exclusion instructions*. An inclusion instruction expresses the fact that two semantic types *may* be used for the same concept or even *should* be used for the same concept. An exclusion instruction expresses the fact that two semantic types *may not* be used for the same concept.

The semantic type **Anatomical Abnormality** is used here to describe the following possible parts of a usage note: (1) specification, (2) inclusion instruction, and (3) exclusion instruction. Below is the UN provided in the UMLS about this semantic type.

UN: Use this type if the abnormality in question can be either an acquired or congenital abnormality. Neoplasms are not included here. These are given the type '**Neoplastic Process**'. If an anatomical abnormality has a pathologic manifestation, then it will additionally be given the type '**Disease or Syndrome**', e.g., "Diabetic Cataract" will be double-typed for this reason.

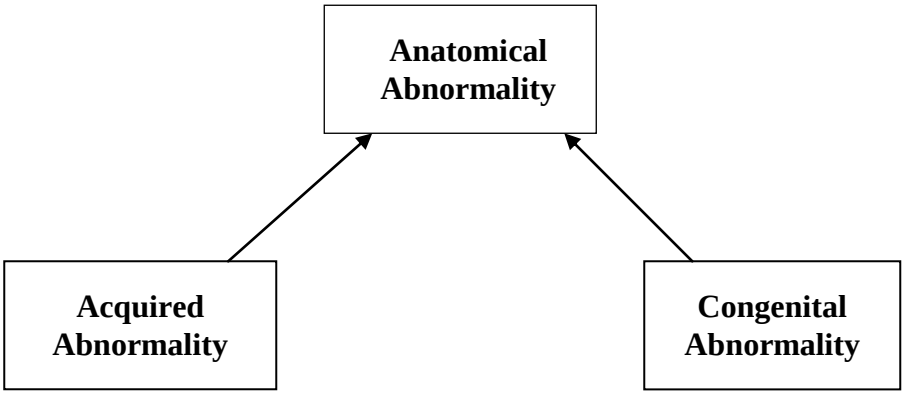

Figure 4.1 Anatomical Abnormality subhierarchy of the Semantic Network.

(1) Specification:

A specification may contain an additional explanation of what a certain semantic type stands for, or a set of requirements to be satisfied by a concept to be assigned this semantic type, or a clarification to distinguish between two semantic types.

In the above usage note of **Anatomical Abnormality** the following part corresponds to a specification. "Use this type if the abnormality in question can be either an acquired or congenital abnormality."

In this case, one needs to realize that, as shown in Figure 4.1, **Acquired Abnormality** and **Congenital Abnormality** are the two children of **Anatomical Abnormality** in the Semantic Network. This specification instruction states that for an abnormality that can be of either kind, the more general parent semantic type **Anatomical Abnormality** should be assigned. This specification implies an exclusion instruction between the two children of **Anatomical Abnormality.**

For example, the abnormalities *"intestinal defect,"* and *"pharyngeal diverticulum"* can be either acquired or congenital. Thus, the semantic type **Anatomical Abnormality** is assigned to them.

(2) Inclusion Instruction:

An inclusion instruction expresses the fact that two semantic types *may* be used for the same concept or even *should* be used for the same concept. In the above UN the following part corresponds to an inclusion instruction: "If an anatomical abnormality has a pathologic manifestation, then it will additionally be given the type '**Disease or Syndrome**'."

Thus, such a concept should be simultaneously assigned **Anatomical Abnormality** and **Disease or Syndrome.** Indeed, the Metathesaurus contains 940 concepts that are assigned these two semantic types, for example, *Dynamic subaortic stenosis.* In the Venn diagram in Figure 4.2, the intersection of extents of concepts, which are assigned **Anatomical Abnormality** and **Disease or Syndrome,** is marked by an "x."

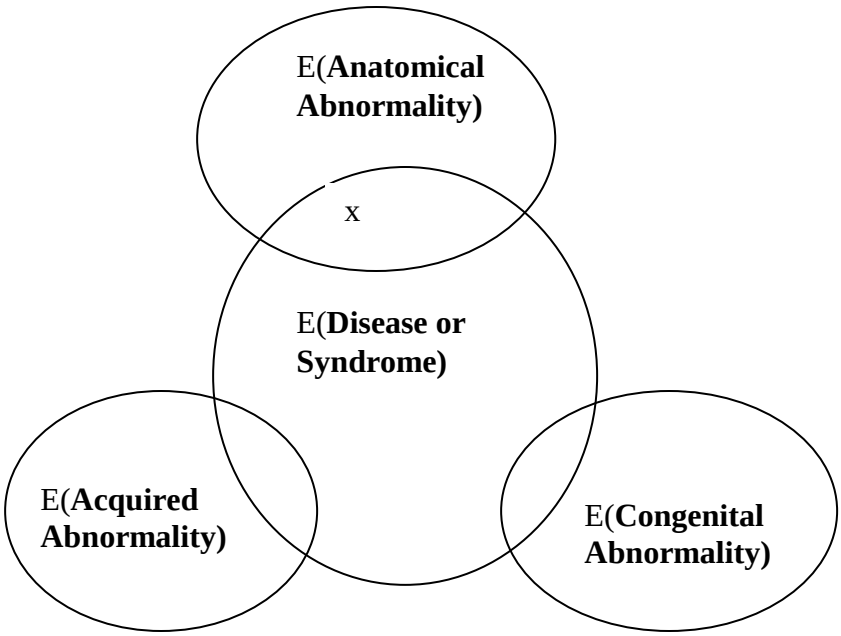

Figure 4.2 The extent of **Disease or Syndrome** intersects the extent of **Anatomical Abnormality** and the extents of its two children.

(3) Exclusion Instruction:

An exclusion instruction expresses the fact that two semantic types *may not* be used for the same concept. In the above usage note of **Anatomical Abnormality** the following part corresponds to an exclusion instruction: "Neoplasms are not included here. These are given the type **Neoplastic Process**." Hence, this exclusion instruction states that no concept is assigned both **Anatomical Abnormality** and **Neoplastic Process.** Thus, the concept *conjunctival erosion* is assigned **Anatomical Abnormality.** On the other hand, *small cell carcinoma of prostate* is assigned **Neoplastic Process.**

4.3.2 Inclusion Rules

In this research, the informal, text-based inclusion instructions of the Semantic Network documentation are mapped into precise, implemented inclusion rules. Explicit, inherited and implicit inclusion rules are distinguished. An *explicit inclusion instruction* is a description of a set of conditions under which it is valid or required for a concept to be assigned two specific semantic types. Explicit inclusion rules are derived from explicit inclusion instructions in the UMLS documentation.

Every inclusion rule has an assigned name, for example **Anatomical Abnormality with Disease or Syndrome Inclusion Rule.** In order to avoid redundant rule names, the two semantic types in a rule name are placed in alphabetical order.

Due to the inheritance of information in the Semantic Network, such a rule may have consequences, going beyond what is expressed by its name. If an explicit inclusion rule is inherited downwards in the Semantic Network, the inherited rule is then referred to as *inherited inclusion rule*.

For the semantic type **Disease or Syndrome,** the following usage note proves that the result of inheriting the **Anatomical Abnormality with Disease or Syndrome Inclusion Rule** is intended: "If an anatomic abnormality has a pathologic manifestation, then it will be given this type as well as a type from the '**Anatomical Abnormality**' hierarchy." (Refer back to Figure 4.1 to see the hierarchy.) In Table 4.1, the three inclusion rules, the numbers of concepts in the intersections of the extents of the semantic types for each rule, and examples of concepts for each rule are presented.

Table 4.1 Inclusion Rules in the **Anatomical Abnormality** Subhierarchy of the Semantic Network

Pair of Semantic Types defining an inclusion rule	**Number of Concepts**	**Example Concepts**
(Anatomical Abnormality; Disease or Syndrome)	940	*Fistula of Uterus;* *Dynamic subaortic stenosis*
(Congenital Abnormality; Disease or Syndrome)	1,392	*Atelocardia;* *Caroli Disease*
(Acquired Abnormality; Disease or Syndrome)	930	*Diabetic cataract;* *Drug-induced peptic ulcer*

An *implicit inclusion rule* cannot be derived from an inclusion instruction in the UMLS documentation. Rather, the fact that an implicit inclusion rule holds for a pair of semantic types needs to be mined from the fact that there are many Metathesaurus concepts assigned exactly this pair of semantic types. It is unlikely that all these assignments are incorrect, and therefore it may be concluded that these two semantic types may occur together. Based on the previous experience with auditing the UMLS for incorrect semantic type assignments [25, 27, 29], a pair of semantic types that has six or more assigned concepts typically defines an implicit inclusion rule.

An interesting case of an inclusion rule stating inclusion for a whole family of pairs is encountered for semantic types that are descendants of the semantic type **Chemical** in the Semantic Network. Its definition contains the following instruction: "Almost every chemical concept is assigned at least two types, generally one from the structure hierarchy and at least one from the function hierarchy." This definition implies a whole "family" of explicit inclusion rules between semantic types in the subhierarchy of **Chemical Viewed Structurally** and semantic types in the subhierarchy of **Chemical Viewed Functionally.** Furthermore, the phrase "… and *at least one* from the function hierarchy" also hints at another interesting family of inclusion rules: A chemical concept may be assigned three

semantic types: two from the **Chemical Viewed Functionally** subhierarchy and one from the **Chemical Viewed Structurally** subhierarchy.

4.3.3 Exclusion Rules

There are three categories of exclusion rules corresponding to the above three categories of inclusion rules, and an additional category called *redundancy exclusion rules. Explicit exclusion rules* are derived from explicit exclusion instructions in the UMLS documentation. Inheritance may spread an explicit exclusion rule of a pair (**A**; **B**) of semantic types to all pairs of semantic types (**C**; **D**), such that **C** is a descendant of **A** and **D** is a descendant of **B** in the hierarchy of the Semantic Network. (In this case, children are included among descendants. In addition, **A**=**C** or **B**=**D** may also hold, but not both.) The results of this inheritance process are *inherited exclusion rules*. *Implicit exclusion rules* are defined based on the following reasoning. If there is not a single concept in the over 2.6 million concepts of the 2011AB release of the UMLS that is assigned a certain pair of semantic types, then it is quite likely that this pair consists of two semantic types that should not occur together, because their combination does not categorize any existing concept in biomedicine. The status of an implicit exclusion rule may change, if such a concept is discovered, but only after an investigation and approval process of a senior UMLS editor, authorizing such a decision.

As for inclusion rules, names are assigned to exclusion rules. Previously, it was shown that the text of the usage note of **Anatomical Abnormality** contained an explicit exclusion instruction, excluding the use of the semantic type **Neoplastic Process** together with it. The corresponding rule is named the **Anatomical Abnormality excluding Neoplastic Process Rule.** The semantic types in the rule name are again in alphabetical

order. A few interesting exclusion rules of the different categories will be reviewed in the subsections below.

4.3.3.1 Explicit Exclusion Rules. As an example of an explicit exclusion rule, the children of **Finding** (**Laboratory or Test Result** and **Sign or Symptom**) are mutually exclusive by definition. This implies the **Laboratory or Test Result Excluding Sign or Symptom** Rule.

In the UMLS documentation it is made explicit that the **Anatomical Abnormality Excluding Neoplastic Process Rule** also applies to the children of **Anatomical Abnormality.** (**Neoplastic Process** has no children). Because of this, there should be no concepts in the Metathesaurus that are simultaneously assigned semantic types from the **Anatomical Abnormality** subhierarchy and **Neoplastic Process**. Surprisingly, however, there were a few such concepts in earlier releases of the UMLS, as Table 4.2 shows for version 2007AC. The last column in Table 4.2 shows the corrected semantic type assignments for those concepts in both the 2009AA and 2011AA releases of the UMLS.

Table 4.2 Two Previous Violations of Exclusion Rules in the Metathesaurus and their Corrections

Illegal Pair of semantic types in 2007AC	**Number of Concepts in 2007**	**Concepts with Illegal Assignments**	**Corrected semantic type Assignment of Concept in the UMLS in 2009AA and 2011AA**
(Anatomical Abnormality; Neoplastic Process)	1	*Acquired arteriovenous aneurysm*	**Pathologic Function**
(Congenital Abnormality; Neoplastic Process)	1	*Congenital melanocytic nevus*	**Neoplastic Process**

4.3.3.2 Inherited Exclusion Rules. Examples of inherited exclusion rules will be discussed in the Results Section, in Subsection 4.4.2.2.

4.3.3.3 Redundancy Exclusion Rules. According to the instructions of the National Library of Medicine, redundant assignments of semantic types are prohibited [58] in the UMLS. In other words, if one semantic type is assigned to a concept, then the parent and (if they exist) ancestors of this semantic type may not be assigned to this concept. Thus, it is possible to create a list of pairs of a semantic type and each of its ancestors (including the parent). Every element in this list defines an exclusion rule.

For example, the semantic type **Neoplastic Process** has the parent **Disease or Syndrome**. Its non-parent ancestors are **Pathologic Function, Biologic Function, Natural Phenomenon or Process, Phenomenon or Process** and **Event.** Thus the pairs (**Neoplastic Process; Disease or Syndrome**), (**Neoplastic Process; Pathologic Function**), (**Neoplastic Process; Biologic Function**), etc. are prohibited combinations. Each of these pairs defines a redundancy exclusion rule.

The Semantic Network contains 88 leaf semantic types, i.e., semantic types without children. Each leaf defines a unique path, starting at the leaf and ending at one of the two roots, **Entity** or **Event**. The root nodes of the Semantic Network are at level zero. Each child of a *node at level m* is considered to be *at level m+1*, thus a level number can be assigned to every node in the Semantic Network. Furthermore, a path from a node **A** at level *m* to its root will contain *m* nodes (excluding **A** itself). This numbering is convenient and is the reason for the choice that the root is assigned the level 0 instead of 1.

Under these assumptions, a semantic type at level *m* excludes all the *m* semantic type(s) above it. This holds true for leaf nodes and for non-leaf nodes. Thus, to compute the

total number of prohibited pairs of semantic types, the distribution of semantic types over levels is needed. Given that the Semantic Network has semantic types at levels 0 to 7, the total number of prohibited pairs *(PP)* can be computed as the product of the number *S(m)* of semantic types at a level *m* with the level number *(m)*, summed over all levels.

$$PP = \sum_{m=1..7} m * S(m) \qquad (1)$$

4.3.3.4 Implicit Exclusion Rules. When given *n* elements, there are *n*(n-1)/2* ways to choose a pair out of these *n* elements, assuming that pairs are order independent, and an element cannot form a pair with itself. Hence, there are potentially 133*(133-1)/2 = 8778 pairs of semantic types. Out of this total of 8778 distinct semantic type pairs, there are only 199 pairs for which concepts have been assigned this combination of two semantic types in the UMLS in version 2011AB.

If a pair of semantic types is not assigned to any concept, i.e., the intersection of their extents is empty, then one should wonder whether this pair should be defined as exclusive. However, with 8579 (=8778–199) candidate pairs such an investigation is difficult. For some of these pairs exclusion rules of the other categories were discussed earlier. But those amount only to a small fraction of the 8579 possibilities.

A pair of semantic types that is not assigned to any concept is assumed to define an implicit exclusion rule. This is similar to the closed world assumption in logic programming, which states that if a fact is not explicitly known, it is assumed not to hold (negation as failure) [93].

4.3.4 Implementation of the Inclusion and Exclusion Rules in a Computer System

An algorithm *adviseEditor* that is passed two or more semantic types as input and returns the rule-category that applies to these semantic types was developed. For reasons of exposition, the description and the algorithm for the simplest case, where only two semantic types are assigned to a concept, will be discussed first. At the end of this section an explanation is given how the system is extended to handle cases where a concept is assigned more than two semantic types.

Redundancy exclusion is the result of a pair of semantic types standing in an ancestor/descendant (or parent/child) relationship in the Semantic Network. Thus, the test for this case is expressed in the algorithm below by ((S1 is an ancestor of S2) OR (S2 is an ancestor of S1)). For the purpose of the algorithm, parents are treated as ancestors. Explicit inclusion rules and explicit exclusion rules cannot be found algorithmically at the current state-of-the-art, as they are based on natural language descriptions in the UMLS documentation. Thus, the list of pairs (**S1**; **S2**) and their mirror images (**S2**; **S1**) that fall into the explicit inclusion and explicit exclusion rule-categories were found by manual research and then pre-stored in two arrays of semantic type pairs, called Explicit_Inclusions_Array and Explicit_Exclusions_Array.

Cases of inclusion and exclusion that are based on inheritance are processed by looking upward in the Semantic Network, with the purpose of finding semantic types that are parents or ancestors that could be the source of inheritance of a specific inclusion or exclusion rule. Thus they do not need to be pre-stored.

Some pairs of semantic types may be categorized in contradictory ways, due to different rules. For example the pair (**Anatomical Abnormality**; **Neoplastic Process**) is

explicitly excluded in the UN of the semantic type **Anatomical Abnormality**. However, the same pair may also be categorized by an inherited inclusion rule, since the pair (**Anatomical Abnormality**; **Disease or Syndrome**) is categorized with an explicit inclusion rule, due to a remark in the UN of **Anatomical Abnormality** about concepts that should be also assigned **Disease or Syndrome,** and because **Disease or Syndrome** is the parent of **Neoplastic Process.** A similar contradiction may also occur between an explicit exclusion rule and cases of implicit inclusion or "more research required." In all such cases, the explicit rule (either inclusion or exclusion) should override the other kinds of rules. In the algorithm below this preference is implemented by checking for explicit inclusion and explicit exclusion before checking for other options such as inheritance. The symbol ∈ is read as "is in." Two vertical bars | | define the number of elements of the set in between them.

Algorithm **adviseEditor**(S1 SemanticType, S2 SemanticType) {
if (S1= S2)
  {return 'Input not valid'}
if ((S1 is an ancestor of S2) OR (S2 is an ancestor of S1))
  {return 'Prohibited by Redundancy Exclusion'}
else if (S1, S2) ∈ Explicit_Inclusions_Array
  {return 'Permitted by Explicit Inclusion'}
else if (S1, S2) ∈ Explicit_Exclusions_Array
  {return 'Prohibited by Explicit Exclusion'}
else if (any_ancestor(S1), any_ancestor(S2)) ∈ Explicit_Inclusions_Array
  {return 'Permitted by Inherited Inclusion'}
else if (any_ancestor(S1), any_ancestor(S2)) ∈ Explicit_Exclusions_Array
  {return 'Prohibited by Inherited Exclusion' }
else if (|Extent(S1) ∩ Extent(S2)| >= 6)
  {return 'Most likely Permitted by Implicit Inclusion'}
else if (|Extent(S1) ∩ Extent(S2)| = 0)
  {return 'Most likely Prohibited by Implicit Exclusion'}
else if (|Extent(S1) ∩ Extent(S2)| is between 1 and 5)
  {return 'More Research Required.
  Check all Concepts that are assigned both S1 and S2.
  If *at least one* is simultaneously, correctly assigned S1 and S2,
    this pair is Permitted by Implicit Inclusion.

```
        If they are all wrongly assigned either S1 or S2 or both,
          this pair is  Prohibited by Implicit Exclusion.' }
}
```

This algorithm is a concise summary of the computer implementation described in Section 4.4 and Chapter 5. However, a database lookup table was utilized to accelerate the performance of the *adviseEditor* system. For example, the line |Extent(S1) ∩ Extent(S2)| >= 6 requires a multi-step computation. The two vertical bars | | indicate that the number of elements of the set between them is returned. Similarly, the line (any_ancestor(S1), any_ancestor(S2)) ∈ Explicit_Inclusions_Array requires an extensive computation. Such results were stored in a database lookup table. The algorithmic notation hides these complications from the reader.

The *adviseEditor* algorithm was executed for every pair of distinct semantic types from the Semantic Network, and the rule-category for each pair was recorded. The total number of occurrences of each rule-category was then computed. These numbers will be reported in Section 4.4. While testing the algorithm, contradictions between rule-category assignments and actual concept assignments in the Metathesaurus were found. These contradictions will be reported in Section 4.4.

How about cases where a concept is assigned more than two semantic types? The case of a concept assigned three semantic types will be discussed in detail. The cases of more semantic types will be handled analogously, as will be explained later.

Let **S1**, **S2** and **S3** be the three semantic types assigned to a concept *C*. (**S1**; **S2**; **S3**) is a triple of semantic types. In the documentation of the UMLS the possibility of an exclusion rule for three or more semantic types is not mentioned. However a triple (**S1**; **S2**; **S3**) is excluded if any of the three pairs (**S1**; **S2**), (**S1**; **S3**) or (**S2**; **S3**) is excluded. Hence,

when considering a triple (**S1**; **S2**; **S3**) *adviseEditor* will test each of the three pairs for explicit exclusion, inherited exclusion and redundancy exclusion. If any of these rules holds for any of the three pairs, the triple is also excluded according to the most stringent rule-category of all the excluded pairs. (In this context redundancy exclusion is more stringent than explicit exclusion, which in turn is more stringent than inherited exclusion.)

With regard to inclusion rules for triples the situation is different. The definition of **Chemical** contains the following instruction: "Almost every chemical concept is assigned at least two types, generally one from the structure hierarchy and at least one from the function hierarchy" (see Section 4.3.1). This implies the possibility of an inclusion rule for triples (**S1**; **S2**; **S3**) where **S1** is a descendant of **Chemical Viewed Structurally** and **S2** and **S3** are descendants of **Chemical Viewed Functionally**. Such assignments of three semantic types occur only in the subtree rooted at **Chemical**. No other possibility of an *inclusion rule* for three or more semantic types is mentioned, which eliminates explicit inclusion and inherited inclusion rules for triples, unless one semantic type is a descendant of **Chemical Viewed Structurally** and two are descendants of **Chemical Viewed Functionally**.

What about other kinds of triples? If any of the three semantic types is not a descendant of **Chemical**, then the triple is categorized as implicit exclusion, since there are no concepts with such triples in the UMLS. All concepts assigned more than two semantic types are chemical concepts.

For cases of three descendants of **Chemical** that do not follow the pattern of the above inclusion rule, e.g., there could be two structural and one functional semantic type, first, their three pairs are tested for explicit, inherited or redundancy exclusion as described

above. If no pair is excluded, these triples are handled just like pairs of semantic types. If a triple is assigned to more than five concepts, it defines an implicit inclusion rule. If no concept is assigned such a triple, it defines an implicit exclusion rule. Finally, if a triple is assigned to between one and five concepts, its status will be "more research required." There are only 178 triples of semantic types assigned to concepts. Most of them follow the pattern of one structural chemical and two functional chemical semantic types of the above explicit inclusion rule. The few remaining triples are stored in a database lookup table where they are listed with corresponding numbers of concepts, allowing fast processing.

An interesting research issue arose out of the fact that sometimes a quadruple (4) or quintuple (5) of semantic types is assigned to one or more concepts. If the combination of four semantic types is allowed, then any three of those four (or five) must also be allowed together.

For the quadruple case there are four different possibilities to choose three semantic types from them. For the quintuple case, the number of ways to choose three out of five is computed by: 5*4 / (5-3)! = 20/2 = 10 possibilities. There are only 31 quadruples of semantic types assigned to concepts in the UMLS. Furthermore, only triples that do not follow the pattern of one structural and two functional semantic types need to be considered. The number of triples added to the database lookup table in this way is quite limited, since most of these triple are already in the database lookup table, due to their independent existence as triples of semantic types assigned to concepts.

For example, for the quadruple (**Amino Acid, Peptide, or Protein; Pharmacologic Substance; Immunologic Factor; Indicator, Reagent, or Diagnostic Aid**) assigned to 146 concepts, three triples follow the pattern of one structural and two

functional chemical semantic types, so only one triple consisting of the last three functional semantic types needs to be considered. But this triple already appears independently in the UMLS, assigned to 94 concepts.

The only quintuple in the UMLS, (**Amino Acid, Peptide, or Protein; Pharmacologic Substance; Biologically Active Substance; Indicator, Reagent, or Diagnostic Aid; Hazardous or Poisonous Substance**) is assigned to only one concept *131I-TM-601***.** The *adviseEditor* system categorizes this quintuple as "Explicit Exclusion," because one of its pairs **Pharmacologic Substance** and **Hazardous or Poisonous Substance** is categorized as "Explicit Exclusion." In other words, there is not a single valid quintuple in the UMLS, and therefore no triples derived from a quintuple were added to the database lookup table.

The details of processing the quadruples are analogous to the treatment of those triples that do not follow the above mentioned explicit inclusion rule for triples. For brevity, these details are not discussed here. Since there are currently no cases of six semantic types assigned to a concept (for the whole UMLS), such a case is not incorporated into the *adviseEditor* system. The implementation of the procedure for handling between three and five semantic types was a straightforward extension of the code for pairs, and therefore no code is provided.

4.3.5 Evaluation of the AdviseEditor System

The *adviseEditor* system is only needed for UMLS concepts assigned more than one semantic type. In order to evaluate the effectiveness of the *adviseEditor* system, a sample of concepts is generated as follows. Pairs of non-chemical semantic types such that there is at least one and there are at most five concepts with those pairs assigned are selected. This

sample was processed with the *adviseEditor* system. The sample concepts were also reviewed by a human auditor. These review results were used to evaluate the performance of the *adviseEditor* system.

This choice of concepts for the sample is based on the fact that combinations of semantic types assigned to just a few concepts as problematic are considered. Such combinations of semantic types will be assigned "more research required" by *adviseEditor*. Those are the kinds of concepts where the *adviseEditor* system is more likely to fail and needs to be tested. In contrast, the system is expected to perform relatively better for combinations of semantic types assigned to many concepts, such as for example the 658 concepts assigned the semantic types **Vitamin** and **Pharmacologic Substance**.

The problematic nature of the former kind of combinations is expressed by the fact that the "more research required" result is returned by the *adviseEditor* system only after all the other choices have been tested. Thus, even though a concept with two assigned semantic types may fulfill the conditions of "more research required," the two semantic types may also fulfill more stringent conditions, such as explicit exclusion. Indeed, this was found to be the case for several concepts in this sample, as will be described in Section 4.7.

4.4 Results

4.4.1 Inclusion Rules for Chemical Semantic Types

For brevity, not all inclusion rules, but only two especially interesting cases are covered.

4.4.1.1 Inclusion Rules between Chemical Viewed Structurally & Chemical Viewed Functionally Semantic Types. As explained in Section 4.3.1, there is a family of

explicit inclusion rules where the first semantic type is a descendant of **Chemical Viewed Structurally** and the second is a descendant of **Chemical Viewed Functionally**. There are 10 descendants of **Chemical Viewed Structurally** and 20 of **Chemical Viewed Functionally**. Hence, the total number of explicit inclusion rules for this family is 10 x 12 = 120. For example, there are 82,059 concepts assigned the pair (**Organic Chemical**; **Pharmacologic Substance**).

