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

NJIT
New Jersey's Science &
Technology University

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

STRUCTURAL ANALYSIS AND AUDITING OF SNOMED HIERARCHIES USING ABSTRACTION NETWORKS

by
Yue Wang

SNOMED is one of the leading healthcare terminologies being used worldwide. Due to its sheer volume and continuing expansion, it is inevitable that errors will make their way into SNOMED. Thus, quality assurance is an important part of its maintenance cycle.

A structural approach is presented in this dissertation, aiming at developing automated techniques that can aid auditors in the discovery of terminology errors more effectively and efficiently. Large SNOMED hierarchies are partitioned, based primarily on their relationships patterns, into concept groups of more manageable sizes. Three related abstraction networks with respect to a SNOMED hierarchy, namely the *area taxonomy*, *partial-area taxonomy*, and *disjoint partial-area taxonomy*, are derived programmatically from the partitions. Altogether they afford high-level abstraction views of the underlying hierarchy, each with different granularity. The area taxonomy gives a global structural view of a SNOMED hierarchy, while the partial-area taxonomy focuses more on the semantic uniformity and hierarchical proximity of concepts. The disjoint partial-area taxonomy is devised as an enhancement of the partial-area taxonomy and is based on the partition of the entire collection of so-called *overlapping concepts* into singly-rooted groups.

The taxonomies are exploited as the basis for a number of systematic auditing regimens, with a theme that complex concepts are more error-prone and require special attention in auditing activities. In general, group-based auditing is promoted to achieve a

more efficient review within semantically uniform groups. Certain concept groups in the different taxonomies are deemed “complex” according to various criteria and thus deserve focused auditing. Examples of these include strict inheritance regions in the partial-area taxonomy and overlapping partial-areas in the disjoint partial-area taxonomy.

Multiple hypotheses are formulated to characterize the error distributions and ratios with respect to different concept groups presented by the taxonomies, and thus further establish their efficacy as vehicles for auditing. The methodologies are demonstrated using SNOMED’s Specimen hierarchy as the test bed. Auditing results are reported and analyzed to assess the hypotheses. With the use of the double bootstrap and Fisher’s exact test (two-tailed), the aforementioned hypotheses are confirmed. Auditing on various complex concept groups based on the taxonomies is shown to yield a statistically significant higher proportion of errors.

STRUCTURAL ANALYSIS AND AUDITING OF SNOMED HIERARCHIES USING ABSTRACTION NETWORKS

by
Yue Wang

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

Department of Computer Science

May 2012

Copyright © 2012 by Yue Wang

ALL RIGHTS RESERVED

APPROVAL PAGE

STRUCTURAL ANALYSIS AND AUDITING OF SNOMED HIERARCHIES USING ABSTRACTION NETWORKS

Yue Wang

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

Dr. Michael Halper, Dissertation Co-Advisor Date
Director, Information Technology Program, NJIT

Dr. James Geller, Committee Member Date
Professor, Computer Science Department, NJIT

Dr. Narain Gehani, Committee Member Date
Professor, Computer Science Department, NJIT

Dr. Kent A. Spackman, Committee Member Date
Chief Terminologist, IHTSDO

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

BIOGRAPHICAL SKETCH

Author: Yue Wang

Degree: Doctor of Philosophy

Date: May 2012

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2012

- Master of Science in Health Informatics,
  University of Texas-Houston Health Science Center, Houston, TX, 2002

- Master of Science in Computer Science,
  Shanghai Jiao Tong University, Shanghai, P. R. China, 2000

- Bachelor of Science in Computer Engineering,
  Soochow University, Suzhou, P. R. China, 1997

Major: Computer Science

Presentations and Publications:

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.

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

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.

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

To my loving husband, Yesheng Li, who has been supportive in this long journey.

To my wonderful children, Raymond and Jaylen, who are the joy of my life.

To my parents, 王振刚 *and* 甘盛莲*, who have always believed in me and encouraged me in all my endeavors.*

ACKNOWLEDGMENT

I would like to thank the following people who helped make this dissertation possible.

I am heartily thankful to my co-advisor, Dr. Yehoshua Perl, for his excellent guidance, patience, and persistence over the years. I will benefit from what I learnt from him my whole life.