4.4.1.2 Pairs of Chemicals Viewed Functionally Inclusion Rules. As explained in Section 4.3.1, there is a family of explicit inclusion rules where both semantic types are descendants of **Chemical Viewed Functionally**. **Chemical Viewed Functionally** has 12 descendants. The total number of potential explicit inclusion rules in this case is (12 * 11)/2 = 66. Table 4.3 shows the numbers of concepts in intersections of descendants of **Chemical Viewed Functionally** with each other. Column headers are identical to row names and are abbreviated as needed. The children of **Pharmacologic Substance** and **Biologically Active Substance** are listed following them, respectively. The first column in Table 4.3 shows that **Pharmacologic Substance** has intersections with large extents with most other semantic types in the **Chemical Viewed Functionally** subhierarchy. The only empty intersection is with **Receptor**.

The intersection of **Pharmacologic Substance** with **Antibiotic** in Table 4.3 is marked “redundant," since the assignment of **Antibiotic** to a concept makes the assignment of **Pharmacologic Substance** to this concept redundant. Out of 66 pairs of semantic types, only 27 are actually assigned to concepts. The difference between the 66 explicit inclusion rules and the 27 non-empty intersections reinforces the fact that explicit inclusion rules enable a combination of semantic types, but the option is not always materialized.

The same observation holds true for the family of inclusion rules in Section 4.4.1.1. For some of the 120 rules there are currently no concepts. For example, the pair (**Receptor**; **Organic Chemical**) is not assigned to any concept.

Figure 4.3 shows a three dimensional view of a matrix consisting of intersections of extents of pairs of semantic types from the **Chemical Viewed Functionally** subhierarchy**.** The number of concepts in an intersection is expressed by the height of the corresponding bar. In order to better differentiate the heights of the bars, a logarithmically scaled *z* axis is used.

As can be seen in Figure 4.3, **Pharmacologic Substance** has intersections with large extents with most other semantic types in the **Chemical Viewed Functionally** subhierarchy (see second row of bars in Figure 4.3, starting from the front).

This figure is symmetrical, having the same set of semantic types on the *x* and the *y* axes. There are no bars in the diagonal (meaningless pairs of a semantic type with itself). However, each pair of semantic types is displayed at both possible locations to simplify the mental retrieval from this three-dimensional view, since by following the horizontal color coding, one can easily see all intersections of a given semantic type. The total number of potential bars in Figure 4.3 is (12 * 12 – 12) = 132. The difference between the 132 potential bars and the 54 visible bars constitutes another way of visualizing the fact that possible pairs of semantic types are not always materialized.

Table 4.3 Intersections of Pairs of Descendants of **Chemical Viewed Functionally** with Each Other

--	Pharma-cologic Sub-stance	Anti-biotic	Biomed-ical or Dental Material	Biologi-cally Active Subs-tance	Neurore-active Substance or Biogenic Amine	Hor-mone	En-zyme	Vi-ta-min	Immu-nolo-gic Factor	Re-ce-pt-or.	Indicat. Rea-gent or Diag. Aid	Hazard or Poison. Subs tance
Pharmacologic Substance	--	--	--	--	--	--	--	--	--	--	--	--
Antibiotic	Redun-d ant	--	--	--	--	--	--	--	--	--	--	--
Biomedical or Dental Material	158	0	--	--	--	--	--	--	--	--	--	--
Biologically Active Substance	803	17	3	--	--	--	--	--	--	--	--	--
Neuroreactive Substance or Biogenic Amine	9	0	0	Redundant	--	--	--	--	--	--	--	--
Hormone	96	0	0	Redundant	12	--	--	--	--	--	--	--
Enzyme	93	0	0	Redundant	0	0	--	--	--	--	--	--
Vitamin	658	0	2	Redundant	0	0	0	--	--	--	--	--
Immunologic Factor	2234	0	0	Redundant	0	0	1	0	--	--	--	--
Receptor	0	0	0	Redundant	0	1	3	0	12	--	--	--
Indicator, Reagent or Diagnostic Aid	479	5	16	3	0	0	1	0	137	0	--	--
Hazardous or Poisonous Substance	97	0	1	498	0	0	10	0	9	0	3	--

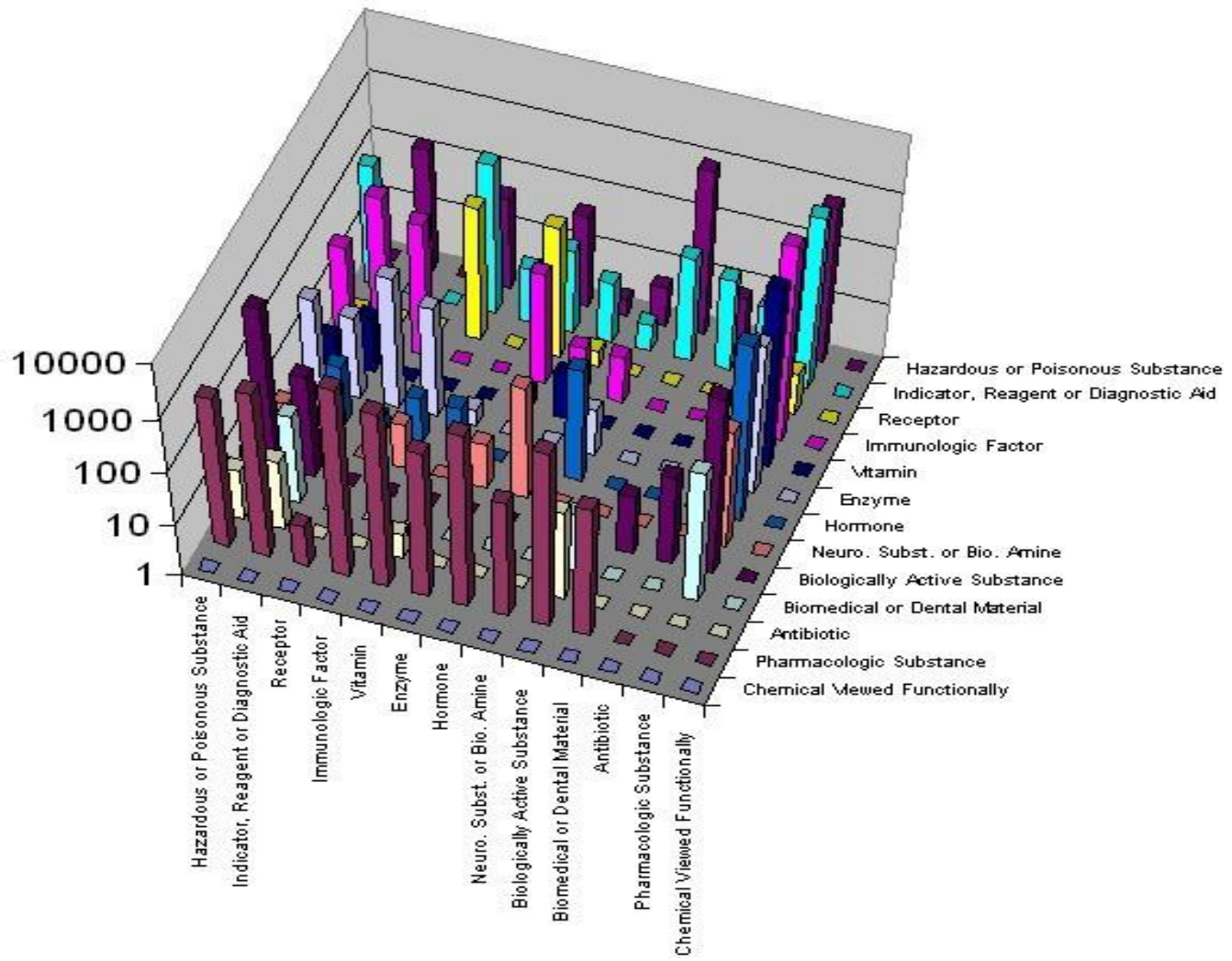

Figure 4.3 Intersections of pairs of functional chemical semantic types.

4.4.2 Exclusion Rules Results

For brevity, only interesting and typical cases of exclusion rules are presented.

4.4.2.1 Explicit Exclusion Rules. The UN of the semantic type **Finding** contains the instruction that "Only in rare circumstances will findings be double-typed with either **'Pathologic Function'** or **'Anatomical Abnormality'.**" this usage note is interpreted to imply two explicit exclusion rules, the **Finding Excluding Pathologic Function Rule** and the **Anatomical Abnormality Excluding Finding Rule**.

For the semantic type **Activity** the UN contains the instruction "In general, concepts will not receive a type from both the '**Activity**' and the '**Behavior**' hierarchies." This expresses the **Activity Excluding Behavior Rule**.

The definition of **Organophosphorus Compound** contains the instruction that "Excluded are phospholipids, sugar phosphates, phosphoproteins, nucleotides, and nucleic acids." This implies four exclusion rules, which are the **Lipid Excluding Organophosphorus Rule,** the **Amino Acid, Peptide or Protein Excluding Organophosphorus Rule** the **Carbohydrate Excluding Organophosphorus Rule** and the **Nucleic Acid, Nucleoside, or Nucleotide Excluding Organophosphorus Rule.**

Table 4.4 lists eleven pairs of semantic types for which an explicit exclusion rule exists, nevertheless, concepts have been assigned to those pairs. The number of problematic concepts for each exclusion rule is listed in Column 2 and a sample concept is listed in Column 3. All 278 concepts referred to in Table 4.4 have a wrong semantic type assignment, according to an explicit exclusion rule.

Table 4.4 Eleven Pairs Prohibited by Explicit Exclusion, with Concept Assignments

Pairs of semantic types defining an explicit exclusion rule	**# of Conc.**	**Example concept**
(Medical Device; Research Device)	12	C0600364 *Biosensors*
(Nucleic Acid, Nucleoside, or Nucleotide; Organophosphorus Compound)	25	C0674527 *5'-O-phosphonylmethylthymidine*
(Hazardous or Poisonous Substance; Pharmacologic Substance)	97	C0145114 *teleocidin B*
(Element, Ion, or Isotope; Inorganic Chemical)	10	C2347051 *Mn2+*
(Amino Acid, Peptide, or Protein; Organophosphorus Compound)	46	C0064331 *keyhole limpet hemocyanin phosphonamidate conjugate*
(Carbohydrate; Organophosphorus Compound)	46	C0063569 *inositol 1,4,5-triphosphorothioate*
(Lipid; Organophosphorus Compound)	35	C0256611 *EPC-NPH*
(Body Substance; Pharmacologic Substance)	1	C1976001 *Blood product units & Blood product unit*
(Organic Chemical; Inorganic Chemical)	1	C2975881 *Ringerfundin*
(Finding; Pathologic Function)	2	C0267995 *Fluid volume disorder*
(Organic Chemical; Element, Ion, or Isotope)	3	C0302933 *Natural graphite*
TOTAL	278	

The semantic type **Clinical Drug** has a UN with the instruction "Do not double type with **Pharmacologic Substance**, **Antibiotic**, or other chemical semantic types." This defines yet another family of explicit exclusion rules.

4.4.2.2 Inherited Exclusion Rules. If **Finding** excludes **Pathologic Function** (see above)**,** then, by inheritance of explicit exclusion rules, **Finding** should also exclude the descendants of **Pathologic Function,** such as **Disease or Syndrome.** In version 2007AC, many concepts contradicting such exclusion rules existed. These were corrected in version

2009AA. In that version, **Finding** did not have any concepts with a second semantic type assigned to them.

However, in version 2011AA, **Finding** and **Pathologic Function** were assigned to two concepts, in spite of the explicit exclusion rule. Furthermore, **Finding** and **Disease or Syndrome** are both assigned to three concepts, in contradiction to inherited exclusion. In addition, **Finding** is assigned to other groups of concepts that are assigned additional semantic types in contradiction to exclusion rules, as follows: **Finding** and **Sign or Symptom** (1 concept) (redundancy exclusion), **Finding** and **Acquired Abnormality** (1) (inherited exclusion), and **Finding** and **Congenital Abnormality** (2) (inherited exclusion). In total, there are nine new assignments that have been introduced into the UMLS for **Finding,** between version 2009AA and version 2011AA, that are likely to be erroneous. For example, in version 2011AA, E(**Finding)** ∩ E(**Acquired Abnormality)** contains the concept *Flexion contracture of proximal interphalangeal joint.*

In summary, a set of errors was corrected between 2007 and 2009 and then new errors violating these rule-categories were introduced by 2011. This indicates the importance for consulting the *adviseEditor* system before assigning a pair of semantic types to a new concept.

4.4.2.3 Redundancy Exclusion Rules. As noted in Section 4.3.3.3, there are 88 leaves in the two trees in the Semantic Network. Every one of these leaves defines a path to its respective root. In total, there are 2 semantic types at level 0, 4 are at level 1, 20 at level 2, 40 at level 3, 24 at level 4, 19 at level 5, 21 at level 6 and 3 at level 7.

Using formula (1) from Section 4.3, with 4 * 1 + 20 * 2 + 40 * 3 + 24 * 4 + 19 * 5 + 21 * 6 + 3 * 7 gives exactly 502 redundancy exclusion rules, which correspond to about

5.7% of the 8778 pairs of semantic types. This result is in agreement with the result found by the program.

4.4.3 The Rule-Category "More Research Required"

The previous research shows that when there are six or more concepts assigned a pair of semantic types, unless appearing as an explicit exclusion rule or inherited exclusion rule, one can safely assume an implicit inclusion rule [25]. Similarly, one can safely assume an implicit exclusion rule when there are no concepts assigned a pair of semantic types. However, what happens when between one and five concepts have been assigned a specific pair of semantic types?

In such a case, the UMLS editor will need to investigate all those concepts, whether the assignment of these two semantic types is really justified. If all such concepts are modified such that they do not have this pair of semantic types assigned, then the pair will be converted into a case of implicit exclusion. In that case, no new concepts may be assigned this pair of semantic types. On the other hand, if the assignment of these two semantic types is justified for an existing concept, this pair should be transitioned to the status of implicit inclusion rule and may also be assigned to a new concept.

In the 2011AA version of the UMLS, 30 pairs of semantic types assigned the rule-category "more research required" were found.

4.4.4 Numbers of Semantic Type Pairs in Each Rule-category

Table 4.5 shows the numbers of pairs of semantic types (**S1**; **S2**) assigned to each rule-category. The results in rows 1 to 8 follow exactly the order in which the corresponding tests are performed in the algorithm *adviseEditor*. The pairs (**S1**; **S2**) and (**S2**; **S1**) are only counted once.

Table 4.5 Numbers of Semantic Type Pairs in Each Rule-category

Row #	Rule-Category	Number of Occurrences
1	Redundancy Exclusion	502
2	Explicit Inclusion	181
3	Explicit Exclusion	104
4	Inherited Inclusion	30
5	Inherited Exclusion	71
6	Implicit Inclusion	34
7	Implicit Exclusion	7826
8	More Research Required	30

4.4.5 Visualizing the Space of Semantic Type Pairs

While concentrating on an algorithmic treatment of inclusion and exclusion rules, the question naturally arises whether pairs of semantic types could not be displayed as a two-dimensional matrix. Displaying a matrix with 8778 numerical values on 8.5″ by 11″ paper is impossible. However, a diagram approximating such a display using color coding is presented.

Figure 4.4 shows color-coded rule-categories for pairs of semantic types. The 133 semantic types are numbered by the NLM from T001 to T203 (there are gaps). Every point encodes the pair of semantic types defined by its values on the *x* and *y* axes. The diagonal through the origin (T001, T001) defines pairs of identical semantic types.

The semantic type **Entity** (T071) naturally is excluded by the largest number of other semantic types due to redundancy exclusion, as it is the root of the larger of the two trees of the Semantic Network. Thus, the longest orange lines in the diagram are at the row and column of T071. Other long lines are at T051, which correspond to **Event,** the other root of the Semantic Network. Together, these two semantic types are excluded by every other semantic type, except by each other. Thus, the lines at T071 and T051 cover almost the complete *x* dimension and *y* dimension of the diagram.

In Figure 4.4, an area of red marks explicit inclusion, above and to the right of T103 (**Chemical**). This illustrates the inclusion rules among the **Chemical Viewed Functionally** semantic types, discussed in Section 4.4.1.2 and between the **Chemical Viewed Functionally** and the **Chemical Viewed Structurally** semantic types, discussed in Section 4.4.1.1.

4.4.6 Evaluation Study for the Performance of the AdviseEditor System

In order to evaluate the performance of the *adviseEditor* system, a sample of concepts was generated follows. All pairs of non-chemical semantic types in the 2011AA UMLS release were determined, such that there was at least one and there were at most five concepts with those pairs assigned. There are only 32 such pairs in the release. Then all 65 concepts assigned any one of these 32 pairs of semantic types were processed with the *adviseEditor* system. These 65 concepts were also reviewed by a human auditor, Dr. Julia Xu, trained in both medicine and medical terminologies. Dr. Xu is not an expert in chemistry, thus the study was limited to the non-chemical combinations. Naturally, the auditor was not given access to the *adviseEditor* system.

Among the 32 pairs of semantic types audited, the 16 pairs listed in Table 4.6 are new in the 2011AA version of the UMLS. The column Rule-category indicates which category the pair of semantic types in this row belongs to. The column #cpts contains the number of concepts that are assigned this pair of semantic types. Notably, the column Rule-Category indicates a kind of exclusion rule for every pair in Table 4.6, and what kind of exclusion rule it is. Thus, the column #cpts (number of concepts) should ideally contain 0 in every row.

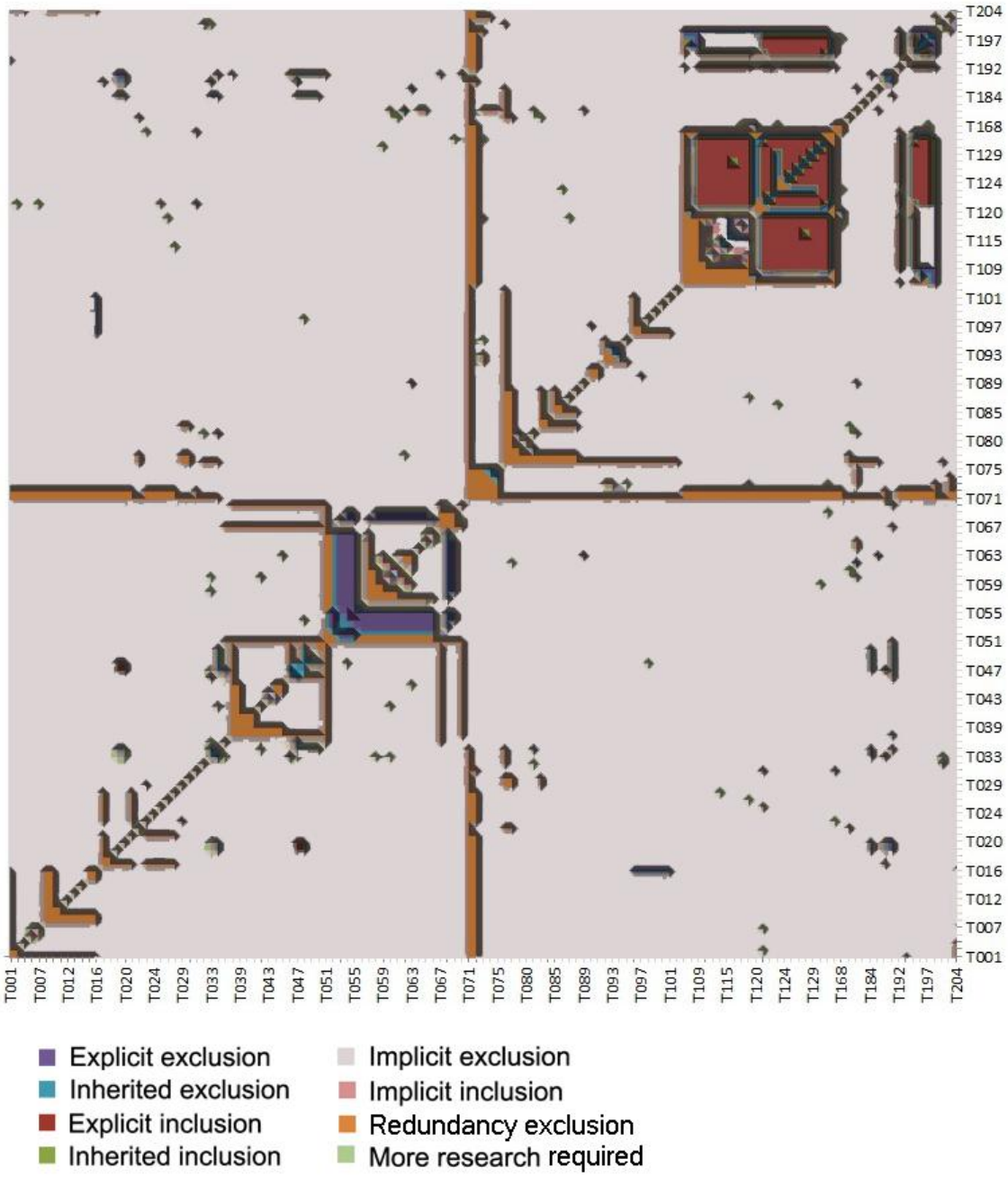

Figure 4.4 Color-coded rule-categories for pairs of semantic types.

The last column, "Appeared in previous UMLS release?" shows whether and when a pair appeared in a previous UMLS release prior to 2010AB, before it disappeared subsequently due to auditing efforts, and (re)appeared in the 2011AA release. Nine out of

the 16 pairs appeared in the past, according to research covering the period from 2006AC to 2010AB.

For six of the 16 rows in Table 4.6, using the *adviseEditor* system would have warned the UMLS editors about introducing erroneous pairs of semantic types for new concepts, because these pairs contradict explicit exclusion, inherited exclusion or redundancy exclusion. For example, **Finding** and **Pathologic Function,** a case of explicit exclusion, are assigned to *Fluid volume disorder*. The auditor suggested assigning **Sign or Symptom** instead. **Congenital Abnormality** and **Finding**, with the category inherited exclusion, are assigned to *Labial hypoplasia*. **Finding** was considered a wrong assignment by the auditor. **Finding** and **Sign or Symptom**, with the category redundancy exclusion, are assigned to the concept C2711130 *Subungual swelling*. The redundant assignment of **Finding** was deemed to be wrong by the auditor.

The other ten of the 16 rows in Table 4.6 are cases of "implicit exclusion." The entries for these rows assume that *adviseEditor* would have been applied *before* the first concept was assigned such a pair when creating the UMLS 2011AA release. However, *after* creating the UMLS 2011AA release the system would have returned "more research required" instead, since in this release such semantic type pairs were already assigned to one or a few concepts (according to the column #cpts). For the purpose of evaluating the *adviseEditor* system, it is assumed that the UMLS editors would have used it when preparing the UMLS 2011AA release.

Table 4.6 New Pairs of Non-chemical Semantic Types with Few (1 to 5) Concepts in 2011AA

Line	Semantic Type A	Semantic Type B	Rule-category	# of cpts	Appeared in prev. UMLS release?
1	**Acquired Abnormality**	**Finding**	Inherited Exclusion	1	2007AC
2	**Body Part, Organ, or Organ Component**	**Substance**	Implicit Exclusion	1	No
3	**Body Substance**	**Pharmacologic Substance**	Explicit Exclusion	1	No
4	**Congenital Abnormality**	**Finding**	Inherited Exclusion	2	2007AC
5**	**Clinical Attribute**	**Finding**	Implicit Exclusion	2	2007AC
6	**Diagnostic Procedure**	**Finding**	Implicit Exclusion	1	No
7	**Disease or Syndrome**	**Finding**	Inherited Exclusion	3	2008AB
8**	**Finding**	**Health Care Activity**	Implicit Exclusion	1	2008AA
9	**Finding**	**Injury or Poisoning**	Implicit Exclusion	2	No
10	**Finding**	**Pathologic Function**	Explicit Exclusion	2	2007AC
11	**Finding**	**Sign or Symptom**	Redundancy Exclusion	1	No
12	**Population Group**	**Mental or Behavioral Dysfunction**	Implicit Exclusion	1	No
13**	**Pharmacologic Substance**	**Plant**	Implicit Exclusion	1	2008AA
14**	**Functional Concept**	**Spatial Concept**	Implicit Exclusion	1	2008AA
15**	**Functional Concept**	**Therapeutic or Preventive Procedure**	Implicit Exclusion	1	2007AC
16	**Bacterium**	**Virus**	Implicit Exclusion	1	No

When the very first assignment of each one of the ten pairs of semantic types to a concept was attempted, "implicit exclusion" would have been the result of *adviseEditor*, which is what appears in Table 4.6.This assignment would only be allowed with an extra level of approval by a senior editor or a team of editors (as will be suggested in Section 4.5). As can be seen below, the auditor would have approved only a few of those pairs, preventing the creation of wrong semantic type assignments. Whenever such a pair would have been approved for one concept, the result of *adviseEditor* would have changed to "more research required" for this pair, because the UMLS would have this pair assigned to a concept at that point in time. If an auditor presents several concepts with the same pair of semantic types (prohibited by implicit exclusion) for approval, then all these concepts will need to be evaluated by the supervisor or team.

Indeed, looking back at Table 4.6, there were six concepts assigned five new pairs of semantic types marked "implicit exclusion," which had appeared in a previous release, but were removed after an audit. (The line numbers of those five pairs are marked by "**.") Considering the fact that only two of these five pairs were accepted by the auditor as correct, namely (**Pharmacologic Substance; Plant)** and **(Functional Concept; Spatial Concept**), there is a high likelihood that approvals would not have been given by the UMLS editors for the other "**" cases either.

Table 4.7, shows in the first row that 3, 8, 1 and 12 concepts, respectively, were categorized by *adviseEditor* as explicit exclusion, inherited exclusion, redundancy exclusion or implicit exclusion. That is, for these 24 concepts, the assigned pairs were deemed wrong by *adviseEditor*. The auditor agreed with 19 (= 2+8+1+8) (79%) of these recommendations of the system, i.e., that these assignments are not acceptable. The auditor

missed one case of explicit exclusion for the concept *Blood product units | Blood product unit* assigned **Body Substance** and **Pharmacologic Substance**.

For "more research required" the issue is different. In this case the auditor agrees with *adviseEditor* whenever s/he considers the pair as acceptable, because there is already a concept with this assignment in the UMLS. It is important to understand that this is an evaluation of the *adviseEditor* system, and not an evaluation of the UMLS. Thus, "more research required" does not mean that the auditor needs to go and check those previous assignments. As indicated in Table 4.7, 68% of the 41 concepts (28/41) categorized by *adviseEditor* as "more research required" were confirmed by the auditor.