I owe my deepest gratitude to my co-advisor, Dr. Michael Halper, who has guided my research and academic writing for the past nine years. Thanks for being open and supportive the whole time.

I am also sincerely grateful to my dissertation committee, Dr. James Geller, Dr. Gai Elhanan, and Dr. Narain Gehani, for their guidance and suggestions. Special thanks go to Dr. Kent A. Spackman, whose knowledge, support and guidance have made this study successful.

Finally, I would like to thank all the people who have lent me help and supported me in innumerable ways during my graduate study.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Motivation

Electronic Health Record (EHR) systems have been widely used in the healthcare industry in pursuit of reduced medical errors, higher-quality care, and improved efficiency. The basis for these products is a standard terminology, which provides a consistent way to index, store, retrieve, and aggregate clinical data across specialties and sites of care. The primary purpose of such a terminology is to support the effective clinical data recording and information exchange so as to improve patient care.

The Systematized Nomenclature of Medicine – Clinical Terms (“SNOMED” for short, hereafter) [1], one of the leading biomedical terminologies, is well structured, highly computerized, and has many merits that make it superior to its peers. This is evidenced, for example, by the fact that it is slated to become an integral component of standardization in health information technology [2]. In one particular application, the encoding of patients’ problems in EHRs by concepts derived from SNOMED has been proposed as part of the requirements for “meaningful use” of such systems [2].

However, due to SNOMED’s large volume and inherent complexity, it is unavoidable that errors will find their way to this large knowledge base, particularly as it continues to expand. As SNOMED underlies decision-support systems, clinical patient records, health care administrative systems, etc., errors in SNOMED may propagate to errors in these systems, which in turn may result in endangering the life or quality of life of a patient.

Given SNOMED’s expanding content and attendant complexity, quality assurance is a critical task facing SNOMED’s maintenance personnel. To this end, the International Health Terminology Standards Development Organisation (IHTSDO) [3] formed the Quality Assurance Committee, which supports the mission of the IHTSDO by advising on issues related to the quality of SNOMED, the quality of related standards for which the IHTSDO has responsibility, and the quality of services provided by the IHTSDO. It is in this committee that SNOMED’s content undergoes a clinical quality assurance process prior to each release. More importantly, automated and semi-automated methodologies that can aid editors in this endeavor and enhance the efficiency and efficacy of SNOMED auditing are invaluable.

The objective of this research is to investigate how computer science techniques can be applied to assist the quality assurance of SNOMED. The structural aspects of the SNOMED hierarchies and their constituent concepts are studied. Three high-level abstraction networks, namely *area taxonomy*, *partial-area taxonomy*, and *disjoint partial-area taxonomy*, are derived programmatically based on analyses of a SNOMED hierarchy’s attribute relationships and their patterns of inheritance. The three taxonomies complement each other in terms of granularity of display, each with different focus. Altogether they serve as a multi-level abstraction of a SNOMED hierarchy, providing a more effective and efficient way for orientation and assessment. Multiple auditing methodologies that make use of these taxonomies are presented in this dissertation. These taxonomy-based auditing regimens are considered semi-automatic, as the taxonomies can aid an auditor by automatically identifying concepts that deserve attention. Importantly, many concept errors were found manifested themselves as structural anomalies at the taxonomy level, and thus the taxonomies proved to be effective building blocks for automated auditing regimens.

1.2 Background and Literature Review

1.2.1 SNOMED

SNOMED [4-6] is a clinical terminology developed as a joint venture between the College of American Pathologists (CAP) and the UK's National Health Service (NHS). It was formed by merging, expanding, and restructuring an earlier version of SNOMED (i.e., SNOMED RT) and the UK's Clinical Terms Version 3 (CTV3). In 2007, the SNOMED intellectual property rights were transferred from the CAP to the IHTSDO.

SNOMED's concepts are organized in 19 top-level hierarchies, each with a unique root called a top-level concept. Above all these top-level concepts sits a single concept called SNOMED Concept, which serves as the root of the entire terminology. Each concept is a descendant of SNOMED Concept via a sequence of IS-A (subsumption) relationships passing through exactly one top-level concept.