Table 4.7 Results of *AdviseEditor* System and Auditor's Evaluation of the Results of the *AdviseEditor* System

	Explicit Exclusion	**Inherited Exclusion**	**Redundancy Exclusion**	**Implicit Exclusion**	**More Research Required**	**Total**
# of concepts categorized by *adviseEditor*	3	8	1	12	41	65
# of concepts confirmed by auditor	2	8	1	8	28	47
# of concepts not confirmed by auditor	1	0	0	4	13	18

Based on Table 4.7, the performance of the *adviseEditor* system for the given sample was calculated. The calculation used the determination of the auditor as a gold standard.

The accuracy (the proportion of the assessments of the system which are confirmed by the auditor) is (2 + 8 + 1 + 8 + 28)/65 = 47/65 = 0.72. The precision (the ratio of the semantic type assignments reported as correct by the system, as confirmed by the auditor, to all concepts reported as correct by the system) is 28/41 = 0.68. The recall (the ratio of semantic type assignments reported as correct by the system, as confirmed by the auditor, to all correct concepts) is 28/(28 + 1 + 4) = 28/33 = 0.85. The F-measure (harmonic mean) is F = 2 * Recall * Precision / (Recall + Precision) = 2 * 0.85 * 0.68 / (0.85+0.68) = 0.76.

The sample used in this study is too small to establish statistical significance. However, the size of this sample could not be increased, because all 65 relevant concepts from the UMLS 2011AA release are already included, as explained in Section 4.3.7.

4.5 Discussion

It is interesting to note the ratio of explicit versus inherited rules, namely, 181:30 for inclusion rules and 104:71 for exclusion rules, according to Table 4.5. Intuitively, one would expect the number of inherited rules to be larger than the number of explicit rules. The reason for that is that if an explicit rule is stated between the semantic types **X** and **Y**, and if **X** has *m* descendants and **Y** has *n* descendants, then there may be *m***n* inherited rules between descendants of **X** and **Y**.

However, the reality is different. One reason for that is that many explicit rules are stated between semantic types that are leaves in SN, or between semantic types with just one or two descendants. The potential exceptions regarding descendants of **Chemical Viewed Functionally** or between them and descendants of **Chemical Viewed**

Structurally are not listed as inherited, since explicit rules are given in the documentation for these two subhierarchies.

An interesting observation from Figure 4.4 is that areas of inherited exclusion (blue) appear adjacent to areas of explicit exclusion (purple). A similar observation can be made for the corresponding inclusion rules (appearing as green and red). The interpretation of this observation is that the semantic types for which inherited rules hold typically appear after (in the UMLS numbering scheme) the semantic types for which the explicit rules are stated.

For some implicit exclusion rules it is surprising that they were not made explicit. For example, the UMLS/SN definition for **Fish** is: "A cold-blooded aquatic vertebrate characterized by fins and breathing by gills. Included here are fish having either a bony skeleton, such as a perch, or a cartilaginous skeleton, such as a shark, or those lacking a jaw, such as a lamprey or hagfish."

The Linnaean system of classification for animals assumes the exclusiveness of parallel branches. The above definition does not state that fish and mammals are considered exclusive in the animal kingdom tree. Therefore, the **Fish Excluding Mammal Rule** cannot be discerned from the Semantic Network itself. This is a case of specialization of a parent semantic type into several children in the Semantic Network, done with the intention that the extents of all sibling semantic types should be disjoint. In other words, being a sibling implies the existence of an exclusion rule. This pattern is repeated in the taxonomy of life forms. For the semantic types **Vertebrate, Animal** (and **Organism**) it is known from the animal kingdom categories that their children are exclusive.

If any concepts were to have two assignments of semantic types from parallel branches of the part of the Semantic Network that mimics the animal kingdom categorization, then this would be a serious error. In version 2007AC there was one such pair. The two semantic types **Invertebrate** and **Alga** were assigned to 19 concepts, e.g., *Euglena*, *Plankton*, and *Discoplastis spathirhyncha*. This violation has been corrected. Subsequently, these two semantic types were removed from the Semantic Network, and thus no concepts can have those assignments in 2011AA.

Around 2009, the NLM implemented an automatic QA procedure which removes redundant semantic type assignments before each release of the UMLS [58]. Hence, there are in general no more illegal semantic type pairs due to redundant assignments in the UMLS, although *adviseEditor* exposed one case (see Table 4.6).

The evaluation showed a relatively high performance of the *adviseEditor* system, exposing many semantic type assignments in contradiction to UMLS rules. In Section 4.7, the reference standard used was not perfect, but this is not unusual when dealing with human decisions about complex choices.

The use of the described *adviseEditor* system as a mechanism can support the process of assigning semantic types to new concepts added to the UMLS or updated due to integration of a new release of a source terminology. This system can inform UMLS editors concerning whether a specific combination of semantic types is permitted or prohibited, rather than considering the assignment of one semantic type in isolation from other existing assignments. The use of the *adviseEditor* system, categorizing a pair of semantic types as permitted, prohibited, etc., is expected to prevent insertions of new erroneous semantic type assignments, and also to expedite the editors' work. Considering

the shortage of human expert editors for terminologies in general and for the UMLS in particular, expediting the editorial process will free up editors to work on other relevant tasks.

Should the situation arise that a new concept is assigned a pair of semantic types from the implicit exclusion rule-category, then this assignment and the concept itself need to be carefully investigated to determine whether they are valid. It is proposed that a policy be enforced that no "ordinary" editor of the UMLS should be permitted to assign such a pair of semantic types to a concept. Rather, the approval of a supervisor or the vote of a team of editors should be required for such an assignment. If approval is granted, then this pair will be categorized as "more research required," until six concepts have been assigned this combination.

Having the *adviseEditor* system in use by UMLS editors would have warned them concerning the introduction of categorization errors and would have avoided the resource-intensive efforts to correct them. It is especially noteworthy that many of these erroneous combinations of semantic types in Table 4.6 were reintroduced after already having been corrected and removed once before.

Obviously, an assignment of a pair of semantic types violating any of the other categories of exclusion rules will always be denied. As noted in Section 4.4, semantic type assignments that contradict explicit exclusion rules were found in the UMLS. The comparisons of two versions (2007AC and 2009AA) of the UMLS showed encouraging results, in that many of those erroneous assignments had disappeared. However, in 2011AA new problems were introduced. This shows the urgency of using a system such as *adviseEditor* for approving new pairs of semantic types.

Some small intersections, categorized as "more research required" turned out to be legitimate combinations of semantic types. Over time, their extents have increased and may increase further with the addition of new concepts into the UMLS. When there are six concepts assigned such a combination, it will be categorized as "implicit inclusion."

Table 4.8 Large Intersections of Extents

Functionally Viewed Chemical Semantic Type	**Structurally Viewed Chemical Semantic Type**	**Size of Intersection Extents**
Pharmacologic Substance	**Lipid**	1475
Pharmacologic Substance	**Carbohydrate**	2053
Pharmacologic Substance	**Inorganic Chemical**	2096
Pharmacologic Substance	**Nucleic Acid, Nucleoside, or Nucleotide**	2351
Hazardous or Poisonous Substance	**Organic Chemical**	2749
Pharmacologic Substance	**Steroid**	3110
Antibiotic	**Organic Chemical**	3414
Receptor	**Amino Acid, Peptide, or Protein**	4018
Biologically Active Substance	**Organic Chemical**	4321
Indicator, Reagent, or Diagnostic Aid	**Organic Chemical**	4684
Pharmacologic Substance	**Amino Acid, Peptide, or Protein**	6796
Immunologic Factor	**Amino Acid, Peptide, or Protein**	14064
Enzyme	**Amino Acid, Peptide, or Protein**	25250
Biologically Active Substance	**Amino Acid, Peptide, or Protein**	46708
Pharmacologic Substance	**Organic Chemical**	82059

Altogether, there are 199 pairs of semantic types that have been assigned to concepts. The sizes of the intersections of their extents vary from 1 to 82,059. The 15 pairs of semantic types with the largest extent intersections and the numbers of concepts in the intersections of their extents are shown in Table 4.8. These are all intersections with more

than 1300 concepts. Each of these intersections involves one semantic type which is a **Chemical Viewed Functionally** and one semantic type which is a **Chemical Viewed Structurally.** These largest intersections demonstrate the prominence of the family of inclusion rules defined by **Chemical Viewed Structurally** and **Chemical Viewed Functionally** in Section 4.3.2.

The Semantic Network is viewed as an "abstraction network" for the Metathesaurus of the UMLS. In recent years, "abstraction networks" were derived for several other terminologies, e.g. taxonomies for SNOMED CT and NCIt [59, 60, 65], a schema for the Medical Entity Dictionary (MED) of Columbia [94] and the Specialty Chemical Semantic Network for the **Chemical** component of the UMLS Metathesaurus [95]. Chapters 3 and 9 of this dissertation discuss other aspects of abstraction networks.

In summary, the *adviseEditor* system reflects the extensive semantic type knowledge that was implemented in the UMLS over a long period of time by numerous editors. In this way, the *adviseEditor* system may also serve as a channel for making the valuable experience of generations of UMLS editors available to the current and future UMLS staff members.

4.6 Conclusions

In the past, there was no systematic account of all combinations of semantic types that are either supposed to be exclusive or supposed to be inclusive. Rather, this information was distributed throughout definitions and usage notes of semantic types. Furthermore, many exclusion rules were not made explicit, as they were assumed to be "obvious" based on some outside source of information, such as the Linnaean taxonomy of animals.

All such rules have been collected and organized into eight rule-categories. Those rule-categories are implemented in the *adviseEditor* system that categorizes pairs, triples, quadruples and quintuples of semantic types, and the numbers of members for each rule-category have been computed.

Many interesting cases of the 8778 possible combinations of pairs of semantic types were discussed. Furthermore, examples of concepts that violate the given exclusion rules were presented. Some of those erroneous semantic type assignments to concepts were introduced only recently. Hopefully, the presented *adviseEditor* system will be used in the future when extending the UMLS with new concepts, to avoid the introduction of such invalid semantic type assignments.

The work described in this chapter has been published in the Journal of Biomedical Informatics [46]. In the next chapter, a Web-based tool that implements the adviseEditor algorithm will be presented.

CHAPTER 5

ADVISEEDITOR – A UMLS SEMANTIC TYPES ASSIGNMENT ADVISER

5.1 Introduction

In Chapter 4, a rule-based support system for multiple UMLS semantic type assignments was introduced. In this chapter, a Web-based tool AdviseEditor is presented that was designed and developed to help UMLS editors to determine the legitimacy of a combination of semantic types for a new concept. AdviseEditor can be used in interactive mode for single concepts or in batch mode for many concepts. The interactive utility supports instant determination whether a tuple (e.g. a triple) of semantic types to be assigned to a concept is permitted or prohibited. The batch processing utility supports processing of a file with many concepts.

5.2 System Design

The Unified Medical Language System contains over 2.9 million concepts derived from more than 170 source terminologies in version 2013AA. Its Semantic Network provides a compact semantic abstraction layer with 133 broad categories called semantic types. A concept in the UMLS is always assigned one or more semantic types. However, a number of problems may occur when two or more semantic types are assigned to the same concept. One semantic type assignment may contradict another one, indicating an inconsistency in the semantic type assignments. For example, in the 2011AA release of the UMLS, C1976001 *Blood product units | Blood product unit* is assigned the semantic types **Body Substance** and **Pharmacologic Substance**. According to the usage note of

Pharmacologic Substance, the semantic type **Body Substance** should not be assigned together with **Pharmacologic Substance**. Thus, this assignment is erroneous and should not exist. AdviseEditor was developed to prevent erroneous semantic type assignment before NLM releases a new version of the UMLS. In Chapter 4, eight rule-categories for multiple semantic type assignments were identified, namely Redundancy Exclusion, Explicit Exclusion, Explicit Inclusion, Inherited Exclusion, Inherited Inclusion, Implicit Exclusion, Implicit Inclusion, and More Research Required. For each rule-category, AdviseEditor returns a corresponding suggestion, e.g. "Prohibited" for "Redundancy Exclusion," "Explicit Exclusion" and "Inherited Exclusion," "Permitted" for "Explicit Inclusion" and "Inherited Inclusion," etc.

5.3 Functionality of the System

In the home page of AdviseEditor, a user can choose amongst five sub-interfaces: interactive utilities for two, three, four, and five semantic types, respectively and a batch processing utility. No legitimate case of more than five semantic types for one concept has ever been observed. In the interactive utilities, the user can choose semantic types from drop down menus and AdviseEditor will return the rule-category that applies to this combination (see Figure 5.1). In the batch processing utility, the user can process a file of concepts and their semantic type assignments and view the output of rule categories for all of them (see Figure 5.2). The current AdviseEditor system can be accessed at http://nat.njit.edu/NATServlet/. In future work, it is planned to extend the batch processing utility to support import of concepts in different file formats and export of the AdviseEditor

output as XML file. The work described in this chapter was published in the Journal of Biomedical Informatics [46].

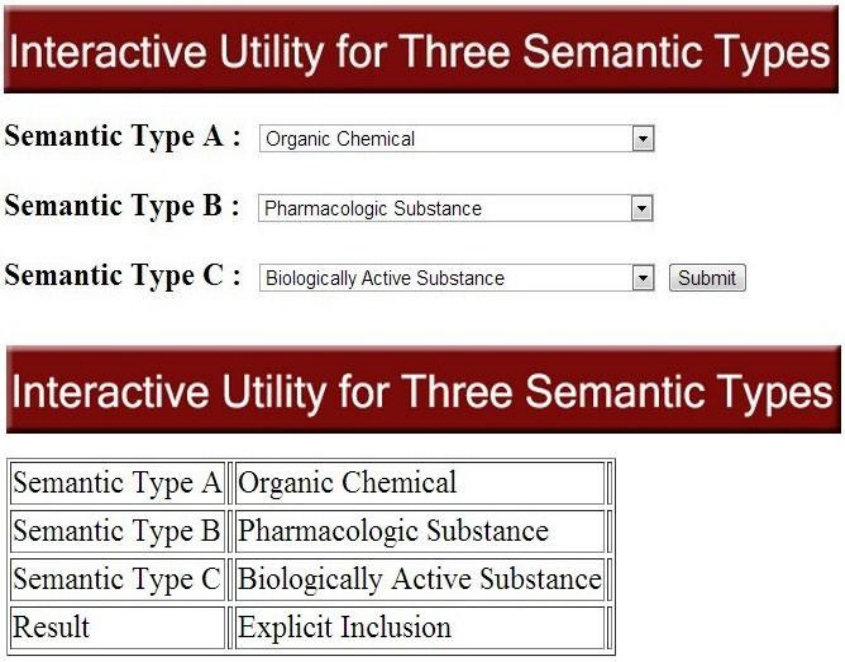

Figure 5.1 Sample input and output of the interactive utility.

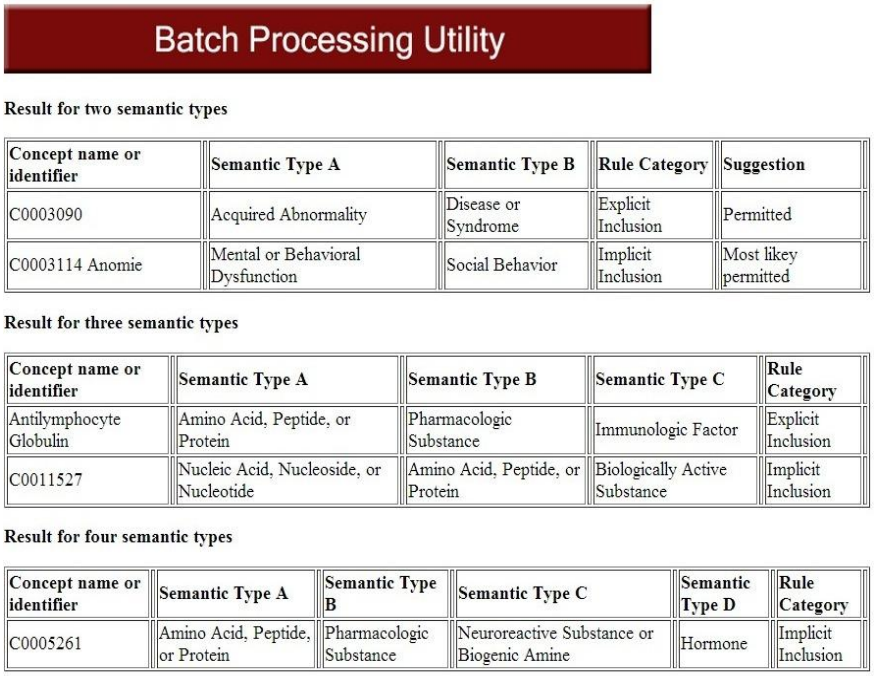

Figure 5.2 Sample output of the batch processing utility.

CHAPTER 6

THE READINESS OF SNOMED CT CONCEPT DESCRIPTORS FOR PRIMARY CARE

6.1 Introduction

Concept descriptors are important in promoting the use of controlled medical terminologies. Among these descriptors, synonyms are particularly important, as indicated by Chute et al. [96]. Synonyms may be even more important when it comes to interface terminologies. In fact, Rosenbloom et al. [97] speculate that one of the cornerstones for usability of clinical interface terminologies is the adequacy of synonymy. Not only is the extent of synonym coverage important, but so is the depth. Medical concepts are often referred to using numerous names, acronyms, and various levels of local variation. While SNOMED CT has emerged internationally as a leading terminology, it surprisingly has a relative paucity of synonyms. Of course, a reference terminology is not necessarily expected to include all synonyms, but only 36% of SNOMED CT's concepts have assigned synonyms, for an average of 0.51 synonyms per concept (103,996 out of a total of 291,205, January 2010 release). In a recent survey [39], more than half of the SNOMED CT users responding indicated that expanding synonym coverage is important to them. Missing synonyms were reported as the second most encountered deficiency in SNOMED CT (after missing concepts) by 17% of the respondents.

Making synonym adequacy more critical is the fact that the HITECH regulations [40, 41] and the "meaningful use" initiative portend nearly exponential growth in the adoption of Electronic Health Record (EHR) systems in the near future [1, 41]. In fact, SNOMED CT is slated to become the exclusive encoding system for problem lists by 2015

[1]. This puts SNOMED CT front and center, and a much wider range of users is expected to interact with SNOMED CT-based content in clinical applications. Such users will expect correct and appropriate synonyms to allow for ease of differentiation between similarly worded concepts in order to efficiently select the clinical concepts that best apply to their patients. Errors in synonyms, lack of synonyms, or insufficient concept information to decipher the exact meaning of concepts' descriptors may prove detrimental to widespread clinical adoption.

In the integration of SNOMED CT into the UMLS, there were numerous cases where two or more SNOMED CT concepts were mapped to the same UMLS concept [98]. Specifically, this happened for 13.4% of SNOMED CT's concepts. Fung et al. [98] also highlight the fact that the methodology of synonym integration may inadvertently increase ambiguity. While Fung et al. [98] provide the reasoning for such occurrences, they did not systematically explore synonyms within SNOMED CT itself. This further raises questions about whether SNOMED CT concept descriptors offer sufficient information for effective clinical differentiation.

In this chapter, an evaluation of concept descriptor issues across SNOMED CT from a practical use perspective is attempted. Four random samples from different SNOMED CT concept populations are utilized in the study. Of particular interest are SNOMED CT concept pairs mapped into UMLS concepts due to shared term patterns. A simulated clinical scenario involving various term-based searches for concepts was used to assess whether SNOMED CT's synonyms and other descriptors provide sufficient differentiation to enable concept selection between similar concepts.

6.2 Background

Each SNOMED CT concept has (i) a fully specified name (FSN) that includes the semantic tag in parentheses, e.g., **hematoma (morphologic abnormality)** (in this chapter, concept names are denoted by bold typeset), and (ii) a preferred term (PT) (e.g., **hematoma**). In many instances, the FSN and the PT are identical except for the semantic tag, which captures the semantic category to which the concept belongs. PTs are meant to capture the common word or phrase used by clinicians to name concepts [99].

Occasionally, SNOMED CT concepts may be accompanied by one or more synonyms. Synonymous terms are intended to convey identical or nearly identical meaning [100], assuming similar semantics of certain words. In SNOMED CT, synonyms are acceptable alternatives to the preferred terms, and both are not necessarily unique [99]. Acronyms are also considered synonymous terms in SNOMED CT. For example, **COPD** and **COLD** are among the 15 synonyms of the concept **chronic obstructive lung disease (disorder)**. SNOMED CT claims to include a large number of synonyms that provide flexibility of expression [99, 101]. On top of the included synonyms, SNOMED CT also offers a "word equivalent" table as part of its Developer Toolkit. This table supports enhanced searches that take into account semantically similar words and provides commonly used abbreviations without greatly increasing the volume of synonyms [102]. Thus, SNOMED CT strives to create a practical balance between synonym explosion on one hand and limited expressivity on the other hand.

In prior research to identify whether the UMLS is a reliable source for enhanced SNOMED CT synonymy, particularly regarding concepts covered by the NLM's published problem lists [103, 104], there were many cases where problematic synonyms in

the UMLS were associated with instances where two SNOMED CT concepts were mapped to the same UMLS concept. For example, **Ectopic beats** and **Premature beats** are two distinct SNOMED CT concepts that are both mapped to the UMLS concept **Premature cardiac complex**. This is a known issue; as discussed by Fung et al. [98], the incorporation of SNOMED CT into the UMLS resulted in numerous instances of more than one SNOMED CT concept being mapped to the same UMLS concept. Several reasons are attributed to such occurrences [98]: (a) strict separation of hierarchies in SNOMED CT results in very similar concepts residing under different roots, (b) fine granularity in SNOMED CT, (c) ‘‘NOS’’ (“Not Otherwise Specified”) concepts, and (d) cases of missed synonymy in SNOMED CT. As an example of (a), concepts with the SNOMED CT semantic tags {disorder} and {morphologic abnormality} may be considered synonymous by the UMLS. Clear errors that were detected during the editorial process by UMLS staff were reported to the editors of SNOMED CT. Although, as noted, the causes of most of these occurrences were explained in [98], they may still present a problem from a clinical use perspective, especially considering the size and fine granularity of SNOMED CT.

6.3 Methods

A simulated clinical scenario was used to assess whether SNOMED CT’s concept descriptors (especially its synonyms) provide sufficient differentiation to enable possible concept selection between similar terms. The evaluation was carried out with respect to single concepts or pairs of concepts within four randomly selected samples, described below. The scenario involves a clinical user performing a series of term-based searches for clinical content and being provided in the process with choices of concepts, displayed with

the most closely matched PT or synonym according to the physician's search term and the application's search algorithm. The search mechanism of SNOMED CT's CliniClue browser [105] was utilized as the search tool. The functionality of CliniClue is similar to other acceptable standalone search tools or search mechanisms within clinical applications which may or may not use subsets of SNOMED CT. CliniClue offers several search options and the default "Words – any order" option was used without any constraints, together with the "Flat list" results display option. Exact string matches are displayed at the top of the returned subset.

The user was instructed to evaluate no more than the topmost twenty items of any returned search results and to focus on exact matches. For example, there exist two aspirin concepts, **aspirin (product)** and **aspirin (substance)**. If the user were to search by typing "aspirin" into the search tool, the highest ranking results would be these two seemingly identical **aspirin** concepts. In CliniClue, and most likely in any built-in search tool within clinical applications, search results are displayed without their respective semantic tags (e.g., {product} and {substance} shown for the **aspirin** concepts). Without additional information, the hypothetical user would have difficulty discerning which of the concepts is appropriate for his clinical need.

The degree of difficulty that a user may face in making such a decision about whether a concept resulting from a search is appropriate for clinical use was quantified. The analysis was performed even when the concepts were presented with their FSNs, which include the semantic tags (e.g., {finding}, {morphologic abnormality}, etc.). The evaluation took into consideration SNOMED CT's principle that PTs and synonyms are not required to be unique. A four-point Likert scale was used, where 0 indicates a

non-issue, 1 indicates a minimal issue, 2 indicates a moderate issue, and 3 indicates a significant/critical issue. In light of typically scarce terminology auditing resources, the evaluation involved a single auditor. To minimize the subjectivity of the evaluation, the results of the four-point scale were converted into a yes/no decision where Grades 0–2 are considered a "no (issue)" and Grade 3 is a "yes." Thus, for example, the synonyms **Arteriovenous catheterization** and **Arteriovenous cannulation** were marked as Grade 3 because they were assigned to the concept **Direct arteriovenous anastomosis**.

In accordance with Fung et al. [98], four data sets were defined. Sample *A* ("Same String Pairs" – "SSP") consists of 65 pairs of SNOMED CT concepts such that the concepts of each pair are mapped to the same UMLS concept and share an identical string across their synonyms and/or their PTs. For example, two SNOMED CT concepts **repair of penis {procedure}** (Concept ID: 81474006) and **balanoplasty {procedure}** (Concept ID: 307240001) are mapped to the same UMLS concept **Repair of penis (procedure)** (UMLS CUI: C1094740). The concept **repair of penis {procedure}** has a synonym **balanoplasty**. Thus, these two concepts share the same string "balanoplasty."

Sample *B* ("No Shared String Pairs" – "NoSSP") comprises 81 concept pairs where each member of a pair is again mapped to the same UMLS concept, but in this case the pair members have completely different strings from one another across their synonyms and their PTs. For example, two SNOMED CT concepts **memory impairment {finding}** (Concept ID: 386807006) and **amnesia {finding}** (Concept ID: 48167000) are mapped to the same UMLS concept **Amnesia** (UMLS CUI: C0002622). These two SNOMED CT concepts do not have any PTs and synonyms with the same string.

Sample *C* ("Synonym Control" – "SynCtrl") consists of 50 individual SNOMED CT concepts with at least one synonym that does not share a UMLS concept with any other SNOMED CT concept. Sample *D* ("Ctrl") is made up of 100 individual SNOMED CT concepts without regard to their number of synonyms.

Table 6.1 General Synonym Characteristics in SNOMED CT and the Concept Samples

	General SCT	**Concepts with Synonyms**	**Sample *A* (SSP)**	**Sample *B* (NoSSP)**	**Sample *C* (SynCtrl)**	**Sample *D* (Ctrl)**
# of concepts	291,205	103,996	130	162	50	100
% concepts w/synonyms	35.7%	100%	68.5%	50.6%	100%	31%
Avg # of synonyms	0.51	1.42	1.39	1.22	2.80	0.51
Avg # of synonyms for concepts w/synonyms	1.42	1.42	2.05	2.40	2.80	1.65
Min / max # of synonyms	0 / 27	1 / 27	0 / 7	0 / 8	2 / 8	0 / 5

Table 6.2 Grade 3 Findings across the Four Samples

	Sample *A* (SSP)	**Sample *B* (NoSSP)**	**Sample *C* (SynCtrl)**	**Sample *D* (Ctrl)**
#	65 (pairs)	81 (pairs)	50	100
Grade 3 Issues	40	14	1	1
% Grade 3 Issues	62%	17%	2%	1%
Synonym Errors	7	–	1	–
Duplicate Concepts	8	7	–	–
Container Classes	11	3	–	–
Other	14	4	–	1

The four randomly selected data sets used in the study were derived from the January 2010 release of SNOMED CT. All samples were chosen to be mutually disjoint, i.e., no concept appears in more than one of them. Excluded from Samples *A* and *B* are concept pairs with FSNs that appear identical, but with one member having the SNOMED CT semantic tag {substance} and the other having {product}, or one having {disorder} and

the other {morphologic abnormality}. This restriction is due to the common occurrence of this kind of situation. An example is **aspirin (product)** and **aspirin (substance)**, both of which share "aspirin" as their PT. Many such pairs can be disambiguated by using well-curated subsets. If such pairs were allowed to dominate Samples *A* and *B*, they might mask other potential issues.

6.4 Results

All evaluations of the four samples were conducted by Dr. Gai Elhanan, a medical informaticist experienced in curation and auditing of large terminologies. Table 6.1 provides general information regarding the synonym content of the concepts in the samples compared to the general population of SNOMED CT concepts. For example, in the general concept population, there are 291,205 (active) concepts (Column 1). Among these, 35.7% have synonyms, with an overall average of 0.51 synonyms per concept. Those concepts having synonyms have an average of 1.42 of them. The concept with the most synonyms has 27 synonyms. For Sample *A* (comprising 65 pairs or 130 concepts), 68.5% of the concepts were found to have synonyms, with an overall average of 1.39 synonyms per concept. The average is 2.05 synonyms for those concepts with synonyms. The maximum number of synonyms for a concept is seven.

Table 6.2 summarizes the findings with respect to each sample. As discussed above, only Grade 3 findings ("significant or critical issues") are displayed. Overall, 442 unique SNOMED CT concepts were evaluated (146 concept pairs, 150 individual concepts). As can be seen in the Table 6.2, 62% (40) of Sample *A* concept pairs were deemed to harbor significant issues (Grade 3): synonym errors, duplicate concepts,

"container classes" (i.e., concepts that are too general), and other issues. In seven pairs, at least one of the concepts was found to contain an erroneous synonym. For instance, balanoplasty is a surgical repair of the glans penis. Therefore, it is an incorrect synonym for the concept **repair of penis (procedure)**, a general concept representing any repair on any part of the penis. Thus, if users were to search for "balanoplasty" in a SNOMED CT-based clinical application, they would be faced with two "balanoplasty" options: (a) **Repair of penis**, and (b) **Balanoplasty**. Without further querying of SNOMED CT's content, they would not be able to differentiate between the two and may select a concept that does not correctly describe the circumstances of their patient. In eight pairs, the concepts were deemed to be duplicates. For example, the concept **oxygen nasal cannula (physical object)** and the concept **nasal oxygen catheter, device (physical object**) co-exist, with the latter having the synonym "nasal oxygen cannula."

In 11 of the pairs, issues resulted from the fact that one or both of the involved concepts were container classes that serve to group together and subsume collections of more refined, sibling concepts. More specifically, one of the concepts was more general than the other, yet shared a synonym. As an example, **Family Megapodiidae (organism)** is the parent of **megapode (organism)**, but the former has the synonym "megapode." Fourteen other concepts, although they did not contain any of the above described issues, still lacked sufficient clarity to resolve potential clinical confusion. For example, a search for "tachycardia" returned two concepts, **tachycardia** as a {disorder} and **tachycardia** as a {finding}. The fine differentiation between a finding and a disorder may escape the common user.

For Sample *B*, 17% of the pairs exhibited Grade 3 issues. Seven pairs were considered duplicates; three showed container-class issues; and four others still resulted in potential confusion. Samples *C* and *D* each had only one concept considered to exhibit a Grade 3 issue.

The differences in the numbers of Grade 3 problems between Sample *A* and Sample *B*, and between Samples *A* or *B* and Samples *C* or *D* were all statistically significant (Fisher's Exact Test, 2-Tail p-value < 0.001 for all).

6.5 Discussion

The findings of this chapter indicate that specific subsets of SNOMED CT concepts may exhibit significant synonym issues. However, the general population of SNOMED CT concepts with synonyms (35.7%) carries a relatively low rate (2%) of major issues with the overall quality of its synonyms. This finding is not in contradiction with the opinions collected in a recent survey of SNOMED CT users [39], where most of the issues raised were with missing synonyms and lack of synonyms, and not necessarily about erroneous ones. It should also be remembered that the relative paucity of SNOMED CT synonyms contributes towards this low rate and that the samples intentionally excluded most issues that may arise from the strict separation of hierarchies in SNOMED CT. However, when a specific population of concepts was examined, i.e., concepts that were deemed similar enough to be mapped to the same UMLS concept, a significantly higher rate of issues could be found. This subset (13.4%; see [98]) of SNOMED CT concepts is not negligible and deserves closer scrutiny. Such issues may lead users of SNOMED CT-based clinical applications to erroneously select a concept that does not necessarily apply to their patient.

This, of course, may lead to subsequent errors by medical personnel and incorrect application of decision support and analytical tools.

From IHTSDO's perspective, most of these issues, in all likelihood, do not represent true problems. SNOMED CT's 19 mutually exclusive hierarchies and its fine granularity virtually guarantee that strings with similar or identical word structure with different semantics will reside under different roots. Indeed, SNOMED CT's User Guide [99] explicitly indicates that synonyms and PTs are not necessarily unique. As a result, the vast majority of SNOMED CT concepts with two mappings in the UMLS fall under such a category. For example, almost all drug names exist separately as {substance} and {product} concepts in SNOMED CT and correspondingly are almost invariably mapped to the same concept in the UMLS.

As much as this arrangement is logical within SNOMED CT's structure, it may present significant difficulties to the average user within clinical applications and even to software vendors. SNOMED CT is no longer considered the product of an "academic exercise." Due to successful leadership and adoption initiatives, SNOMED CT has already passed the tipping point of clinical adoption. The accelerating adoption of EHRs and the regulatory emphasis on standardized encoding of clinical problems within such applications [1, 40, 41] will lead to an increased exposure of novice users to SNOMED CT, especially in primary care settings. These users cannot be expected to know the inherent structure and underlying logical modeling of such a terminology, and will be oblivious to many of the finer principles described in the SNOMED CT User Guide. Nor can it be assumed that such users have the desire to use terminological tools to discern the differences between the meanings of SNOMED CT's concepts *prima facie*.

Institutions like Kaiser Permanente (KP) and the Veterans Administration (VA) spent years and significant financial resources to reach the point where they can utilize aspects of SNOMED CT for their clinical needs [106, 107]. And while large EHR vendors can possibly match such an effort, most vendors of the approximately 1000 currently certified complete EHRs [108] cannot reasonably muster such an effort. Many of the findings and examples presented in this study involve hierarchies that can be expected to be commonly used (diagnoses/findings, procedures). In this work, readily identifiable issues such as **aspirin (product)** and **aspirin (substance)** are excluded. Such issues can easily be dealt with using well-defined subsets of SNOMED CT. However, even with the use of limited subsets such as the CORE [103] and VA/KP [104] problem list subsets of SNOMED CT, or commercially available well-curated subsets, the described problems can still be expected to present themselves.

A scenario is presented where a community physician wishes to record the fact that a patient is undergoing chemotherapy. Intuitively, a user will type "chemotherapy" into a hypothetical search tool within the EHR that relies on SNOMED CT terms. Using CliniClue [105], for example, the two top-most terms returned are both synonyms named chemotherapy, each related to two different concepts: (1) **antineoplastic chemotherapy regimen (procedure)**, and (2) **administration of antineoplastic agent (procedure)**. Clearly, there are more than subtle differences between these two concepts. Since both concepts belong to the same hierarchy, limiting the search to a specific subset is not likely to eliminate the confusion. How is a hypothetical user to select the correct one if all s/he is presented with are two identical strings, one a PT and the other a synonym? Should only the PT be presented to her/him, or the FSN, or all of them? Obviously, there is no simple

answer, but in this case, even with exposure to all the available information, the decision might be difficult and frustrating, and may require dwelling on the finer details of SNOMED CT's conceptual representations.

Another observation is related to the way that SNOMED CT uses container-class concepts clearly created to subsume a group of other concepts under the "same roof." As an example, **cow's milk specific immunoglobulin E antibody measurement (procedure)** is a child of **milk specific IgE antibody measurement (procedure)**. Each of the two concepts has at least one synonym indicating that they are related to cow's milk RAST (lgG radioallergosorbent test). However, a closer examination reveals that for the parent, this is an error since goat and sheep milk RAST tests are children as well. Although this can be considered simply a synonym error, the phenomenon observed here and in other cases is, most likely, that a concept that was formerly a leaf node became a container class. Such instances can be algorithmically detected and avoided altogether by a disciplined editorial approach. It is proposed that IHTSDO formulate special editorial rules for container classes, especially for ones that are not specific enough to be used clinically, and thus may not require synonyms. Unintended use of higher-level concepts can lead to reasoning mistakes by algorithmic decision-support systems.

Such scenarios were hypothesized when the samples were evaluated. While most of the findings of this research are not likely to be recognized by IHTSDO (except for potential duplicate concepts or erroneous synonyms) they may confound everyday clinical users. While not knowing how often such issues may arise under different clinical settings, the expanded role of SNOMED CT subsets suggests that the identified issues should be systematically addressed for better encoding and wider clinical adoption.

Aside from obvious errors that should be corrected, the most plausible mechanism that SNOMED CT offers to deal with such issues is the local extension [99]. However, the extension mechanism requires a resource intensive, coordinated effort [109, 110], most likely, on a national level, and may still not resolve the majority of issues. If a hypothetical physician were to record the gender of a female patient by typing the term "female" into CliniClue, s/he would be presented with both "female" as **female structure (body structure)** and "female" as **female (finding)**. Such complexity is by design and is not likely to be resolved by local extensions. Similar situations are presented where the involved concepts carry the semantic tags of {finding} and {disorder}, the semantics of which is essential for problem lists. The selection made under clinical circumstances carries with it significance beyond the common meaning of the string that represents it. Each concept has a different conceptual representation, and future reasoning engines may be compromised and draw different conclusions due to hasty selections made under sub-optimal conditions.

Although SNOMED CT is a reference terminology and is not expected by IHTSDO to serve as an interface terminology *per se*, many others have attempted to utilize it that way. The dangers associated with the ambiguities described above should be addressed. However, the prospect of creating a dedicated, clinically specific extension that addresses such issues, as well as many others—within a practical timeframe—is not promising although some issues, such as the use of container class concepts can be addressed algorithmically. Some of the issues highlighted in this study demonstrate a schism that already exists within SNOMED CT between reference and interface uses.

Therefore, the complexity of a reference terminology, such as SNOMED CT, should be balanced against its clinical usefulness during the creation and editing process.

This study is qualitative, with a subjective aspect associated with the simulation and review process by a single expert. For that reason, only Grade 3 findings were exposed. As the examples above show, Grade 3 findings were non-arbitrary, clear-cut issues. However, medical decision making is often subjective as well, and it has been the experience, over many years of providing feedback to both the UMLS and the SNOMED CT governing bodies, that the error correction and content introduction process is not entirely objective and structured. Nevertheless, the findings of this dissertation exposed aspects of SNOMED CT usability that were not considered before. Future work is required to systematically address and eliminate such confounding scenarios.

Our selection of the CliniClue search mechanism, although arbitrary, represents a reasonable approach and may affect only some types of errors, while others are independent of the search-and-display algorithm. Even though other search algorithms may display search results in a different manner, CliniClue is the prominent tool to view and investigate SNOMED CT [39] and offers a practical and satisfactory solution. It is unlikely that many of the vendors of the more than 1000 currently certified complete EHRs will offer significantly better tools to explore medical ontologies.

A PubMed search reveals that the literature related to auditing of SNOMED CT synonyms is scant, with only two immediately relevant studies [98, 101]. Despite historical claims [99, 101] the overall paucity of synonyms mandates that a significant effort be directed at improving their coverage and depth. This is especially relevant for leaf-node concepts that are more likely to be used in clinical circumstances. This is particularly true

in the short term for the proposed problem lists. Addressing specific subsets of SNOMED CT concepts, such as those covered by this study, can provide a good starting point.

6.6 Conclusions

SNOMED CT exhibits a low overall rate of synonym errors. However, its hierarchical structure and synonym content result in murky areas where non-expert users may find it difficult to choose the correct concepts in clinical settings. In this chapter, a simulated clinical scenario was utilized to demonstrate some of the difficulties that could be encountered and samples of SNOMED CT's conceptual content were evaluated in this regard. While IHTSDO does not consider SNOMED CT as an interface terminology, there is no immediately available alternative. Thus, it is desirable that IHTSDO should pay closer attention to practical clinical use cases and formulate editorial policies to better address practical clinical needs and reduce structural complexity. Clearly marking container class concepts that are not intended for clinical use and possibly removing synonyms from such concepts might serve as a start. In light of SNOMED CT's increasing role in primary care, more attention should be focused on pragmatic usability aspects.

The work described in this chapter has been published in the Proceedings of the 2nd International Workshop on Managing Interoperability and Complexity in Health Systems [47]. In the next chapter, a study to approach the semantic harmonization problem by categorizing the relationships between structurally congruent concepts from pairs of terminologies, with SNOMED CT being one member of every pair, will be presented.

CHAPTER 7

ANALYZING CONGRUENT CONCEPTS FROM PAIRS OF METATHESAURUS TERMINOLOGIES FOR SEMANTIC HARMONIZATION

7.1 Introduction

In this chapter, semantic harmonization is approached by analyzing the relationships between *structurally congruent* concepts from pairs of terminologies in the UMLS. Auditing of terminologies may uncover problems such as omissions [111]. Previously, algorithmic and mixed human-computer auditing methods for the UMLS and some of its source terminologies have been developed [27, 30]. Auditing may also discover concepts that are synonymous in real life but are coded as different in the UMLS. Occasionally two terminologies in overlapping domains "cut the world at different joints," which makes ontology alignment [112] and ontology integration difficult. In such a situation, the same conceptual knowledge may be classified in (often orthogonal) different ways, which are called "alternative classifications." In this chapter, the use of structural congruency in pairs of terminologies is presented to alert a human auditor to possible cases of harmonization and correction. SNOMED CT (abbreviated as "SNOMED") is the focus of this chapter due to the importance of its concepts.

7.2 Background

Bodenreider [113] observed that it is the policy in the UMLS that 'PAR' represents an explicit parent-child relationship in a source, and 'RB' indicates an implied one (as interpreted by the UMLS editorial team). In this chapter, explicit hierarchical relationships

are the focus, thus only terminologies in the UMLS with 'PAR' links annotated with 'INVERSE-IS-A' relationship attributes were chosen.

7.3 Methods

The methods used in this chapter are based on comparing two medical terminologies from the UMLS. The targets of the investigation are formally defined as follows.

Definition: The concepts X (from Terminology 1) and Y (from Terminology 2) are called "structurally congruent" if:

a) Both concepts X and Y have the same parent A in Terminology 1 and in Terminology 2.
b) Both concepts X and Y have the same child B in Terminology 1 and in Terminology 2.
c) The concept X does not appear anywhere in Terminology 2.
d) The concept Y does not appear anywhere in Terminology 1.
e) There is no synonymy relationship and no hierarchical relationship between X and Y (in the UMLS).

Figure 7.1 shows an abstract layout of two structurally congruent concepts to elucidate the above definition. It is hypothesized that there are **six possible cases** for how X and Y may relate to each other.

1) The concepts X and Y are alternative classifications. That means that concept A may be validly assigned X and Y as its children. However, these two assignments are indicative of two different ways of clustering the grandchildren of A. Furthermore, concept B may be

correctly classified as a child of X and as a child of Y. However, Terminology 1 omits the classification by Y and Terminology 2 omits the classification by X. In this and the next chapter the symbol "→" will be used to stand for "IS-A."

2) It holds that B → Y → X → A. In other words, Y may be inserted as a child of X into Terminology 1, thereby adding more detailed information to Terminology 1. Similarly, X may be inserted as a parent of Y into Terminology 2. Such insertions should only be done with approval of a subject matter expert.

3) It holds that B → X → Y → A. This is the mirror case of Case 2) in that now X may be inserted as a child of Y into Terminology 2 and Y may be inserted as a parent of X into Terminology 1.

4) Concept X is a real world synonym of concept Y, which was previously not recognized by the UMLS editors.

5) There might be a structural error in Terminology 1, e.g., X is not really a child of A.

6) There might be a structural error in Terminology 2.

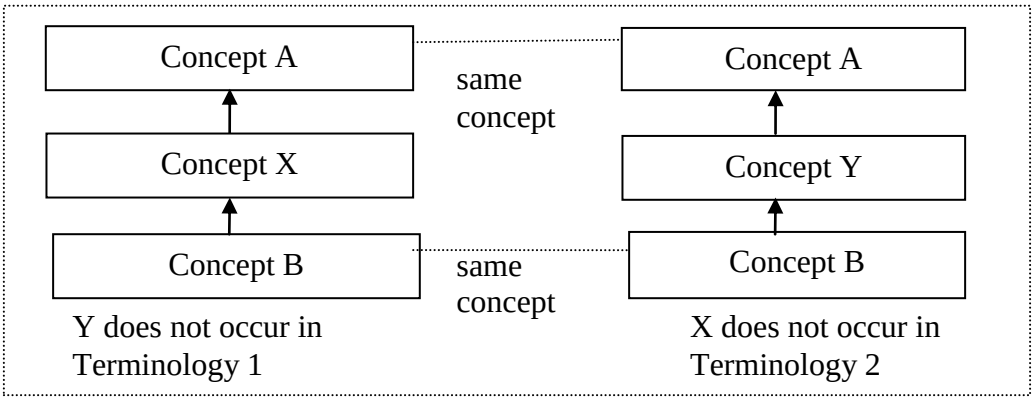

Figure 7.1 An abstract layout of structurally congruent concepts.

Every one of these six cases may be utilized in a human review, possibly leading to an improvement and harmonization of both terminologies. To further probe the potential of this idea, the following study was performed. Six terminologies were selected from the 2012AB release of the UMLS to function as reference terminologies for SNOMED. (Note: It is a *coincidence* that there are six cases and six terminologies.) They are MEDCIN3_2012_07_16, National Cancer Institute Thesaurus (NCI2012_02D), Gene Ontology (GO2012_04_03), Medical Entities Dictionary (CPM2003), UMDNS: product category thesaurus (UMD2012) and Foundational Model of Anatomy Ontology (FMA3_1). Only English-language terminologies using the "PAR" relationship annotated with "IS-A" relationship attributes were chosen. The University of Washington Digital Anatomist (UWDA) was excluded due to its similarity with the FMA3_1 terminology.

Algorithms were implemented for finding all structurally congruent pairs of concepts from pairs of terminologies with one terminology taken from the list of six reference terminologies, the other one being the July 2012 version of SNOMED. The UMLS is well known to contain many cycles [113, 114], which were eliminated during processing.

7.4 Results

Table 7.1 shows the numbers of pairs of congruent concepts of the six reference terminologies relative to SNOMED and the sizes of the samples randomly chosen for human review as follows. The third column shows the number of pairs of congruent concepts found by the program. For reference terminologies with over 100 pairs of congruent concepts, random samples of 50 were chosen for human review; for the other

terminologies, all of the congruent concepts were reviewed. In total, 181 /1384 = 13.1% of all the congruent concept pairs discovered by the program were reviewed.

Table 7.1 Comparison of SNOMED CT with Six Reference Terminologies

Reference Terminology	**Size of Terminology**	**# of Pairs of Congruent Concepts**	**Sample Size**
MEDCIN3_2012_07_16	279529	655	50
NCI2012_02D	95523	582	50
FMA3_1	82062	116	50
UMD2012	15956	18	18
GO2012_04_03	61925	6	6
CPM2003	3078	7	7
Total	--	1384	181

Gai Elhanan, a medical informaticist and MD with many years of experience in auditing terminologies reviewed the sample. Table 7.2 shows the results according to the six cases defined in Section 7.3. The results show that 64.6% are alternative classifications. Another 14.4% + 7.2% = 21.6% fall into the category where the congruent concept in the reference terminology could be imported into SNOMED, and vice versa.

Table 7.2 Human Review Results by Reference Terminology

Reference Terminology	**Sample Size**	**Alterna. Classifi-cation**	**Y IS_A X**	**X IS_A Y**	**Error in Termi-nology 1**	**Error in Termi-nology 2**	**Synonym**
MEDCIN3_ 2012_07_16	50	34	6	3	--	1	6
NCI2012_ 02D	50	33	8	4	--	1	4
GO2012_ 04_03	6	2	--	4	--	--	--
CPM2003	7	5	--	--	--	--	2
UMD2012	18	9	1	--	--	--	8
FMA3_1	50	34	11	2	1	--	2
Total	181	117	26	13	1	2	22
Percentage	100%	64.6%	14.4%	7.2%	0.6%	1.1%	12.2%

Figure 7.2 shows an example where congruent concepts were identified as alternative classifications. Thus, *Eleventh posterior intercostal vein* in the FMA is a classification by ordinality, while in SNOMED *Lower right posterior intercostal vein* is a classification by position.

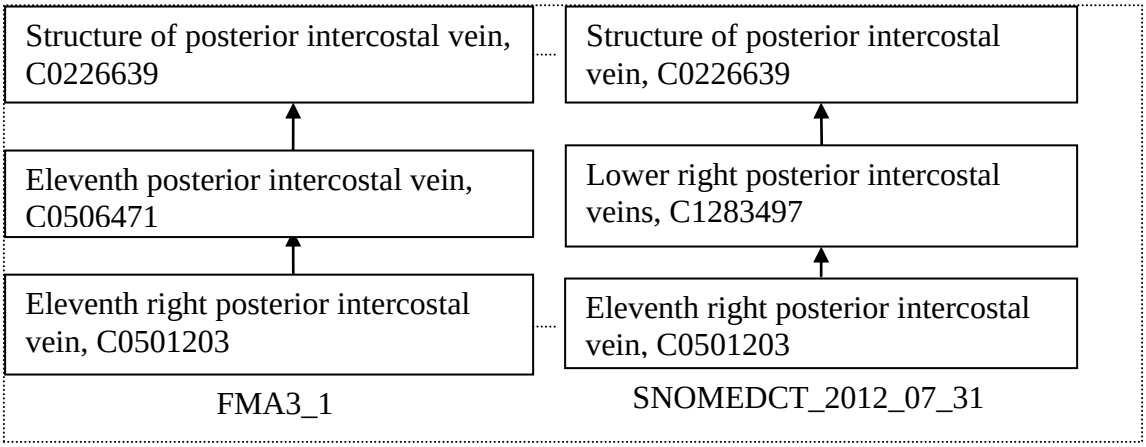

Figure 7.2 An example of alternative classification.

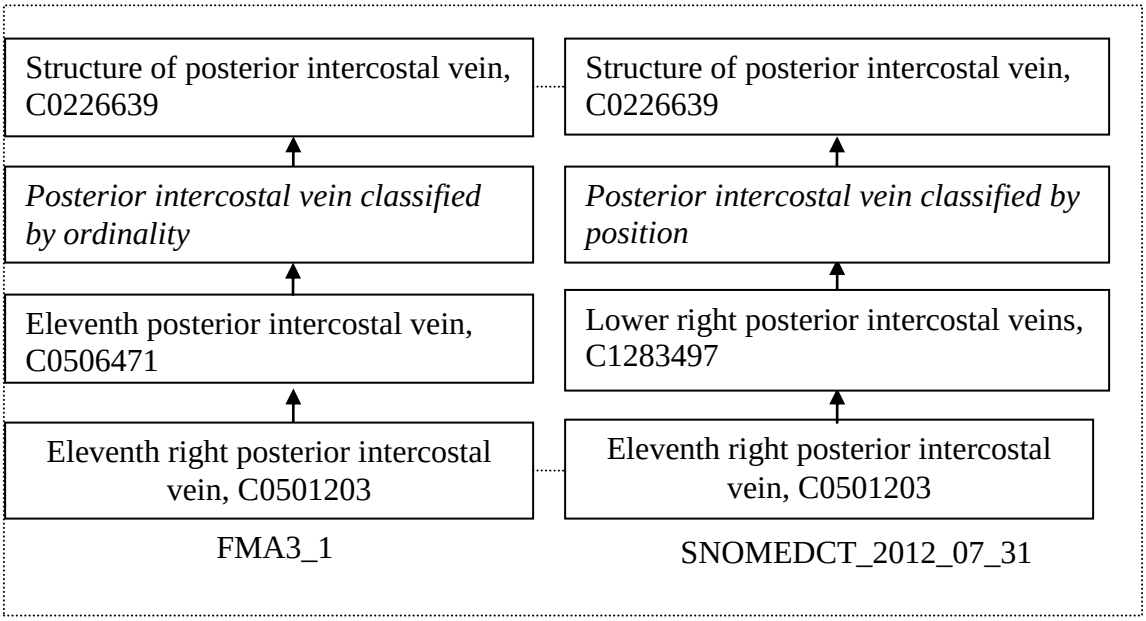

Figure 7.3 An example of making explicit an implicit assumption of the ontology designers.

The discovery of alternative classifications is useful, because it makes explicit the implicit assumptions of the ontology designers how they are viewing the world. This view could then be codified in the ontology if the ontology commonly uses container concepts. Figure 7.3 shows the utilization of the findings in Figure 7.2 by adding two new container concepts (with labels shown in *Italics*.) The curators of both ontologies will need to decide if they want to include one or both alternative views in their ontologies.

Figure 7.4 shows a case where one congruent concept was deemed a parent of the other by the auditor. In this example, the congruent concept *Finding by Site or System* can be a parent of *Finding by site*, thus the congruent concept *Finding by Site or System* from FMA may be added as a parent of *Finding by site* in SNOMED, and vice versa, if this is desirable in the judgment of the owners of the FMA and/or SNOMED. The structure after the import is shown in Figure 7.5.

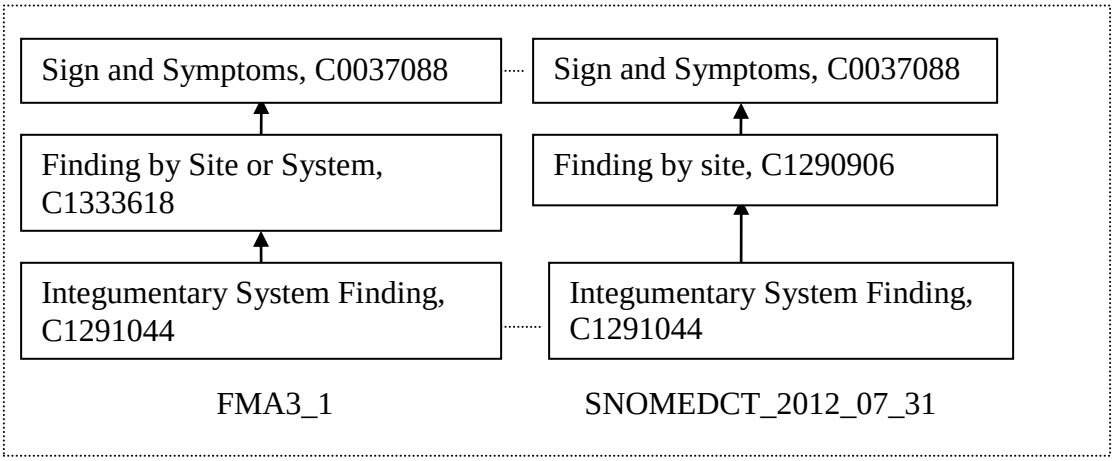

Figure 7.4 An example of one structurally congruent concept being a parent of the other.

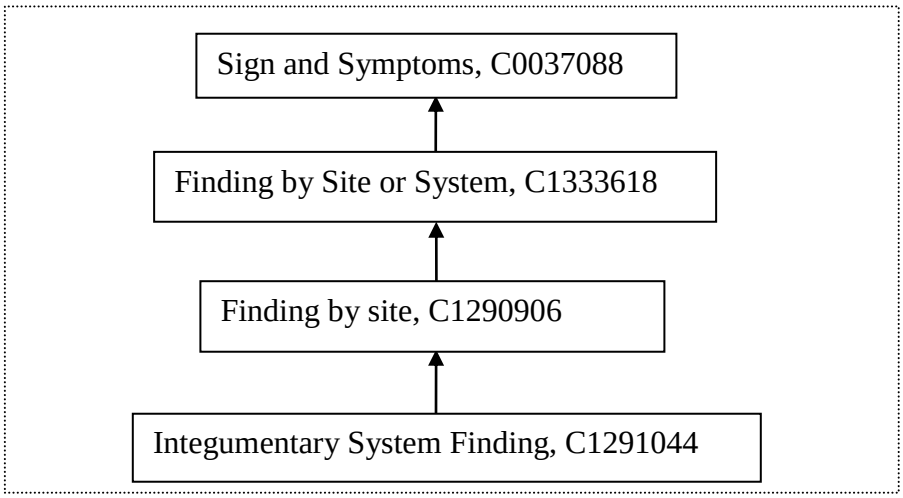

Figure 7.5 An example of importing a structurally congruent concept.

The congruent concepts *Chemical Viewed Structurally* from CPM and *Chemical categorized structurally* from SNOMED are deemed synonyms that were not recognized before by the auditor (Figure 7.6) and should be merged.

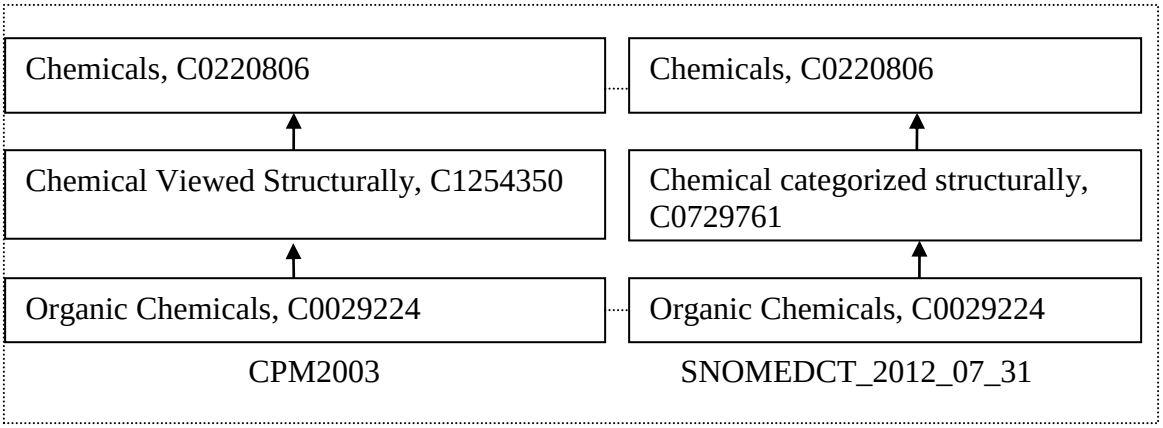

Figure 7.6 An example of one middle concept being a synonym of the other.

During the review of the sample, a few errors within terminologies emerged. The concept from SNOMED *Artificial Implant* was deemed incorrect by the auditor because it should not be considered as "artificial." This structure is shown in Figure 7.7.

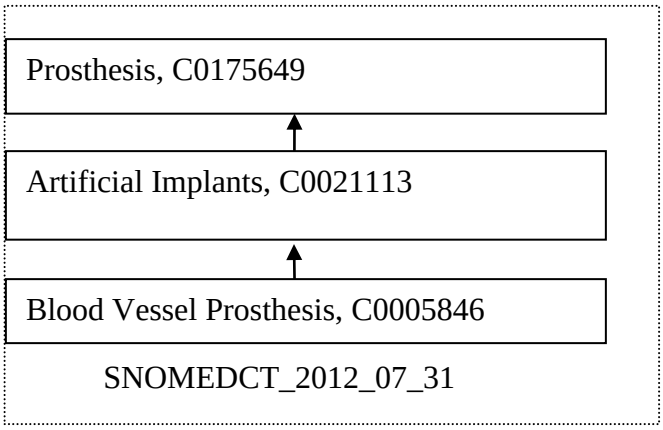

Figure 7.7 An example of an error found in SNOMED CT.

7.5 Discussion

The UMLS provides many concept pairs from different terminologies, where algorithmically made structural observations raise the question how to harmonize those concepts. In this chapter, one such structural observation "structurally congruent concepts" was identified and the different ways how such a congruency can be resolved were indicated. However, the semantic harmonization cannot be done without the consent of terminology curators. Moreover, modeling differences between terminologies make semantic harmonization difficult. For UMD2012 (Table 7.2), eight pairs of congruent concepts were found to be synonyms. For GO, more cases where one congruent concept is a potential parent of the other were found than alternative classifications. For the cases 2) and 3), relevant work in MIREOT [115] defines a set of guidelines for importing classes from external ontologies. However, it only supports OBO foundry ontologies (OWL format). In this paper, all the terminologies are in UMLS RRF format. Thus, the import guidelines introduced in MIREOT cannot be used here directly. A possible limitation of this work is that it uses SNOMED concepts and all reference terminology concepts in the

formats that they were provided in by the UMLS. There may be differences between the original concept representation of SNOMED (or the reference terminologies) and the representation of SNOMED that is accessible through the UMLS.

7.6 Conclusions

Six terminologies of the UMLS were compared with SNOMED with respect to structurally congruent concepts. In a sample study it was found that the great majority of cases corresponded to alternative analysis situations (117 out of 181, corresponding to 64.6%). The second most common situation indicated the possibility of adding more detail to SNOMED CT or the reference terminologies (39 out of 181, corresponding to 21.6%). In 22 cases new synonyms were discovered, and three pairs of concepts indicated errors. The work in this chapter was limited to pairs of structurally congruent concepts. However, there are cases of configuration that involve three, four and even more intermediate path concepts. An analysis of these cases will be presented in the next chapter.

CHAPTER 8

ANALYSIS OF *M:N* TRAPEZOIDS FROM PAIRS OF METATHESAURUS TERMINOLOGIES FOR SEMANTIC HARMONIZATION

8.1 Introduction

In Chapter 7, structurally congruent concepts from pairs of Metathesaurus terminologies were analyzed, with SNOMED CT being one terminology in every pair. Six kinds of configurations were observed, e.g., alternative classification, suggested parents, and suggested synonyms.

This chapter extends the approach from Chapter 7 to configurations with n ($n > 1$) intermediate concepts in one or both of the two terminologies.

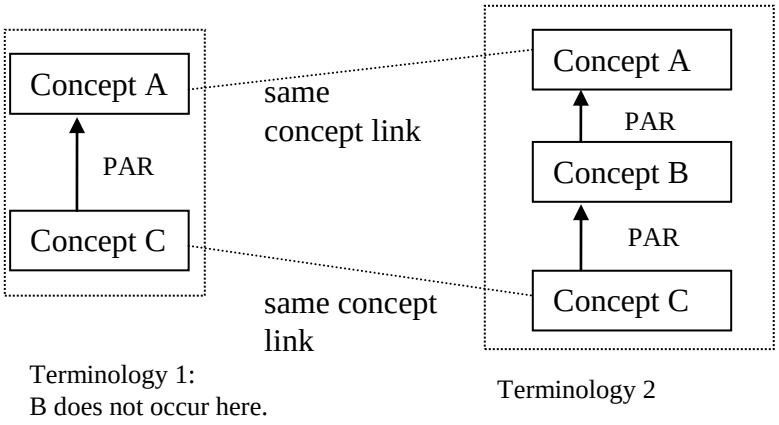

Figure 8.1 The basic layout of a vertical density difference.

Figure 8.1 shows excerpts from two "hypothetical" terminologies. The concepts A and C are assumed to be identical in both of them based on their CUIs. However, Terminology 2 has an additional concept B located on a path of PAR (parent) links from C to A. Note that it is assumed that B does not appear *anywhere* in Terminology 1.

A concrete example following the configuration in Figure 8.1 is shown in Figure 8.2. Figure 8.2 compares small structures from two terminologies, which were both extracted from the Unified Medical Language System's (UMLS) [18, 19, 116, 117] Metathesaurus [20, 21]. Figure 8.2 shows a case where SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [34-37] has an additional concept *Fracture of bone forefoot* between *Fracture of foot* and *Metatarsal bone fracture* compared with MEDCIN. Thus, the ratio of PAR-path lengths is 1:2. Due to the shape defined by the PAR links and the two dotted lines ("same concept") indicating identity of the concepts from two source terminologies in the UMLS, this configuration will be referred to as *1:2 trapezoid* in the balance of the chapter. SNOMED CT is always Terminology 2 and is always displayed in the right "half" of a figure comparing two terminologies, thus *2:1 trapezoids* are also well defined.

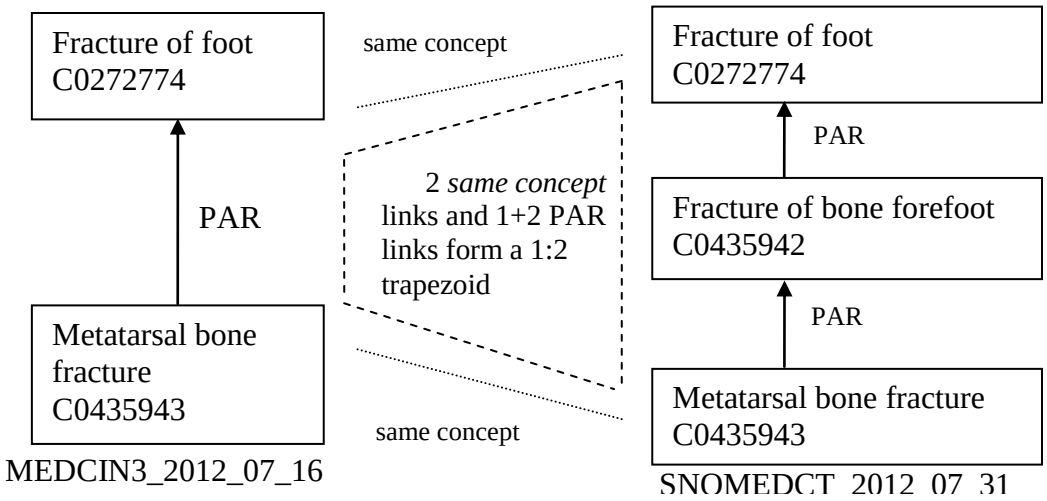

Figure 8.2 An example of a 1:2 trapezoid.

The version of SNOMED CT used is dated July 31, 2012. The distinctions between Terminology 1 and Terminology 2 in Figure 8.2 appear in the vertical (PAR) structures and are interpreted as a *density difference* (as described in Chapter 7) of the two terminologies.

Omissions in terminologies are undesirable, and locating them is one of the goals of work in terminology auditing [111]. In past work, methods have been developed to recognize certain omissions in the UMLS and some of its source terminologies [27, 30, 36]. One may interpret the lack of "Fracture of bone forefoot" (C0435942) from MEDCIN3_2012_07_16 as an undesirable omission, but this will be up to MEDCIN's curators.

Having established the scope and descriptive terminology of this chapter, its objectives are best expressed by three questions:

1) How often does the phenomenon of density differences between medical terminologies occur, limited to the precise configurations described above? In other words, is a density difference an outlier or are they common?

2) How far reaching are density differences between one target terminology and one or several reference terminologies? In other words, are there only 1:2 and 2:1 trapezoids, or are there 1:3, 1:4, 2:3, 2:4 etc. and 3:1, 4:1, 3:2, 4:2 etc. trapezoids?

3) What are the relationships of intermediate path concepts of *m*:*n* trapezoids, where m >= 2, n >= 2? Can these relationships be used to enhance the semantics and coverage of a terminology? Can they contribute to the semantic harmonization of the two source terminologies.

8.2 Methods

Six terminologies, which are the same as in Chapter 7, were selected from the UMLS (Version 2012AB) to function as reference terminologies for SNOMED CT. Only English-language terminologies making use of the 'PAR' relationship and "INVERSE-IS-A" relationship attribute were chosen, as this study relies on paths of PAR links. Algorithms were designed for finding the numbers of *m:n* trapezoids for pairs of terminologies, one taken from the list of six reference terminologies, the other one being the July 2012 version of SNOMED CT. The algorithms were implemented in the Oracle Relational Database Management System (RDBMS) native programming language PL/SQL. The problem of cycles was addressed by adding tests that guaranteed that no concept appeared twice in a path. This required a large number of tests for longer paths, but the effect of these tests on overall performance was acceptable. Furthermore, as the programs developed for this research are typically executed only once (per UMLS release) or a very few times, no additional effort was put into optimizing the code.

It should be noted that multiple parents may lead to overlapping trapezoids, which could in turn lead to counting the same intermediate path concepts repeatedly. Thus, intermediate path concepts are collected, duplicates are eliminated, and counts of additional concepts are adjusted in the algorithms.

8.3 Results

8.3.1 Analysis of *1:k* and *k:1* Trapezoids

Table 8.1 below shows the comparison of SNOMED CT with six reference terminologies, each of which could potentially contribute new concepts to SNOMED CT. When the numbers in columns 3 and 4 were calculated, duplicate concepts were eliminated.

Table 8.1 Comparison of SNOMED CT with Six Reference Terminologies that could Contribute Concepts to SNOMED CT

Reference Terminology	**Size of Refer. Termin.**	**Additional Concepts in Reference Terminology**	**Additional Concepts in SNOMED**	**Number of *1:k* Trapezoids**	**Number of *k:1* Trapezoids**
MEDCIN3 _2012_07_16	279529	325	2635	2389	514
GO2012_04_03	61925	41	0	0	19
FMA3_1	82062	158	491	536	149
NCI2012_02D	95523	505	2604	2161	608
CPM2003	3078	25	237	180	19
UMD2012	15956	24	35	42	16

The first column shows the name of the reference terminology and the second its size. The third column defines the number of concepts that the reference terminology could contribute. SNOMED CT could also contribute concepts to five out of the six reference terminologies. The numbers of those are in the fourth column. Columns 5 and 6 list the total numbers of *1:k* trapezoids and *k*:*1* trapezoids. A path "on the right side" in a 1:3 trapezoid indicates that there are two concepts in SNOMED CT that could be contributed to the reference terminology.

Table 8.2 shows the numbers of observed *1*:*k* and *k:1* trapezoids, ordered by increasing values of *k*. The table shows that *1:k* trapezoids were found with *k* up to 9. For the mirror image case, *k:1* trapezoids, instances were found up to $k = 8$. Columns 2 and 6 show the running time of computing each kind of trapezoid. Columns 3 and 6 show numbers of distinct concepts in each kind of trapezoid.

Table 8.2 Comparison of SNOMED CT with Six Reference Terminologies by Trapezoid Size

Path Length Ratio of Reference Terminology: SNOMED	Running time	Number of Trapezoids	Additional Concepts in SNOMED	Path Length Ratio of Reference Terminology: SNOMED	Running time	Number of Trapezoids	Additional Concepts in Reference Terminology
1:2	3m 7s	5308	2602	2:1	31s	1311	758
1:3	53s	2144	1933	3:1	11s	228	270
1:4	41s	875	1080	4:1	11s	36	85
1:5	32s	557	736	5:1	3m 58s	21	43
1:6	3m 48s	261	507	6:1	2m 55s	2	10
1:7	3m 37s	112	223	7:1	2m 46s	0	0
1:8	3m 52s	41	81	8:1	2m 57s	1	7
1:9	4m 15s	6	26	9:1	2m 33s	0	0
1:10	3m 56s	0	0	10:1	3m 24s	0	0
1:11	4m 59s	0	0				

Figure 8.3 shows an example of a 2:1 trapezoid. The pair of concepts *Vaccines*, C0042210, and *Hepatitis A Vaccine, Inactivated*, C0795623, exists in the NCI2012_02D and SNOMED. In SNOMED, *Hepatitis A Vaccine, Inactivated* is a child of *Vaccines*. In NCI the two concepts are separated by *Vaccines, Inactivated*, C0042210. Thus *Vaccines, Inactivated*, is a concept that might be imported into SNOMED.

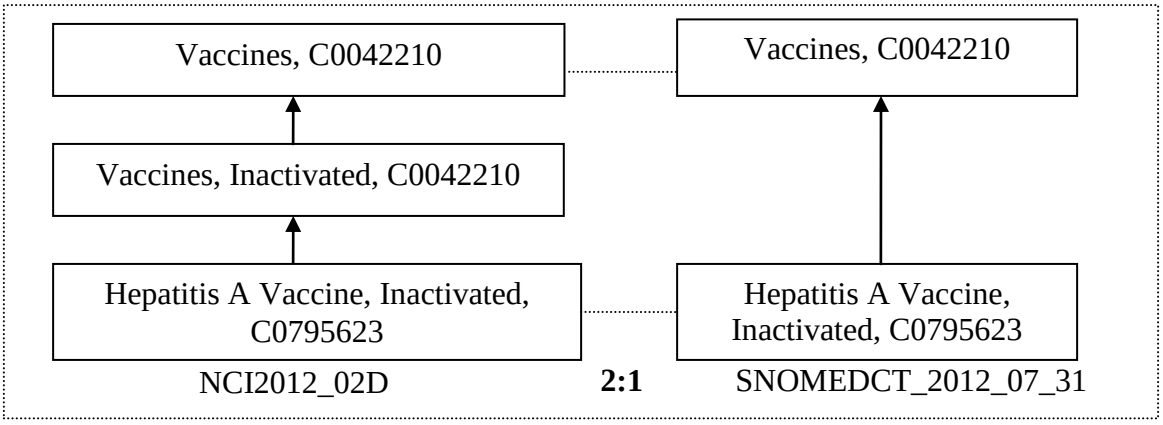

Figure 8.3 An example of a 2:1 trapezoid that suggests a concept import into SNOMED.

Figure 8.4 shows an example of a 3:1 trapezoid. Both the FMA and SNOMED CT contain the concepts "Connective Tissue" and "Loose areolar connective tissue." There is a direct link between them in SNOMED, but there are two concepts "Irregular connective tissue" and "Loose connective tissue" between them in the FMA. Thus it should be considered to import these two concepts into SNOMED CT.

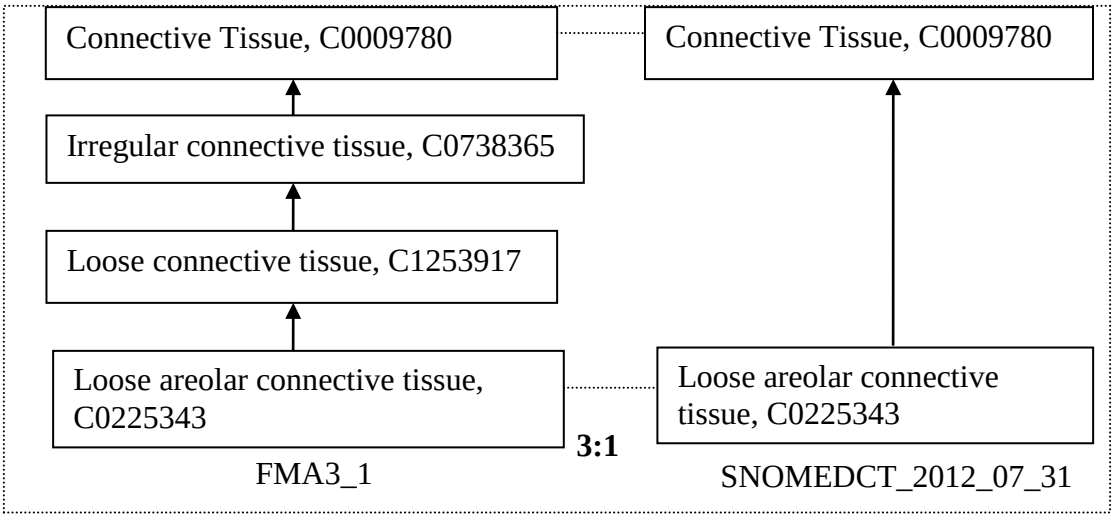

Figure 8.4 An example of a 3:1 trapezoid that suggests two concept imports into SNOMED CT.

Table 8.3 shows two more high density examples in a space-conserving table format, namely an 8:1 trapezoid with GO as the reference terminology, and a 1:9 trapezoid with SNOMED having a much higher density than MEDCIN. Whether one wishes to import all those concepts depends on the domain and goals of SNOMED and of the reference terminologies.

8.3.2 Analysis of *m*:*n* Trapezoids

With the consent of the designers of a terminology, intermediate path concepts in 1:*k* and *k*:1 trapezoids could be automatically imported to the terminologies missing those concepts. However for the cases of *m*:*n* trapezoids where $m > 1$ and $n > 1$, the intermediate path concepts cannot be automatically imported due to implicit relationships between intermediate path concepts.

Table 8.3 Two Examples of High-ratio Trapezoids

Reference Terminology	**SNOMED CT**
8:1	
GO2012_04_03	*SNOMEDCT_2012_07_31*
Immune System Processes, C1817756 immune effector process, C1817420 leukocyte mediated immunity, C1817894 lymphocyte mediated immunity, C1817899 B cell mediated immunity, C1155251 immunoglobulin mediated immune response, C1155250 type II cellular hypersensitivity, C1820377 type IIa hypersensitivity, C1327446	Immune System Processes, C1817756
Antibody-Dependent Cellular Cytotoxicity, C0003272	Antibody-Dependent Cellular Cytotoxicity, C0003272
1:9	
MEDCIN3_2012_07_16	*SNOMEDCT_2012_07_31*
Biliary Tract Surgical Procedures, C0005427	Biliary Tract Surgical Procedures, C0005427 Bile duct operation, C0400634 Repair of bile duct, C0193566 Repair of hepatic duct, C1280034 Anastomosis of hepatic ducts, C0193540 Anastomosis of hepatic duct to gastrointestinal tract, C0193531 Hepatojejunostomy, C0193425 Roux-en-Y hepaticojejunostomy, C0585537 Kasai procedure, C1536401
Portoenterostomy, Hepatic, C0032722	Portoenterostomy, Hepatic, C0032722

Structurally congruent concepts in Chapter 7 can be considered as intermediate path concepts in 2:2 trapezoids (in both terminologies, there are two "PAR" links attached to the same two concepts that are shared by both terminologies.). The relationship attached to structurally congruent concepts must be determined by human experts. Similarly, for *m*:*n* trapezoids where m >= 2 and n >= 2, the relationships of intermediate path concepts from both terminologies need to be determined by domain experts. As the values of *m* and

n grow, the possible relationships between intermediate concepts in trapezoids become more and more complex. In this section, 2:3 and 3:2 trapezoids will be analyzed. Samples of such configurations have been reviewed by a human expert, and the results are presented.

For intermediate path concepts X, Y and Z in a 2:3 trapezoid, as can be seen in Figure 8.5, it is hypothesized that there are **six possible cases** of how X, Y, and Z may relate to each other. Additionally, errors might be found in Terminology 1 and Terminology 2. Thus, eight hypotheses are defined.

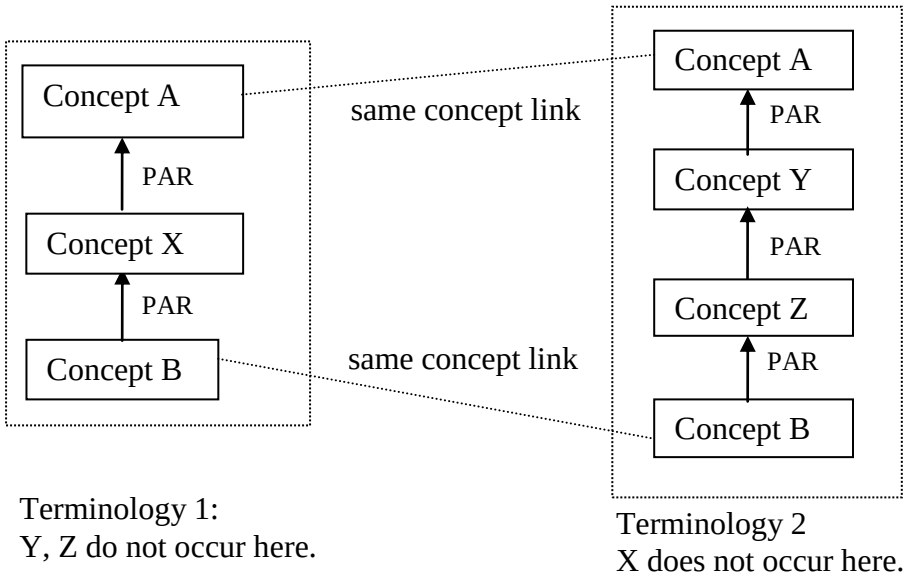

Figure 8.5 The layout of 2:3 trapezoids.

1) The concepts X and Y are alternative classifications. That means that concept A may be validly assigned X and Y as its children. However, these two assignments are indicative of two different ways of clustering the grandchildren of A. Furthermore, concept B may be correctly classified as a child of X and as a child of Z. However, Terminology 1 omits the classification by Y and Terminology 2 omits the classification by X.

2) It holds that B → Z → Y → X → A. In other words, X may be inserted as a child of A and a parent of Y into Terminology 2, thereby adding more detailed information to Terminology 2. Similarly, Y may be inserted as a child of X into Terminology 1, and Z maybe inserted as child of Y into Terminology 1. Such insertions should only be done with approval of a subject matter expert.

3) It holds that B → Z → X → Y → A. In other words, X may be inserted as a child of Y and a parent of Z into Terminology 2, thereby adding more detailed information to Terminology 2. Similarly, Y may be inserted as a parent of X into Terminology 1, and Z may be inserted as a child of X into Terminology 1. Such insertions should only be done with approval of a subject matter expert.

4) It holds that B → X → Z → Y → A. In other words, X may be inserted as a child of Z and a parent of B into Terminology 2, thereby adding more detailed information to Terminology 2. Similarly, Z may be inserted as a parent of X into Terminology 1, and Y may be inserted as a parent of Z into Terminology 1. Such insertions should only be done with approval of a subject matter expert.

5) Concept X is a real world synonym of concept Y, which was previously not recognized by the UMLS editors.

6) Concept X is a real world synonym of concept Z, which was previously not recognized by the UMLS editors.

7) There might be a structural error in Terminology 1, e.g., X is not really a child of A.

8) There might be a structural error in Terminology 2, e.g., Y is an unrecognized synonym of Z.

For intermediate path concepts X, Y and Z in a 3:2 trapezoid, as can be seen in Figure 8.6, eight hypotheses are defined.

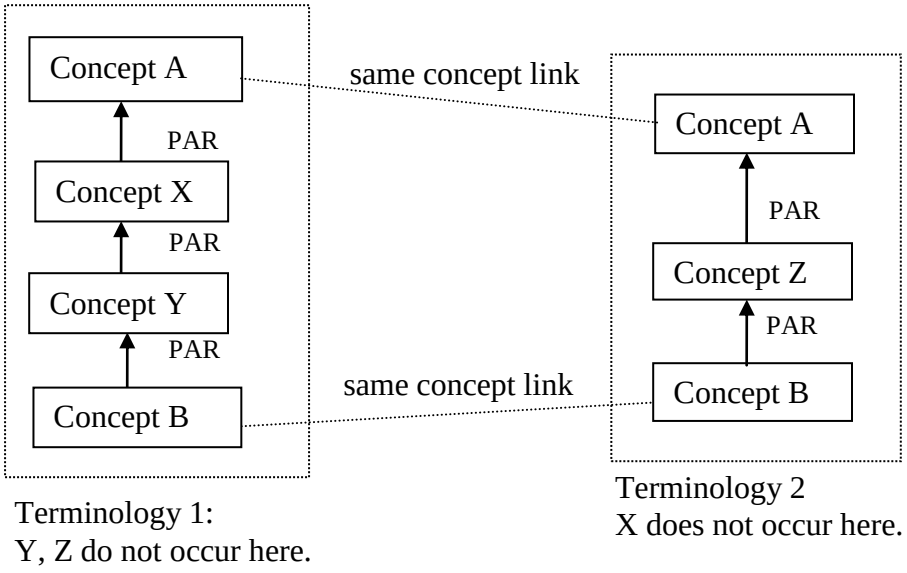

Figure 8.6 The layout of 3:2 trapezoids.

1) The concepts X and Z are alternative classifications.

2) It holds that B → Y → X → Z → A.

3) It holds that B → Y → Z → X → A.

4) It holds that B → Z → Y → X → A.

5) Concept Z is a real world synonym of concept X.

6) Concept Z is a real world synonym of concept Y.

7) There might be a structural error in Terminology 1,

8) There might be a structural error in Terminology 2.

Table 8.4 shows the number of 2:3 and 3:2 trapezoids found. In order to analyze the relationships of intermediate path concepts in the trapezoids, for 2:3 trapezoids, random samples of 50 trapezoids were chosen from each of MEDCIN, NCI, and FMA, all of which have more than 100 2:3 trapezoids. For 3:2 trapezoids, random samples of 50 trapezoids

were chosen from MEDCIN and NCI. If fewer than 100 trapezoids were found, i.e., for UMD, GO, FMA and CPM, all the trapezoids found were reviewed.

Table 8.4 Results for 2:3 and 3:2 Trapezoids of SNOMED CT and Reference Terminologies

Reference Terminologies	Size of Reference Terminology	2:3	Sample Size	3:2	Sample Size
MEDCIN3_2012_07_16	279529	634	50	121	50
NCI2012_02D	95523	354	50	229	50
FMA3_1	82062	106	50	38	38
UMD2012	15956	1	1	12	12
GO2012_04_03	61925	1	1	3	3
CPM2003	3078	5	5	2	2
Total	528073	1101	157	405	155

Dr. Yan Chen, a PhD who specialized in techniques for auditing medical terminologies and has training in Sports Medicine reviewed the sample. Table 8.5 shows the results according to the eight hypotheses for intermediate path concepts in 2:3 trapezoids. The results show that 40.8% are alternative classifications. Another 7.0% + 8.3% + 11.5% = 26.8 % fall into the three categories where the intermediate path concepts in the reference terminology could be imported into SNOMED, and vice versa.

Table 8.6 shows the results for 3:2 trapezoids according to the eight hypotheses for intermediate path concepts. The results show that 56.8% fall into alternative classifications. Another 9.7% + 3.9% + 7.1% = 20.7% fall in the three categories where the intermediate path concepts in the reference terminology could be imported into SNOMED and vice versa.

Table 8.5 Human Review Results of 2:3 Trapezoids

Reference Terminology	Sample Size	Alter. Classi-fication	Z → Y → X	Z → X → Y	X → Z → Y	X is a synonym of Y	X is a synonym of Z	Error in Termin-ology 1	Error in Termin-ology 2
MEDCIN 3_2012_07_16	50	21	5	4	7	3	7	3	--
NCI2012_02D	50	23	4	7	6	6	4	--	--
GO2012_04_03	1	1	--	--	--	--	--	--	--
CPM2003	5	3	--	--	--	1	1	--	--
UMD2012	1	--	--	--	--	1	--	--	--
FMA3_1	50	16	2	2	5	6	18	--	1
Total	157	--	--	--	--	--	--	--	--
Percentage	100%	40.8%	7.0%	8.3%	11.5%	10.8%	19.1%	1.9%	0.6%

Table 8.6 Human Review Results of 3:2 Trapezoids

Reference Terminology	Sample Size	Alter. Classi-fication	Y → X → Z	Y → Z → X	Z → Y → X	Z is a synonym of X	Z is a synonym of Y	Error in Termin-ology 1	Error in Termin-ology 2
MEDCIN 3_2012_07_16	50	32	5	1	4	5	2	1	--
NCI2012_02D	50	28	4	3	2	8	5	--	--
GO2012_04_03	3	1	--	--	2	--	--	--	--
CPM2003	2	--	--	2	--	--	--	--	--
UMD2012	12	2	--	--	1	9	--	--	--
FMA3_1	38	25	6	--	2	2	3	--	--
Total	155	88	15	6	11	24	10	1	0
Percentage	100%	56.8%	9.7%	3.9%	7.1%	15.5%	6.5%	0.6%	0%

Examples of 2:3 trapezoids will be used to illustrate the findings. Figure 8.7 shows an example where intermediate path concepts were deemed as alternative classifications. Thus, *non-infectious skin disorders* in the MEDCIN is a classification by infectiousness, while in SNOMED, *Age, sex or race-related dermatoses* is a classification by patient characterization.

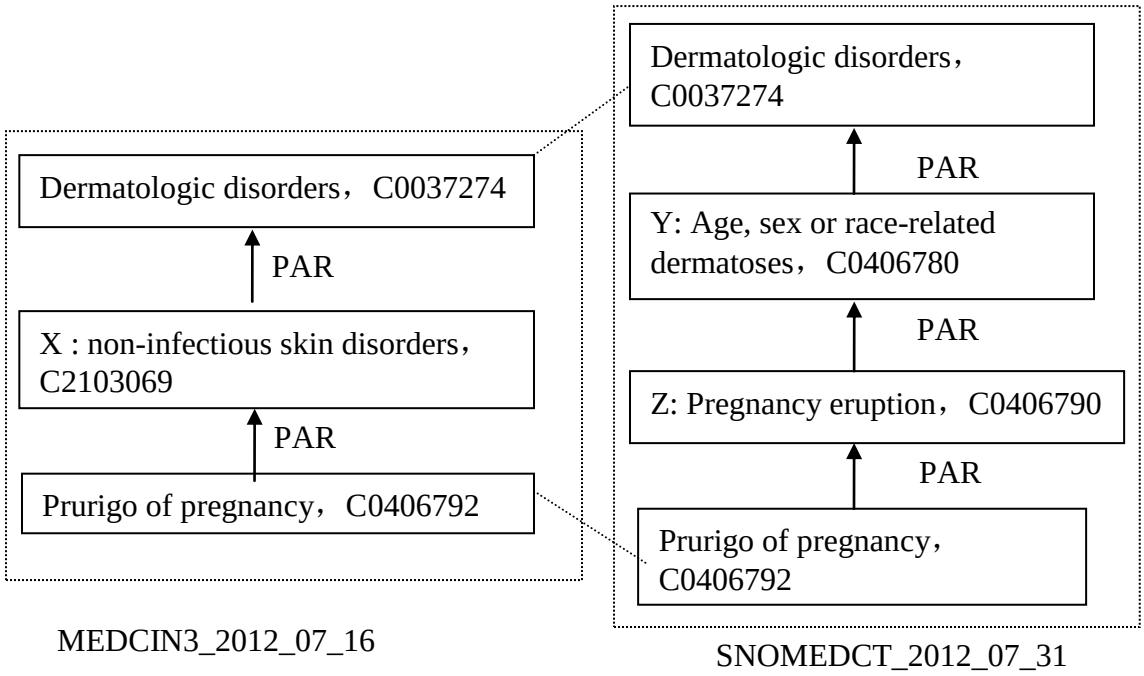

Figure 8.7 An example of alternative classification.

As stated in Chapter 7, the discovery of alternative classifications is useful, because it makes explicit the implicit assumptions of the ontology designers how they are viewing the world. This view could then be codified in the ontology by adding a more general concept XX as a parent of Concept X in Terminology 1 and a more general concept YY as a parent of Concept Y in Terminology 2, respectively. A possible name for XX is "Dermatologic disorder by Infectiousness" and for YY "Dermatologic disorder by population."

Figure 8.8 shows an example where Concept X was identified as a parent of the other Concept Y by the auditor. In this example, the concept *Systemic Congenital Disorder* can be a parent of *Congenital abnormality of lower limb AND/OR pelvic girdle*, thus the intermediate path concept *Systemic Congenital Disorder* from NCI may be added as a parent of *Congenital abnormality of lower limb AND/OR pelvic girdle* in SNOMED, and vice versa, if this is needed by the judgment of the curators of NCI and/or SNOMED. Alternatively, one might say that leaf node *Urogenital Abnormalities* is wrong in both terminologies, as it does not refer to "congenital."

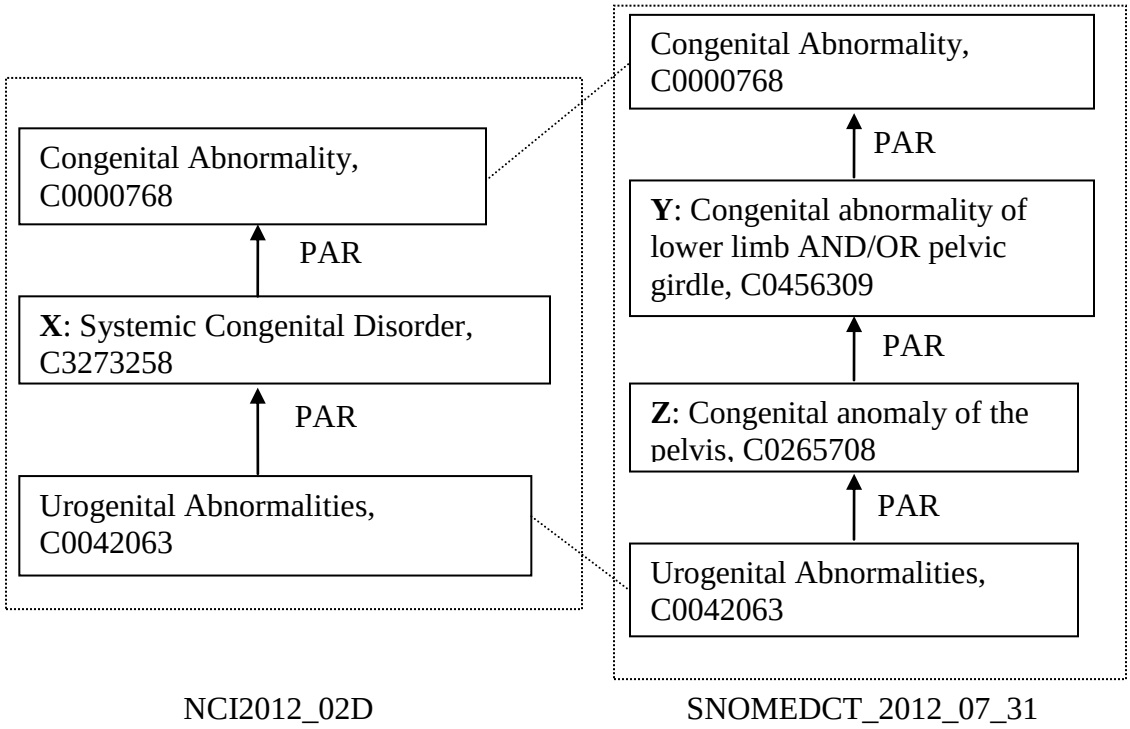

Figure 8.8 An example of Concept X being a parent of Concept Y.

Figure 8.9 shows an example where Concept X was identified as a parent of Concept Z and a child of Concept Y by the auditor. In this example, the concept *pulmonary obstructive disorders* can be a parent of *Bronchial Diseases*, and a child of *Disorder of lower respiratory system*, thus the concept *pulmonary obstructive disorders* from

MEDCIN may be added as a parent of *Bronchial Diseases*, as well as a child of *Disorder of lower respiratory system* in SNOMED, and vice versa. Similarly, in Figure 8.10, *Gadolinium-Chelate*, which was deemed a child of *Magnetic resonance imaging contrast media*, can be added in SNOMED CT accordingly.

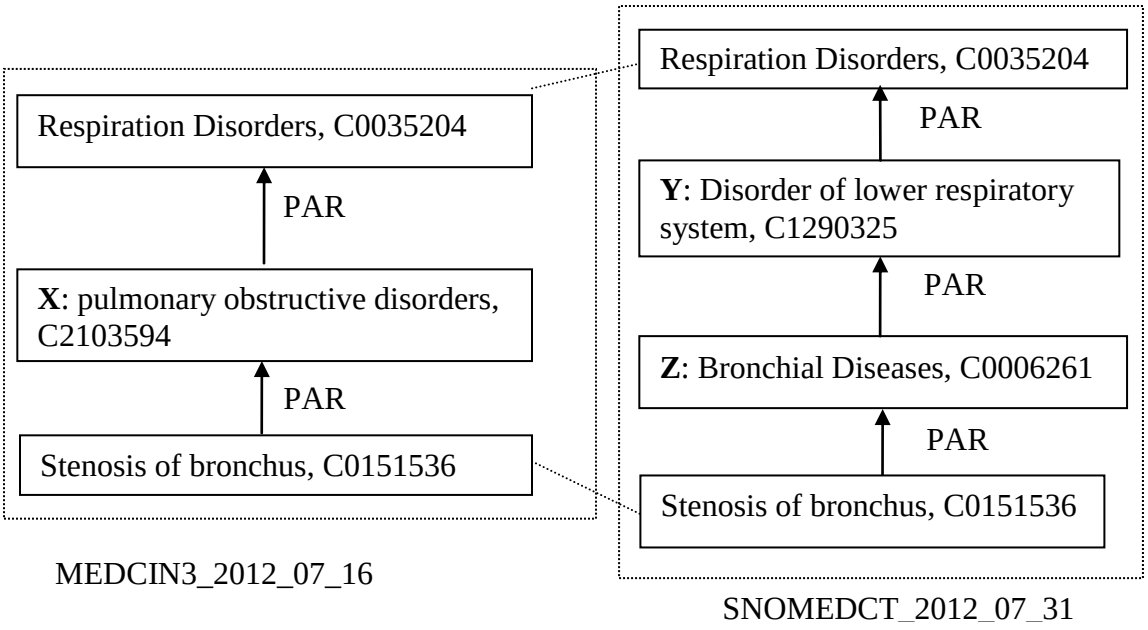

Figure 8.9 An example of Concept X being a parent of Concept Z, and a child of Concept Y.

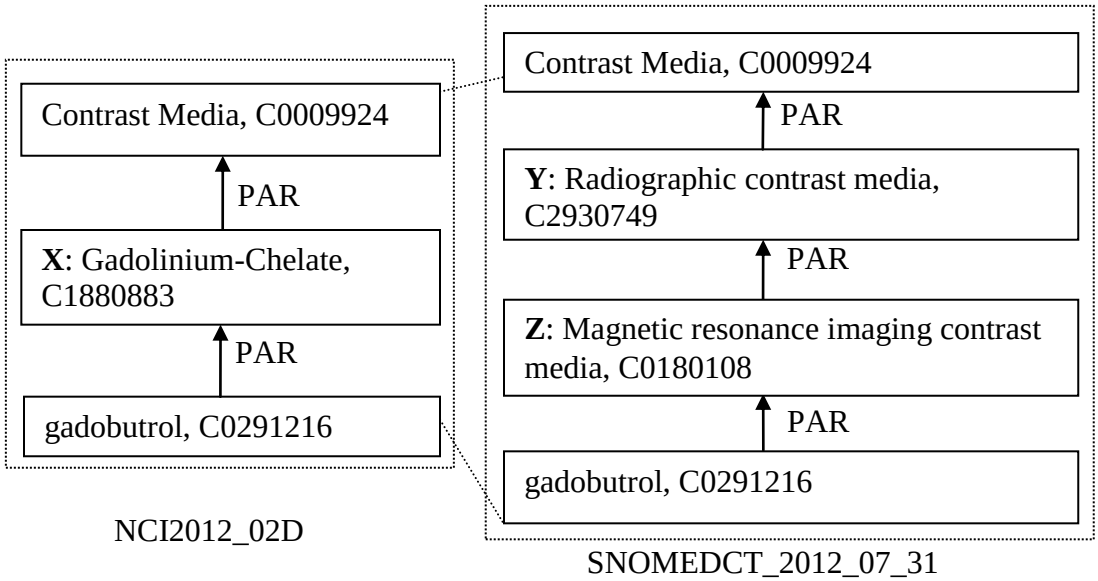

Figure 8.10 An example of Concept X being a child of Concept Z.

Figure 8.11 shows an example where Concept X was identified as a synonym of Y. In this example, the concept *Ingested food* from FMA and *Gastrointestinal Contents* from SNOMED were deemed synonyms that were not previously recognized and thus should be merged.

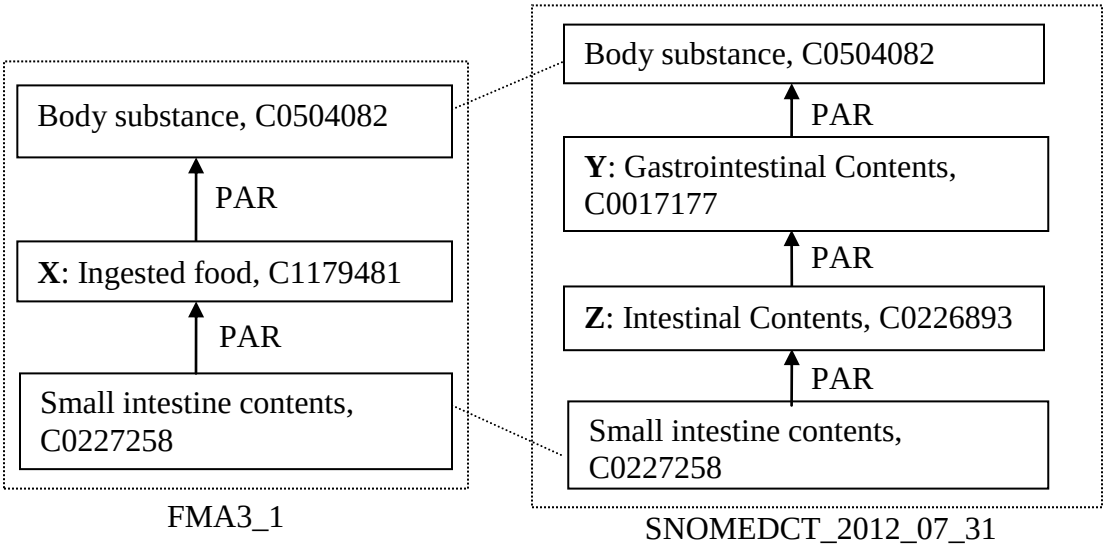

Figure 8.11 An example of Concept X being a synonym of Concept Y.

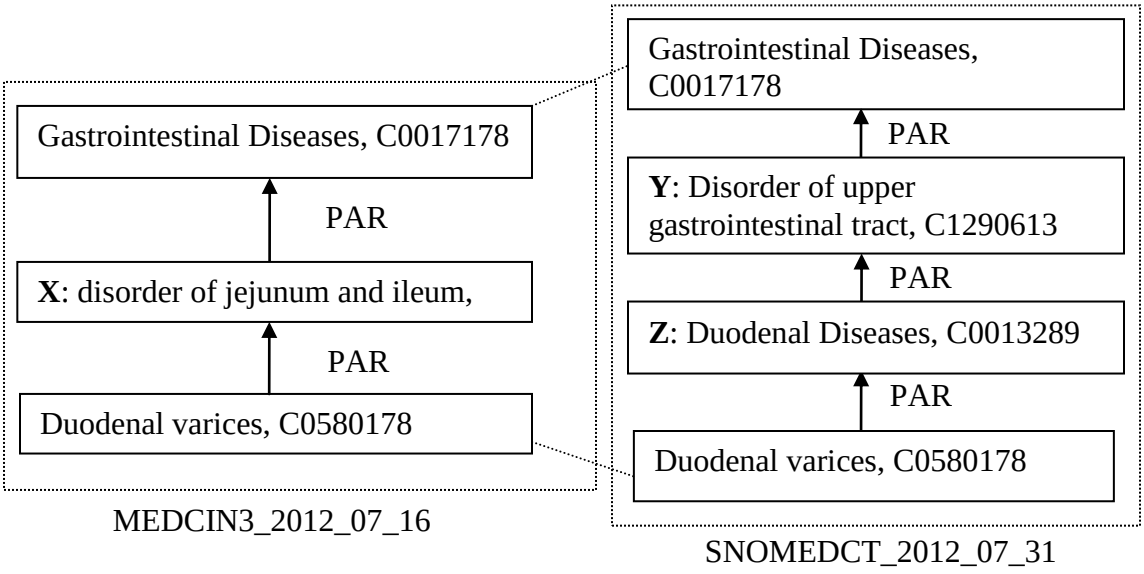

Figure 8.12 An example of an error in Terminology 1.

The auditor determined that there is an error in MEDCIN in Figure 8.12. Duodenum, jejunum and ileum are the three different segments of the small intestine. Duodenal varices are located in the duodenum, not jejunum or ileum. Therefore, Duodenal varices are duodenal diseases instead of disorders of the jejunum and ileum.

8.4 Discussion

For *m*:*n* trapezoids where m >= 3 and n >=3, it becomes difficult for human experts to make judgments as to what the relationships of the intermediate path concepts from two terminologies are, because there is a combinatorial increase in the number of possibilities. Figure 8.13 shows a hypothesized 3:3 trapezoid configuration.

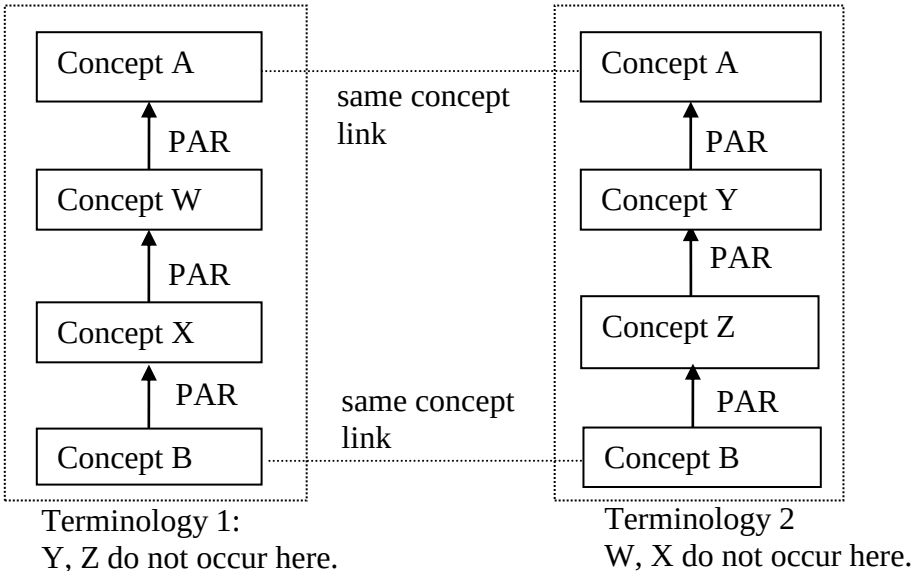

Figure 8.13 The layout of 3:3 trapezoid concepts.

It is hypothesized that there are **13 possible cases** for how W, X, Y, and Z may relate to each other.

1) The concepts W and Y are alternative classifications. That means that concept A may be validly assigned W and Y as its children. However, these two assignments are indicative of two different ways of clustering the grandchildren.

2) It holds that B → Z → Y → X → W → A.

3) It holds that B → Z → X → Y → W → A.

4) It holds that B → X → Z → Y → W → A.

5) It holds that B → Z → X → W → Y → A.

6) It holds that B → X → Z → W → Y → A.

7) It holds that B → X → W → Z → Y → A.

8) Concept W is a real world synonym of concept Y, Concept X is a real world synonym of concept Z.

9) Concept W is a real world synonym of concept Y and X is a parent of Z.

10) Concept W is a real world synonym of concept Y and X is a child of Z.

11) Concept X is a real world synonym of concept Y.

12) Concept W is a real world synonym of concept Z.

13) Concept X is a real world synonym of concept Y. Concept W is a real world synonym of concept Z, i.e., cases 11) and 12) hold at the same time.

14) There might be a structural error in Terminology 1.

15) There might be a structural error in Terminology 2.

With the values of *m* and *n* growing, the number of possible cases of relationships among intermediate path concepts grows even faster. Thus, algorithms that might be used to identify these complex relationships are desired as future work.

A look at Table 8.3 raises the question whether SNOMED CT is possibly too detailed in its modeling. When a clinician is looking for a concept to describe a symptom of a patient in an Electronic Medical Record system, s/he might find it difficult to identify a proper concept due to unnecessarily fine granularity of a terminology. Are the eight intermediate path concepts between "Portoenterostomy, Hepatic" and "Biliary Tract Surgical Procedures" really needed? One may also question whether importing even some of these intermediate path concepts into MEDCIN is really in alignment with the intentions of the MEDCIN designers. Examples were found, where concepts from FMA could be imported into SNOMED CT. However, some of those concepts are likely to be judged as clinically irrelevant, and therefore it is not expected that they would be integrated into SNOMED CT, even if the opportunity exists. In general, while it may appear that concepts are missing in a terminology, this may well be by a choice dictated by the intended domain and application area, and the final determination has to be made by a curator of the terminology.

The biggest limitation of this research is that only vertical configurations of PAR links are used. The UMLS also supports RB (Relationship Broader) links that function in an analogous way to PAR links, but differ in the source of the relationships. Furthermore, the UMLS allows annotating the "REL" (relationship) PAR with additional information ("RELA"). Many PAR relationships do not have any annotation (roughly half of them), but about 20,000 are annotated to indicate a *part* link, distinguishing those from relationships annotated in other ways, e.g., those expressing an "IS-A" link. In this research only IS-A annotations were used. A thorough analysis distinguishing between PAR relationships

with other annotations and comparing the results with paths of RB relationships would provide deeper insights into the phenomenon of density differences.

A further limitation of this work is that it uses SNOMED CT concepts and all reference terminology concepts in the format that they were provided in by the UMLS. There may be differences between the original concept representation of SNOMED CT (or the reference terminologies) and the representation of SNOMED CT that is accessible through the UMLS.

In future work it would also be interesting to investigate the question whether SNOMED CT should be considered too detailed, in the sense that there are too many pairs of concepts without practically significant distinctions between them. Given that the size of a terminology creates a strain on both human users and computer systems, "slimming down" SNOMED CT without losing any valuable information appears to be desirable.

8.5 Conclusions

For 1:*k* and *k*:1 trapezoids, path length ratios of up to 1:9 and 8:1 were observed, i.e., a parent in MEDCIN was separated in SNOMED CT from the MEDCIN child by a path of 9 PAR relationships. With the consent of the owners, SNOMED CT could be extended by importing concepts from the six reference terminologies. Meanwhile six reference terminologies can also be extended by imported concepts from SNOMED CT. For *m*:*n* trapezoids, random samples of 2:3 and 3:2 trapezoids were reviewed by human experts. It is conjectured that for *m*:*n* trapezoids where m >= 3 and n >=3, it would be extremely difficult for human experts to make judgment on the relationships of intermediate path

concepts in the trapezoids. Automatic algorithms to identify relationships of intermediate path concepts in complex trapezoids are desirable and thus worthy of further exploration.

CHAPTER 9

A FAMILY-BASED FRAMEWORK FOR SUPPORTING QUALITY ASSURANCE OF BIOMEDICAL ONTOLOGIES IN BIOPORTAL

9.1 Introduction

Modern biomedical science is impossible without the management and integration of large data sets. Moreover, the proliferation of interdisciplinary research efforts in the biomedical field is fueling the need to overcome terminological barriers when integrating knowledge from different fields into a unified research project. Thus, biomedical research needs the support of well-developed and well-maintained ontologies that provide structured domain knowledge for data integration, natural language processing, and decision support [118, 119].

The National Center for Biomedical Ontology (NCBO) provides an encyclopedic repository of over 300 ontologies within a uniform development and visualization system covering many different domains. As BioPortal ontologies underlie various Health Information Systems (HIS), Electronic Health Record (EHR) systems, Health Information Exchanges (HIEs) and healthcare administrative systems (see Chapter 1), BioPortal is growing in importance. With the BioPortal framework maturing, the time has come to stress the significance of QA methodologies for BioPortal ontologies and to further develop them.

The Web Ontology Language (OWL) and the Open Biological and Biomedical Ontologies (OBO) formats are standards based on description logic (DL) that provide a common model for creating ontologies. Most of the ontologies in BioPortal are released in

one of these two formats, while some ontologies are released in the Rich Release format (RRF).

Abstraction networks are compact networks summarizing the structure and content of ontologies. Abstraction networks have been derived in uniquely tailored ways for various individual ontologies. These abstraction networks include: an object-oriented schema for the Medical Entities Dictionary (MED) [120]; the Refined Semantic Network for the UMLS (see Chapter 2) [27]; and various area taxonomies and partial-area taxonomies for SNOMED CT [60], NCIt [59], the Ontology of Clinical Research (OCRe) [61], the Sleep Domain Ontology (SDO) [62], and the Ontology for Drug Discovery Investigations (DDI). These abstraction networks were shown to support orientation into the ontologies' content and structure and have been used to support their QA. However, it would not be practical to derive a unique type of abstraction network for each individual BioPortal ontology. Because the large majority of BP ontologies are released in OWL (Web Ontology Language) or OBO (the Open Biological and Biomedical Ontologies) formats, many of them share a common underlying structure, such as the usage of domain-defined object properties.

A *family of ontologies* is defined as a set of ontologies satisfying some overarching conditions regarding their structural features. By structural features, knowledge elements of classes of an ontology are referred to, such as kinds of object properties, whether classes with multiple parents exist and whether data properties are distinct from object properties. Unique combinations of structural features can be used to group BioPortal ontologies into a family.

In this chapter, seven families according to combinations of various structural features available in BioPortal ontologies were identified. For example, one family consists of those ontologies with object properties given explicitly defined domains and ranges. Another family contains ontologies with object properties either used as restrictions on classes or given explicitly defined domains and ranges. Details and metrics of structural features for 186 BioPortal ontologies were collected and each ontology was classified into the proper family.

The organization of ontologies into families serves as the foundation for a new family-based QA framework for ontologies, utilizing a uniform abstraction network derivation technique and uniform abstraction network-based QA regimen for a whole family of ontologies. Streamlining the abstraction network derivation and the QA process will result in higher efficiency and lower cost of QA. As an illustration of the abstraction network-based QA framework, it is applied to the Cancer Chemoprevention Ontology (CanCo) [121, 122]. The abstraction network for the CanCo is derived and presented in this chapter. The results of an initial QA review of CanCo based on its abstraction network are given.

Some aspects of the new family-based QA framework presented in this chapter are beyond the scope of this dissertation, although the various aspects of the framework are illustrated using examples. By identifying the structural features defining such families of ontologies, and classifying ontologies into the families, the groundwork for the family-based QA framework is laid. This framework will enable automated abstraction network derivation and semi-automated QA regimens, bringing to bear computer support for the QA of many biomedical ontologies of BioPortal.

9.2 Background

9.2.1 Structural Features of BioPortal Ontologies

In OWL, object properties are important ontological elements, used to relate classes and to represent potential relationships between class instances. In ontologies, object properties are utilized in several ways. Object properties can be given explicitly defined domains and ranges, i.e., *global* limitations on instantiation. An object property's domain and range can consist of any number of classes from the ontology.

Below is an example in Manchester OWL Syntax of an object property with an explicitly defined domain and range taken from CanCo. In this example, the object property is named *has disease location* and has *Disease* class as its domain, and *Organ* class defined as its range. Any instance of *has disease location* must have a domain that is a disease and a range that is an organ.

```
ObjectProperty: has_disease_location
          Domain: Disease
          Range: Organ
```

Object properties can also be utilized in class restrictions, such as in subclass axioms and class equivalence axioms. Class restrictions are a less strict, *local*, limitation on the instantiation of object properties. The use of restrictions is more flexible than rigorously defining the domain of every object property and is a common way object properties are utilized (see Section 9.4).

In this chapter, abstraction networks were derived for OCRe [61] and SDO [62], both available in BioPortal, using object properties to create different types of *area taxonomies* and *partial area taxonomies* (see Section 2.1.3). Taxonomies are abstraction networks that group together classes of similar structure and/or semantics. These

taxonomies are used to support summarization and QA of ontologies by highlighting groups of concepts that have a higher likelihood of errors. For more details on defining taxonomies see Section 9.4.3 and work of Wang et al. [60] and Ochs et al. [61, 62].

The taxonomies derived for OCRe's Entity hierarchy utilized only object properties with explicitly defined domains. For the SDO taxonomies, either object properties with explicitly defined domains or object properties used in class restrictions or both were considered to create three different kinds of taxonomies, each of a different granularity [62]. A preliminary analysis of the Gene Ontology (GO) (with cross maps to ChEBI) showed that taxonomies using object properties used in class restrictions on subclass axioms and in class equivalence axioms can be derived.

Data properties (attributes) are similar to object properties except that the range is a literal value, such as a number or a character string. Like object properties, data properties can be given explicitly defined domains or be used in class restrictions. The previous research has focused on using only object properties to derive taxonomies, but by modifying the abstraction network derivation methodologies, data properties can potentially be used independent of, or in conjunction with, object properties for deriving new kinds of taxonomies. Below is an example of a data property, *has sequence*, taken from CanCo, with a domain consisting of two classes, *Protein* and *Nucleic Acid*, and a range value defined as a character string.

```
DataProperty: hasSequence
    Domain: Protein, NucleicAcid
    Range: xsd:string
```

Ontologies are organized in a hierarchical structure where the more general classes are at the top and the most specific classes are at the bottom. An ontology hierarchy can be organized either as a directed acyclic graph (DAG), where classes can have multiple

superclasses, or as a tree where each class except for the root has exactly one superclass. Hierarchical relationships can be utilized in deriving abstraction networks as demonstrated by Wang et al. [65].

9.3 Methods

As mentioned in Section 9.1, the goal of this research is to create widely applicable uniform abstraction network derivation algorithms and uniform QA methodologies that will work for many ontologies without modification. To accomplish this, ontologies have to be grouped into families that exhibit similar structural features. For these families, an abstraction network can then be derived with the same algorithm for each ontology of a family. The structural features must (a) be common enough to create families of meaningful sizes; and (b) be useful for deriving abstraction networks capable of supporting summarization and QA.

9.3.1 Ontology Classification

Object properties are widely used in ontology development and introduce a significant amount of knowledge into an ontology. Given a set of ontologies, each ontology can be classified into one (or potentially more) families based on the presence or absence of the previously defined structural features. In this chapter, ontologies are classified into seven disjoint families with classification priority given to structural conditions that have been proven useful for deriving abstraction networks in previous research. In each case, taxonomies were successfully shown to support summarization and QA of the underlying ontology. Figure 9.1 illustrates a binary decision tree for classifying ontologies into

families. The diamond boxes represent the conditions and the rectangles represent the seven enumerated families of ontologies, plus the starting point "All ontologies."

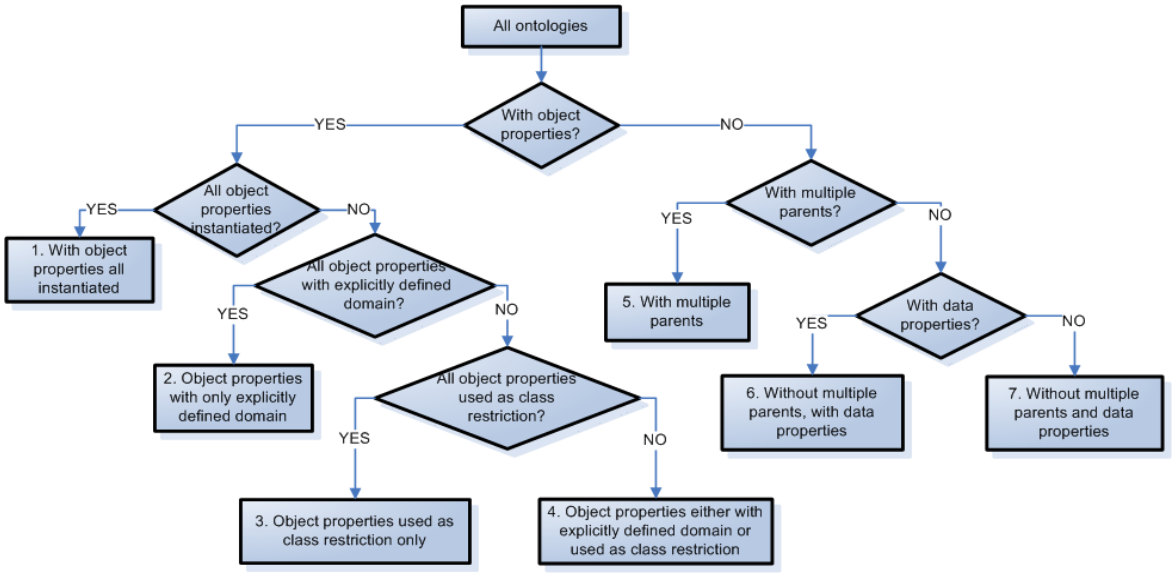

Figure 9.1 A binary decision tree for classifying ontologies into seven disjoint families.

Object properties represent the semantic connections between classes, expressing the domain knowledge of the ontology. The importance of object properties is manifested, for example, in the consideration of classes of ontologies as primitive if they miss some object properties. Thus, they have been chosen to initially separate families into two disjoint groups: those with object properties and those without object properties (Figure 9.1). These high-level groups dictate the *type* of abstraction network that can be derived for the ontologies of a family.

The ontologies that have object properties are further divided into four disjoint subgroups:

1. The first group consists of ontologies that have all their relationships instantiated (namely, SNOMED CT and NCIt).

2. The second group consists of ontologies that have all object properties with only explicitly defined domains.

3. The third group consists of ontologies that have all object properties only used in class restrictions.

4. The fourth group consists of ontologies that have at least one object property with an explicitly defined domain and at least one object property used in a class restriction.

Two of the largest and most used ontologies in BioPortal, SNOMED CT and NCIt, share a similar ontological model based on description logics. The model of these two ontologies is assigned to a separate family, since all their relationships are instantiated, which means that each pair of concepts connected by a relationship is concretely linked. In contrast, all other BioPortal ontologies' object properties are *potential* connections between classes, with only some of them instantiated with concrete links. In previous work, taxonomies were derived for SNOMED CT [60] and NCIt [59] using both their lateral and hierarchical relationships.

The second group of ontologies, which do not have object properties, are divided into three disjoint subgroups.

5. The first subgroup consists of ontologies that have some classes with multiple parents.

6. The second subgroup consists of ontologies that have data properties but have no classes with multiple parents.

7. The third subgroup consists of ontologies that have no data properties and no classes with multiple parents.

In this way, ontologies are grouped into seven disjoint families that exhibit different structural conditions.

9.3.2 Generalizable Design of Abstraction Networks for Families

Previously, abstraction network-based QA was a "one at a time" methodology; the research on developing techniques for deriving abstraction networks and developing QA methodologies was done on a per-ontology basis. The process of abstraction network derivation utilizes structural elements from an ontology to algorithmically create a "summary." Therefore, by deriving abstraction networks using the set of structural features common to all members of a family, abstraction networks can be derived with the same algorithm for each member of the family.

This generalizable abstraction network-based QA methodology will be illustrated by deriving a partial-area taxonomy for the Cancer Chemoprevention BioPortal Ontology (CanCo) [121, 122]. All of the object properties in CanCo are given explicitly defined domains. Therefore, the same taxonomy derivation methodology can be utilized as previously developed for OCRe, since both ontologies belong to Family 2. A review of the different partial-areas of CanCo's taxonomy was performed to demonstrate how anomalies in the taxonomy highlight classes with a high likelihood of modeling problems.

For practical QA work, it is necessary to create software for automatically deriving and visualizing abstraction networks for families of ontologies. In previous work, the Biomedical Layout Utility for SNOMED CT (BLUSNO) [123], a tool for automatically deriving and visualizing abstraction networks for SNOMED CT was developed. The experience with BLUSNO has guided work on the development of a utility called the Biomedical Layout Utility for the Web Ontology Language (BLUOWL) [124]. In an early prototype of BLUOWL, users can select an ontology expressed in OWL from the family of BioPortal ontologies with only object properties with explicitly defined domains, and

BLUOWL generates a partial-area taxonomy on the fly. The resulting abstraction network diagram can be manipulated by the user. The partial-area taxonomy for CanCo (see Figure 9.2) was derived using the BLUOWL prototype.

9.4 Results

Between September 2012 and January 2013 210 distinct BioPortal ontologies were collected, representing 64% of the 330 BioPortal ontologies available at that time. In addition to SNOMED CT and NCIt, only ontologies released in OWL and OBO formats, were considered. Each ontology was converted from the stated view to the inferred view, to utilize all inferable axioms, using the HermiT classifier [125]. Of the 210 ontologies, 24 ontologies were not investigated for various reasons, e.g., inconsistency with the OWL standard, missing imports, compatibility with the classifier, etc.

The remaining 186 ontologies included the Gene Ontology (GO), Foundational Model of Anatomy (FMA), Ontology for General Medical Science (OGMS), Ontology of Clinical Research (OCRe), Sleep Domain Ontology (SDO), Vaccine Ontology (VO), Infectious Disease Ontology (IDO), and others. In total, 115 ontologies were in OWL format, 70 were in OBO format, and two in flat file format.

9.4.1 Commonality of Structural Conditions

As mentioned before, there must be enough ontologies exhibiting a particular structure to meet the criterion that a family should be of meaningful size. Table 9.1 lists the numbers of BioPortal ontologies exhibiting each of the structural features that were utilized to analyze the ontologies. For brevity, in Table 9.1 and onward, abbreviated names for those features are used. For example, object properties with explicitly defined domains are called

domain-defined object properties. If object properties are used in class restrictions, they are called restriction-defined object properties.

From Table 9.1 one can see that there are some ontologies with both kinds of object properties. In fact, 62 ontologies have some domain-defined object properties and some restriction-defined object properties. Nineteen ontologies have only domain-defined object properties and 69 ontologies have only restriction-defined object properties. Furthermore, 71 out of 186 ontologies have data properties.

Table 9.1 Ontologies in the Sample Set which Exhibited a Particular Structural Condition

Characteristic	**# of Ontologies with Characteristic**	**% of Sample (n = 186)**
Object properties (total)	150	80.6
Domains-defined object properties	81	43.5
Restriction-defined object properties	131	70.4
Data properties (total)	71	38.2
Multiple parents (DAG)	110	59.1
No multiple parents (Tree)	76	40.9

For ontologies without object properties, hierarchical structure conditions can be used for abstraction network derivation. There are nine ontologies without object properties having some classes with multiple parents. They are APO, FBSP, HEALTHINDICATORS, HOMHARVARD, HP, IMMDIS, OGMD, PEDTERM and YPO [126].

9.4.2 Members of Families

Table 9.2 lists the families of ontologies which have object properties or instantiated relationships. Since the families were defined as disjoint, the numbers in Table 9.2 are not coming from Table 9.1, but from the disjoint partition described above, e.g., there are 19

ontologies with domain-defined object properties.

Table 9.2 Families for Ontologies that have Object Properties (Relationships)

Family	Structural Condition	# of Ontologies	Samples
1	Ontologies with all relationships instantiated	2	SNOMED CT, NCIt
2	Ontologies with only domain-defined object properties	19	Cancer Chemoprevention Ontology (CanCo), International Classification of Functioning, Disability and Health (ICF), Physical Medicine and Rehabilitation (PMR)
3	Ontologies with only restriction-defined object properties	69	Gene Ontology (GO), Cereal Plant Development (GRO_CPD), Host Pathogen Interactions Ontology (HPIO)
4	Ontologies with some domain-defined object properties and some restriction-defined object properties	62	Sleep Domain Ontology (SDO), Infectious Disease Ontology (IDO)

Table 9.3 lists the families of ontologies that have no object properties. As in Table 9.2, the numbers are computed from the disjoint sets above.

Table 9.4 lists a sample of ontologies from Family 2, i.e., those with only domain-defined object properties. In Section 9.4.3, the CanCo ontology of this family will be used to illustrate how its taxonomy, created automatically by BLUOWL, looks and how QA work for CanCo, based on the CanCo taxonomy, can be performed.

Table 9.3 Families of Ontologies that have no Object Properties (Relationships)

Family	Structural Condition	# of Ontologies	Samples
5	Ontologies that have classes with multiple parents	9	Ascomycete phenotype ontology (APO), Human Phenotype Ontology (HP), Ontology of Glucose Metabolism Disorder (OGMD)
6	Ontologies that have no classes with multiple parents and have classes with data properties	3	Cell Behavior Ontology (CBO), CareLex
7	Ontologies that have no classes with multiple parents and data properties	22	Ontology for General Medical Science (OGMS), Reproductive trait and phenotype ontology (REPO), Sample processing and separation techniques (SEP)

Table 9.4 Sample of Ontologies that have only Domain-defined Object Properties

Ontology Name	# of classes	# of object properties
Animal natural history and life history (ADW)	364	16
Biomedical Resource Ontology (BRO)	487	17
Cancer Chemoprevention Ontology (CanCo)	127	37
International Classification of Functioning, Disability and Health (ICF)	1595	41
Physical Medicine and Rehabilitation (PMR)	137	14
RAPID Phenotype Ontology (RPO)	1544	157
Student Health Record (SHR)	343	35
Syndromic Survaillance Ontology (SSO)	176	11
Top-Menelas	524	296

9.4.3 Illustration for the Cancer Chemoprevention Ontology (CanCo)

To illustrate the viability of the family-based QA framework, the Cancer Chemoprevention Ontology (CanCo) has been chosen. CanCo has 127 classes and 37 object properties. The Basic Formal Ontology (BFO), an upper level BioPortal ontology [127], was fully migrated into CanCo for reuse in its design. The BLUOWL prototype is used to

automatically generate and display the taxonomy of any member of Family 2. The taxonomy for CanCo, generated by BLUOWL, appears in Figure 9.2.

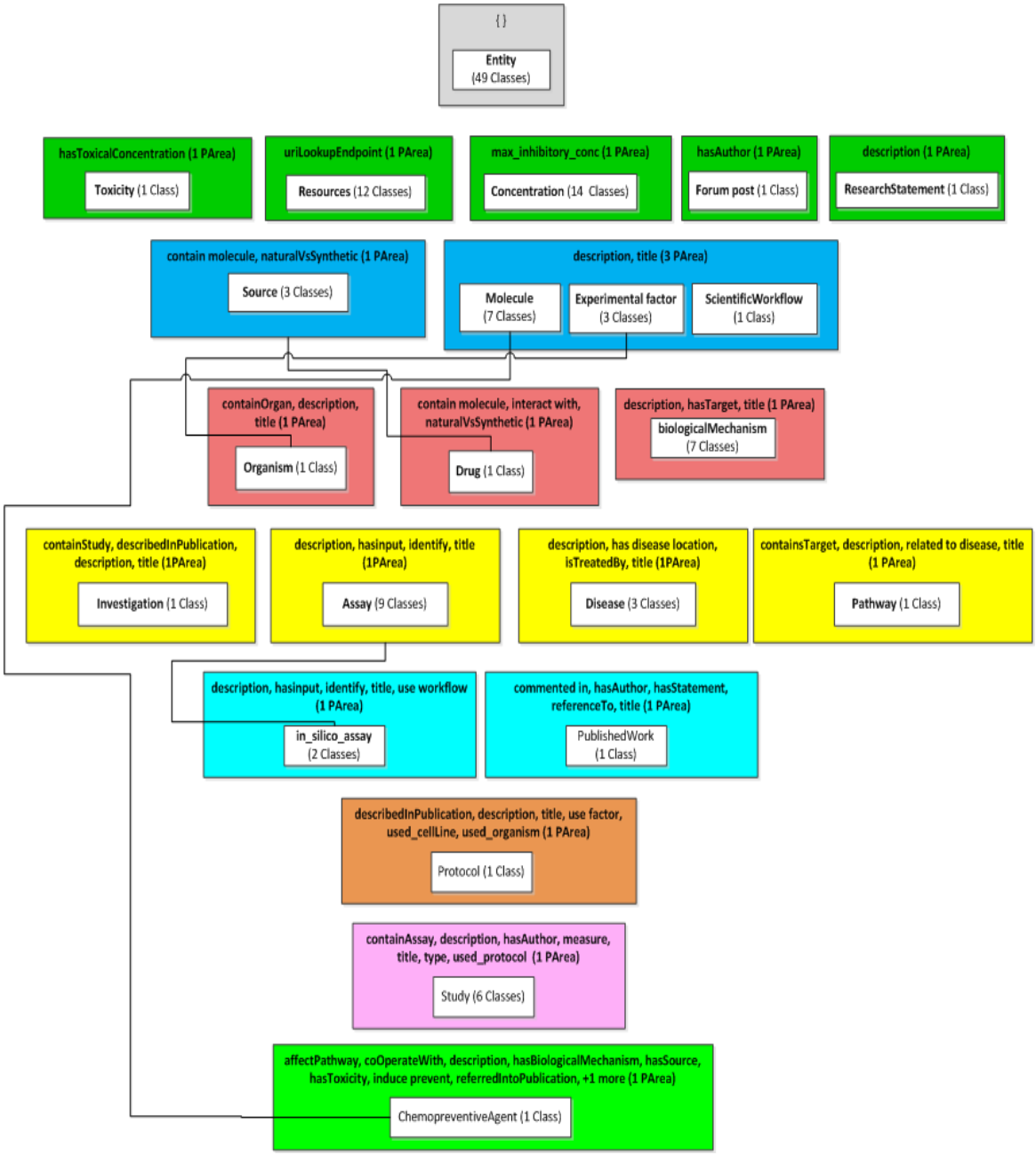

Figure 9.2 Partial-Area Taxonomy of Cancer Chemoprevention Ontology (CanCo).

In the partial-area taxonomy of CanCo, areas for every existing set of object properties in the ontology are organized into color-coded levels based on their numbers of object properties. For example, areas with three object properties appear in red in Level 3. The top level is Level 0 and always consists of one area. This area is the root area of the taxonomy. It summarizes the classes with no object properties. In general, the level number is equal to the number of object properties.

Partial-areas are represented using white boxes within colored area boxes and are labeled using their roots. The lines (between boxes and relationships) are *child-of*. The *child-of* relationships from all but four partial-areas are directed to the root *Entity* partial-area and are not shown to avoid clutter. This indicates that most sets of object properties of areas are disjoint. The only *child-of* relationships shown point to the partial areas *Source*, *Experimental Factor*, *Module* and *Assay*. Except for the area labeled with *description* and *title*, with three *partial-areas*, all areas contain only one partial area.

Most partial areas contain only one class, with the exception of three large ones: *Entity* (49), *Concentration* (14), and *Resources* (12). Medium size partial-areas (5-9 classes) include: *Assay* (9), *Module* (7), *Biological Mechanism* (7) and *Study* (6). These nine partial areas, covering 104 classes, provide an excellent summary of the content and structure of CanCo. Note that the terms “large” and “small” are relative to the overall size of CanCo. In SNOMED CT, a structure with 49 classes would not be considered large.

According to extensive experience with SNOMED CT [60, 84] and NCIt [59] partial-area taxonomies helped to identify anomalies in the modeling, characterizing sets of concepts with a high likelihood of errors. In recent QA work on BioPortal ontologies such as OCRe [61] and the Sleep Domain Ontology [62], it was found that large partial-areas

characterize sets of concepts with a high likelihood of errors. There are three large partial-areas in Figure 9.2. This constitutes an anomaly for CanCo. The second anomaly in the CanCo taxonomy is the unique area (*description, title*) with three partial-areas: *Module* (7), *Experimental factor* (3) and *Scientific workflow* (1). The third anomaly is defined by the few (four) *child-of* relationships not directed to *Entity*.

In the following, it is shown how these anomalies helped expose modeling problems. First, consider the *Entity* (49) root partial-area containing all classes with no object properties. It was found that 39 out of 49 were migrated from BFO, which is modeled without object properties. Closer examination reveals that 20 of them are leaves (classes without children) in CanCo. That means they were not used as the basis for child classes in the chemoprevention domain and should not have been migrated to CanCo at all. The process of "hiding" all 20 such leaves from CanCo would not affect any other CanCo classes. For details on a hiding mechanism for BioPortal ontologies, see [128].

Another modeling problem concerns both large partial-areas *Entity* (49) and *Concentration* (14). The class *Concentration* and all its 13 descendants have the object property *max_inhibitory_concentration*, but its sibling *inhibitory_concentration* and the latter's child *Max_inhibitory_concentration* do not have this object property and are in *Entity* (49). Furthermore the last class name is identical to the object property name. The two redundant classes *inhibitory_concentration* and *Max_inhibitory_concentration* should be removed. The object property *max_inhibitory_concentration* should be removed and replaced by a new data property *concentrationValue* (domain: Concentration, range: float) defined for *Concentration* and inherited to its descendants to store the concentration value for all the various types of concentrations.

Two of the *child-of* relationships not directed to *Entity* raise questions: Why does it hold that a *Drug* IS-A *Source* and why is it true that *Organism* IS-A *Experimental Factor*? *Source* has two children *Natural* and *Synthetic*, which should be renamed *Natural source* and *Synthetic source*, and the assertion *Drug* IS-A *Synthetic source* should be added. Regarding the second *child-of* relationship, the problem was not yet resolved and is considered future work.

It is worth noting the unique area with three partial-areas and the seven classes in the partial-area *Molecule*. One child of *Molecule*, namely *Target*, should be renamed *Biological target* according to its definition. The five children of *Target*, e.g., *Lipid*, *Protein* and *Sugar* are macromolecules. Hence, a class *Macromolecule* should be introduced as child of *Molecule* and become the parent of its current five children. The modeling of the relationships between them and *Biological Target* will be considered in future work.

There are also issues regarding the three children of *Experimental factor*, another partial-area in this area (defined by the object properties description, title), left for future consideration. The curators of CanCo (Dimitrios Zeginis and Konstantinos Tarabanis) [122] have implemented all the above changes, which were incorporated in a new release (version number 0.3) of CanCo in BioPortal. As can be seen, the anomalies found in the CanCo taxonomy helped to detect problems in CanCo's modeling.

9.5 Discussion

The purpose of this chapter was to introduce a family-based QA framework for ontologies, which will enable broad applicability and substantial savings by automating the derivation

of abstraction networks in support of QA work. According to the literature (see, e.g., [128]), current QA techniques for ontologies and taxonomies, target only single ontologies or terminologies. The new framework suggests methods that work uniformly across families of ontologies. This chapter provided a proof of concept for the feasibility of such a framework. It was discussed how families are defined and seven disjoint families covering 186 ontologies of the BioPortal repository were introduced. The definition of these families together with the classification of the 186 ontologies into them provide a proof of concept for the existence of such a framework. Alternative groupings of families are possible, as described in Section 9.5.1.

The two operational aspects of this framework are (1) the automatic family-based uniform derivation of abstraction networks and (2) the utilization of abstraction networks in characterizations of sets of classes with a high likelihood of errors, recognizable by various aspects and anomalies in the appearance of the abstraction networks for a given family of ontologies. Concentrating QA efforts on such anomalous sets will increase the yield of QA work in terms of the ratio of problems found and resolved, to the number of classes reviewed.

For each ontology from Families 2–4 (having object properties), the prototype derivation and display tool BLUOWL is available to automatically create an abstraction network. This has been demonstrated for CanCo, as well as for OCRe [61] and Top-Manelas [129], all in Family 2. An abstraction network for GO (Family 3) was reported on the SABOC website [130]. Abstraction networks for the Family 4 members SDO [62] and DDI [128] have been generated. For Family 1, the BLUSNO tool [131] constructs taxonomies for SNOMED hierarchies [60, 64, 65].

The families of Tables 9.2 and 9.3 were intentionally designed to be disjoint since, for ontologies with object properties, the proper taxonomies will typically have sufficient granularity [62] to support QA. The other observed structural features, e.g., data properties, are not needed for the design of abstraction networks for QA. For the families of Table 9.3 without object properties, an abstraction network can be derived for an ontology with only data properties (Family 6) in a manner similar to that for an ontology with only object properties, so that classes with the same set of data properties are grouped into one area.

Ontologies having no object properties but having some concepts with multiple parents (Family 5) pose difficulties for abstraction network derivation. Due to the lack of object properties, an area taxonomy cannot be derived. A possible alternative abstraction paradigm might exploit overlapping subhierarchies resulting from concepts with multiple parents.

While extensive future work is needed for completing the framework, the presented results show that family-based automatic abstraction network derivation is possible. In the future, BLUOWL will be implemented with a separate module for each family. This tool will be made available for download so that ontology curators can easily derive abstraction networks for their ontologies, on demand.

Regarding family-based QA work, note that for the two ontologies of Family 1, SNOMED CT and NCIt, small sets represented by nodes of the partial-area taxonomies were shown experimentally to have high likelihoods of errors [59, 60]. For OCRe, SDO, and CanCo large partial-areas of the taxonomy were shown to indicate higher concentrations of errors [61, 62]. These examples demonstrate the viability of the QA aspect of the framework introduced in this chapter. However, the properties that make a

partial-area suspicious vary between families and presumably also between members of a family. This is left for future investigation.

9.5.1 Future Work

In this chapter, all the families are disjoint, i.e., each ontology is classified into only one family. While families are defined as disjoint for this study, an ontology may exhibit several structural features, e.g., domain-defined object properties, a hierarchy with multiple parents, and the presence of data properties, as demonstrated in Table 9.1. If an ontology has several structural features, then there are several alternatives how to model it. For example, different types of abstraction networks can be derived, providing additional, independent QA options. If one abstraction network does not work well for QA, others may. If, for example, an ontology has few object properties, yielding too coarse an abstraction network, as was the case for the domain-defined taxonomy for SDO [62], the object properties can be combined with data properties to derive a richer taxonomy. The discovery of further families of ontologies is expected and will be part of future research. One can define a sub-family of ontologies with restriction-defined object properties and classes with multiple parents. Another example for a sub-family would consist of ontologies with few domain-defined object properties and many data properties, e.g., the DermLex BioPortal ontology. In future research, such definitions of sub-families will be explored. The abstraction network derivation and QA for this family-based QA framework will be further developed.

9.6 Conclusions

In this chapter, structural features of 186 BioPortal ontologies were identified, that enabled the classification of these ontologies into families. Using this family classification, it is

possible to derive abstraction networks for whole families of ontologies, which enables a uniform QA methodology for these similar ontologies. A QA review of the Cancer Chemoprevention Ontology (CanCo) was used to illustrate the benefits of a uniform family-based QA methodology.

The research work described in this chapter has been published in the American Medical Informatics Association 2013 Annual Symposium [48].

REFERENCES

[1] US Department of Health and Human Services, Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Records Technology. [cited May 21, 2013]; Available from: http://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17210.pdf

[2] Health IT Adoption Programs. [cited April 8, 2013]; Available from: http://www.healthit.gov/policy-researchers-implementers/health-it-adoption-programs

[3] eClinicalWorks Homepage. [cited October 19, 2013]; Available from: http://www.eclinicalworks.com/

[4] Allscripts Homepage. [cited October 19, 2013]; Available from: http://www.allscripts.com/

[5] Epic Homepage. [cited October 19, 2013]; Available from: http://www.allscripts.com/

[6] DXplain Homepage. 2013 [cited November 25, 2013]; Available from: http://dxplain.org/dxp2/dxp.asp

[7] DiagnosisPro Homepage. [cited October 19, 2013]; Available from: http://en.diagnosispro.com/

[8] VisualDX Homepage. 2013 [cited November 25, 2013]; Available from: http://www.visualdx.com/purchase-redeem/institutional-license

[9] Harvard Pilgrim Health Care Homepage. [cited October 19, 2013]; Available from: https://www.harvardpilgrim.org

[10] Delaware Health Information Network Homepage. [cited October 19, 2013]; Available from: http://www.dhin.org/

[11] Indiana Health Information Exchange Homepage. [cited October 19, 2013]; Available from: http://www.ihie.org/

[12] CareCloud Homepage. [cited October 19, 2013]; Available from: http://www.carecloud.com/

[13] ADS Homepage. [cited October 19, 2013]; Available from: http://www.adsc.com/

[14] NueMD Homepage. [cited October 19, 2013]; Available from: http://www.nuesoft.com/nuemd/medical-software/medical-billing-software.html

[15] Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008:67-79.

[16] White Paper: Interoperability is the Future - 3 Questions Critical to Achieving Healthcare Interoperability. [cited April 3, 2013]; Available from: http://www.interoperabilityshowcase.org/himss13/documents/axway_primer_interoperability_is_the_future_en.pdf

[17] Woolf SH, Kuzel AJ, Dovey SM, Phillips RL, Jr. A string of mistakes: the importance of cascade analysis in describing, counting, and preventing medical errors. Ann Fam Med. 2004 Jul-Aug;2(4):317-26.

[18] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.

[19] Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):1-11.

[20] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993 Apr;81(2):217-22.

[21] Tuttle MS, Sherertz DD, Olson NE, et al. Using META-1, the first version of the UMLS Metathesaurus. Proc 14th Annu Symp Comput Appl Med Care; 1990. p. 131-5.

[22] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods Inf Med. 1995 Mar;34(1-2):193-201.

[23] McCray AT. An upper-level ontology for the biomedical domain. Comp Funct Genomics. 2003;4(1):80-4.

[24] McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. Proc 14th Annu Symp Comput Appl Med Care. Los Alamitos, CA; 1990. p. 126-30.

[25] Gu HH, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med. 2004 May;31(1):29-44.

[26] Gu HH, Hripcsak G, Chen Y, Morrey CP, Elhanan G, Cimino JJ, Geller J, Perl Y. Evaluation of a UMLS Auditing Process of Semantic Type Assignments. Proc AMIA Annu Symp. 2007:294-8.

[27] Gu HH, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. J Am Med Inform Assoc. 2000;7(1):66-80.

[28] UMLS Reference Manual. [cited March 4, 2013]; Available from: http://www.ncbi.nlm.nih.gov/books/NBK9676/

[29] Geller J, Gu HH, Perl Y, Halper M. Semantic refinement and error correction in large terminological knowledge bases. Data and Knowledge Engineering. 2003;45(1):1-32.

[30] Chen Y, Gu HH, Perl Y, Geller J. Structural group-based auditing of missing hierarchical relationships in UMLS. J Biomed Inform. 2009 Jun;42(3):452-67.

[31] Chen Y, Gu HH, Perl Y, Geller J, Halper M. Structural group auditing of a UMLS semantic type's extent. J Biomed Inform. 2009 Feb;42(1):41-52.

[32] SNOMED CT User Guide. [cited April 2, 2013]; Available from: http://www.ihtsdo.org/fileadmin/user_upload/doc/en_us/ug.html

[33] Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P. The Description Logic Handbook - Theory, Implementation and Applications. Cambridge, UK: Cambridge University Press; 2003.

[34] Wilcke JR, Green JM, Spackman KA, et al. Concerning SNOMED-CT content for public health case reports. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):613; author reply -4.

[35] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED Clinical Terms: overview of the development process and project status. Proc AMIA Annu Symp. 2001:662-6.

[36] Wei D, Halper M, Elhanan G, Chen Y, Perl Y, Geller J, Spackman KA. Auditing SNOMED relationships using a converse abstraction network. Proc AMIA Annu Symp. 2009;2009:685-9.

[37] SNOMED CT Homepage. [cited January 10, 2013]; Available from: http://www.ihtsdo.org

[38] Medicare and Medicaid Programs; Electronic Health Record Incentive Program. CMS-0033-P. RIN 938-AP78. [cited January 10, 2013]; Available from: http://healthit.hhs.gov/portal/server.pt/gateway/

[39] Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. J Am Med Inform Assoc. 2011 Dec;18 Suppl 1:i36-44.

[40] 11th Congress ed. USA, American Recovery and Reinvestment Act of 2009. 2009.

[41] Blumenthal D. Launching HITECH. N Engl J Med. 2010 Feb 4;362(5):382-5.

[42] Musen MA, Noy NF, Shah NH, et al. The National Center for Biomedical Ontology. J Am Med Inform Assoc. 2012 Mar-Apr;19(2):190-5.

[43] OWL Web Ontology Language Overview. [cited January 10, 2013]; Available from: http://www.w3.org/TR/owl-features

[44] Resource Description Framework (RDF). [cited March 3, 2013]; Available from: http://www.w3.org/RDF/

[45] OBO Foundry Princeples. [cited April 4, 2013]; Available from: http://www.obofoundry.org/wiki/index.php/Cateory:Accepted

[46] Geller J, He Z, Perl Y, Morrey CP, Xu J. Rule-based support system for multiple UMLS semantic type assignments. J Biomed Inform. 2013 Feb;46(1):97-110.

[47] He Z, Halper M, Perl Y, Elhanan G. Clinical clarity versus terminological order: the readiness of SNOMED CT concept descriptors for primary care. Proc of the 2nd international workshop on Managing interoperability and compleXity in health systems. Maui, Hawaii, USA: ACM; 2012:1-6.

[48] He Z, Ochs C, Agrawal A, Perl Y, Zeginis D, Tarabanis K, Elhanan G, Halper M, Noy N, Geller J. A Family-Based Framework for Supporting Quality Assurance of Biomedical Ontologies in BioPortal. Proc AMIA Annu Symp. Washington, D.C.; 2013:581-90.

[49] Morrey CP. Auditing the Unified Medical Language System and Enhancing the Refined Semantic Network: Dissertation in the Department of Computer Science, New Jersey Institute of Technology. ; 2009.

[50] Geller J, He Z, Elhanan G. Categorizing the Relationship between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization. Submitted for peer review; 2013.

[51] Geller J, He Z, Chen Y. A Comparative Analysis of M:N Trapezoids between Pairs of Metathesaurus Terminologies. In preparation; 2013.

[52] Morrey CP, Geller J, Halper M, Perl Y. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. J Biomed Inform. 2009 Jun;42(3):468-89.

[53] Gu HH, Elhanan G, Perl Y, Hripcsak G, Cimino JJ, Xu J, Chen Y, Geller J, Morrey CP. A study of terminology auditors' performance for UMLS semantic type assignments. J Biomed Inform. 2012 Dec;45(6):1042-8.

[54] Chen Y, Gu HH, Perl Y, Halper M, Xu J. Expanding the extent of a UMLS semantic type via group neighborhood auditing. J Am Med Inform Assoc. 2009 Sep-Oct;16(5):746-57.

[55] Chen L, Morrey CP, Gu HH, Halper M, Perl Y. Modeling multi-typed structurally viewed chemicals with the UMLS Refined Semantic Network. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):116-31.

[56] Definition of UMLS Semantic Types. [cited December 5, 2012]; Available from: http://semanticnetwork.nlm.nih.gov/Download/RelationalFiles/SRDEF

[57] Peng Y, Halper MH, Perl Y, Geller J. Auditing the UMLS for redundant classifications. Proc AMIA Symp. 2002:612-6.

[58] Srinivasan S. Personal Communication. 2009.

[59] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc. 2006;13(6):676-90.

[60] Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007 Oct;40(5):561-81.

[61] Ochs C, Agrawal A, Perl Y, Halper M, Tu SW, Carini S, Sim I, Noy N, Musen M, Geller J. Deriving an abstraction network to support quality assurance in OCRe. Proc AMIA Annu Symp. 2012:681-9.

[62] Ochs C, He Z, Perl Y, Arabandi S, Halper M, Geller J. Refining the Granularity of Abstraction Networks for the Sleep Domain Ontology. Proc the 4th International Conference on Biomedical Ontology. Montreal, QC, Canada; 2013:84-9.

[63] Wei D, Bodenreider O. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies - a case study in SNOMED CT. Stud Health Technol Inform. 2010;160(P2):1070-4.

[64] Wang Y, Halper M, Wei D, Gu H, Perl Y, Xu J, Elhanan G, Chen Y, Spackman KA, Case JT, Hripcsak G. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. J Biomed Inform. 2012 Feb;45(1):1-14.

[65] Wang Y, Halper M, Wei D, Perl Y, Geller J. Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. J Biomed Inform. 2012 Feb;45(1):15-29.

[66] IHTSDO. SNOMED CT and LOINC to be linked by cooperative work. 2013 [cited October 11, 2013]; Available from: http://www.ihtsdo.org/about-ihtsdo/governance-and-advisory/harmonization/loinc/

[67] Weng C, Fridsma DB. A call for collaborative semantics harmonization. Proc AMIA Annu Symp. 2006:1142.

[68] Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: a case study. J Biomed Inform. 2007 Jun;40(3):353-64.

[69] Tao C, Solbrig HR, Chute CG. CNTRO 2.0: A Harmonized Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives. AMIA Summits Transl Sci Proc. 2011:64-8.

[70] Bodenreider O. Strength in numbers: exploring redundancy in hierarchical relations across biomedical terminologies. Proc AMIA Annu Symp. 2003:101-5.

[71] Kumar A, Smith B, Novotny DD. Biomedical informatics and granularity. Comp Funct Genomics. 2004;5(6-7):501-8.

[72] Bittner T, Smith B. A Theory of Granular Partitions. In: Duckham M, Goodchild M, Worboys M, editors. Foundations of Geographical Information Science. London: Taylor & Francis; 2002.

[73] Sun P, Zhang S. Identifying Granularity Differences between Large Biomedical Ontologies through Rules. Proc AMIA Annu Symp 2010. p. 927-31.

[74] Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. Genome Biol. 2005;6(3):R29.

[75] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007 Feb;40(1):30-43.

[76] Schulz S, Boeker M, Stenzhorn H. How Granularity Issues Concern Biomedical Ontology Integration. In Proceedings of the International Congress of the European Federation for Medical Informatics (MIE 2008). Gotenburg, Sweden; 2008: 863-68.

[77] Rector A, Rogers J, Bittner T. Granularity, scale and collectivity: when size does and does not matter. J Biomed Inform. 2006 Jun;39(3):333-49.

[78] Mortensen JM, Horridge M, Musen MA, Noy NF. Applications of ontology design patterns in biomedical ontologies. Proc AMIA Annu Symp. 2012;2012:643-52.

[79] Bail S, Horridge M, Parsia B, Sattler U. The Justificatory Structure of the NCBO BioPortal Ontologies. Proc International Semantic Web Conference. Bonn, Germany; 2011:67-82.

[80] Quesada-Martínez M, Fernández-Breis JT, Stevens R. Extraction and analysis of the structure of labels in biomedical ontologies. Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems. Maui, HI; 2012:7-16.

[81] Ghazvinian A, Noy N, Jonquet C, Shah NH, Musen M. What Four Million Mappings Can Tell You about Two Hundred Ontologies. The Semantic Web - ISWC 2009, Lecture Notes in Computer Science; 2009:229-42.

[82] Ghazvinian A, Noy N, Musen M. How orthogonal are the OBO Foundry ontoloiges? J Biomed Semantics. 2011;Suppl 2(S2).

[83] Vescovo CD, Gessler D, Klinov P, et al. Decomposition and Modular Structure of BioPortal Ontologies. Proc International Semantic Web Conference. Bonn, Germany; 2011:146-61.

[84] Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, Spackman KA. Analysis of error concentrations in SNOMED. Proc AMIA Annu Symp. 2007:314-8.

[85] Wang Y, Wei D, Xu J, Elhanan G, Perl Y, Halper M, Chen Y, Spackman KA, Hripcsak G. Auditing complex concepts in overlapping subsets of SNOMED. AMIA Annu Symp Proc. 2008:273-7.

[86] Chen Y, Gu H, Perl Y, Geller J. Overcoming an obstacle in expanding a UMLS semantic type extent. J Biomed Inform. 2012 Feb;45(1):61-70.

[87] NLM. The Future of the UMLS Semantic Network. 2005 [cited October 8, 2013]; Available from: http://mor.nlm.nih.gov/snw/

[88] Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

[89] Lomax J, McCray AT. Mapping the Gene Ontology into the Unified Medical Language System. Comp Funct Genomics. 2004;5(4):354-61.

[90] Yu H, Friedman C, Rhzetsky A, Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA Symp. 1999:181-5.

[91] Cohen B, Chen Y, Perl Y. Updating the genomic component of the UMLS Semantic Network. AMIA Annu Symp Proc. 2007:150-4.

[92] Perl Y, Chen Z, Halper M, Geller J, Zhang L, Peng Y. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. J Biomed Inform. 2002 Jun;35(3):194-212.

[93] Clark KL. Negation as failure. In: M.L. Ginsberg, editor. Readings in nonmonotonic reasoning. San Francisco, CA: Morgan Kaufmann Publishers Inc. ; 1987. p. 311-25.

[94] Gu HH, Halper M, Geller J, Perl Y. Benefits of an object-oriented database representation for controlled medical terminologies. J Am Med Inform Assoc. 1999;6(4):283-303.

[95] Morrey CP, Perl Y, Halper M, Chen L, Gu HH. A chemical specialty semantic network for the Unified Medical Language System. J Cheminform. 2012;4(1):9.

[96] Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. J Am Med Inform Assoc. 1998 Nov-Dec;5(6):503-10.

[97] Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc. 2006 May-Jun;13(3):277-88.

[98] Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):486-94.

[99] IHTSDO. SNOMED Clinical Terms User Guide-January 2010 International Release. [cited 2012 June 28]; Available from: http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Publications/doc_UserGuide_Current-en-US_INT_20100131.pdf

[100] Merriam-Webster Dictionary. July 2004 ed.

[101] Nash SK. Nonsynonymous synonyms: correcting and improving SNOMED CT. Proc AMIA Annu Symp. 2003:949.

[102] IHTSDO. SNOMED Clinical Terms® Technical Reference Guide-January 2009 International Release. [cited 2012 June 28]; Available from: http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/SNOMED_CT_Publications/SNOMED_CT_Technical_Reference_Guide_20090131.pdf

[103] NLM. The CORE Problem List Subset of SNOMED CT. [cited 2012 June 24]; Available from: http://www.nlm.nih.gov/research/umls/Snomed /core_subset.html

[104] NLM. UMLS Enhanced VA/KP Problem List Subset of SNOMED CT. [cited 2012 June 28]; Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_problem_list.html

[105] CliniClue. [cited June 15, 2012]; Available from: http://www.cliniclue.com/home

[106] Dolin RH, Mattison JE, Cohn S, et al. Kaiser Permanente's Convergent Medical Terminology. Stud Health Technol Inform. 2004;107(Pt 1):346-50.

[107] Lincoln MJ, Brown SH, Nguyen V, et al. U.S. Department of Veterans Affairs Enterprise Reference Terminology strategic overview. Stud Health Technol Inform. 2004;107(Pt 1):391-5.

[108] Certified Health IT Product List. [cited January 10, 2013]; Available from: http://onc-chpl.force.com/ehrcert/CHPLHome

[109] Lee DH, Lau FY, Quan H. A method for encoding clinical datasets with SNOMED CT. BMC Med Inform Decis Mak. 2010;10:53.

[110] Liu J, Lane K, Lo E, Lam M, Truong T, Veillette C. Addressing SNOMED CT implementation challenges through multi-disciplinary collaboration. Stud Health Technol Inform. 2010;160(Pt 2):981-5.

[111] Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. J Biomed Inform. 2009;43(3):274-82.

[112] Shvaiko P, Euzenat J. Ontology Matching: State of the Art and Future Challenge. Knowledge and Data Engineering, IEEE Transactions on. 2013;25(1):158-76.

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

[114] Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naive vs. formal. Proc AMIA Annu Symp. 2005:550-4.

[115] Courtot M, Gibson F, Lister AL, Malone J. MIREOT: The Minimum Information to Reference an External Ontology Term. International Conference on Biomedical Ontology. Buffalo, NY; 2009. p. 87-90.

[116] Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):12-6.

[117] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

[118] Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform. 2008 Jan;9(1):75-90.

[119] Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform. 2006 Sep;7(3):256-74.

[120] Gu HH, Cimino JJ, Halper M, Geller J, Perl Y. Utilizing OODB schema modeling for vocabulary management. Proc AMIA Annu Fall Symp. 1996:274-8.

[121] BioPortal. Cancer Chemoprevention Ontology. 2012 [cited January 4, 2013]; Available from: http://bioportal.bioontology.org/ontologies/3030

[122] Zeginis D, Hasnain A, Loutas N, Deus HF, Fox R, Tarabanis K. A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web Journal. 2013.

[123] Geller J, Ochs C, Perl Y, Xu J. New abstraction networks and a new visualization tool in support of auditing the SNOMED CT content. AMIA Annu Symp Proc. 2012;2012:237-46.

[124] Ochs C, Geller J, Perl Y. BLUSNO: A Biomedical Layout Utility for the OWL-Based Ontologies. In preparation.

[125] Shearer R, Motik B, Horrocks I. HermiT: a highly-efficient OWL reasoner. . Proc the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008); 2008.

[126] BioPortal. List of ontologies in BioPortal. [cited January 4, 2013]; Available from: http://bioportal.bioontology.org/ontologies

[127] BioPortal. Basic Formal Ontology. [cited January 4, 2013]; Available from: http://bioportal.bioontology.org/ontologies/1332

[128] He Z, Ochs C, Soldatova L, Perl Y, Arabandi S, Geller J. Auditing Redundant Import in Reuse of a Top Level Ontology for the Drug Discovery Investigations

Ontology. International Workshop on Vaccine and Drug Ontology Studies. Montreal, QC, Canada; 2013.

[129] BioPortal. Menelas Project Top-Level Ontology. [cited December 2, 2013]; Available from: http://bioportal.bioontology.org/ontologies/TOP-MENELAS

[130] Figures of the taxonomies of the ontologies. [cited March 3, 2013]; Available from: http://cs.njit.edu/~oohvr/SABOC/figures.php

[131] Geller J, Ochs C, Perl Y, Xu J. New abstraction networks and a new visualization tool in support of auditing the SNOMED CT content. Proc AMIA Annu Symp. 2012;2012:237-46.