Descriptions are the terms, or names, assigned to each of SNOMED's concepts. A given concept has one or more associated descriptions. One of them is called the "Fully Specified Name" (FSN), which is a unique phrase that describes a concept in a way that is intended to be unambiguous. All concepts have one description which is designated as a "preferred term" for each language edition. (The preferred term is different for UK English, US English, and, of course, Spanish.) Many concepts have alternative descriptions called "synonyms."

Most of SNOMED's top-level hierarchies represent broad groupings of clinically related concepts. There are Clinical Finding, Procedure, Body Structure, Organism, Pharmaceutical/Biologic Product, etc. Three of the hierarchies, namely, Linkage Concept, Qualifier Value, and Special Concept, serve more specific structural roles in the terminology.

Relationships are the connections between concepts in SNOMED, with every concept having at least one relationship to another concept. Relationships in SNOMED are unidirectional, extending from a source concept to a target concept. There are two general kinds of relationships in SNOMED:

1. IS-A relationships (already noted above), that form the basis of the hierarchies. Each connects a more specific concept (a child) to a more general concept (a parent).

2. Attribute relationships, that characterize and define concepts. Each can take on values (targets) only from a prescribed top-level hierarchy.

A particular attribute relationship comprises its source concept, its relationship type (defined as a separate SNOMED concept in its own right), and a value (another concept). These three together are called the "Object-Attribute-Value" (OAV) triplet. For brevity, "attribute relationship" will be referred to as "relationship," while "IS-A relationship" will be referred to as "IS-A" hereafter.

Relationships in SNOMED are the major interests when the partitioning techniques are applied and abstraction networks are constructed for auditing. These relationships serve in definitional capacities. For example, the concept *Ear problem* (in the Clinical Finding hierarchy) has the relationship *finding site* to the concept *Ear structure* (in the Body Structure hierarchy) specifying that *Ear structure* is the site of *Ear* *problem*. Some hierarchies introduce multiple relationships. For example, the Specimen hierarchy defines five kinds of relationships, namely, *specimen substance*, *specimen procedure*, *specimen source morphology*, *specimen source topography*, and *specimen source identity*.

1.2.2 Abstraction Networks

In the course of extensive research on terminologies and ontologies over the past 20 years, it has become apparent that their maintenance (including auditing) is greatly enhanced by high-level abstraction networks, particularly those derived from partitions, i.e., groupings of concepts into smaller, more manageable collections. SNOMED's designers decided to organize their terminology in 19 top-level hierarchies as of its most recent release (July 2011). With respect to the UMLS, an abstraction feature was considered paramount and the Semantic Network was thus built as one of its fundamental knowledge sources [7, 8].

A refined Semantic Network (SN) of the UMLS was presented in [9], which offers a partition of the UMLS Metathesaurus into disjoint sets of concepts with similar semantics, not offered by the SN. Later work went even further with a proposal for an additional layer of abstraction, a partition of the SN's semantic types into various subject areas [10, 11]. The notion of metaschema of the SN [12, 13] was introduced as another form of additional level of abstraction. As a matter of fact, most of the papers in a special issue of the Journal of Biomedical Informatics on Structural Issues in UMLS Research [14] utilize the interplay between the SN and the Metathesaurus [15, 16] in one way or another. An abstraction network for the Medical Entities Dictionary (MED) [17] was presented that partitions it into disjoint sets of concepts of similar structure and semantics in [18, 19]. The usefulness of the schema for structural orientation and auditing was demonstrated in [20].

Beyond the field of medical informatics, the Suggested Upper Merged Ontology (SUMO) [21, 22], developed toward the IEEE Standard Upper Ontology, and the mapping of WordNet [23] to SUMO [24] have been conceived in this spirit of abstraction.

1.2.3 Terminology Auditing

Auditing large terminologies is a serious challenge facing the biomedical informatics community. Terminologies are typically huge in size and have high complexity, making comprehensive audits very difficult— indeed, overwhelming—tasks.

A variety of systematic auditing techniques have been proposed and applied to SNOMED. Its conceptual coverage and its completeness have been assessed using comparative approaches involving external sets of clinical terms [25-27]. An evaluation of the semantic completeness of SNOMED's content has also been done using a formal concept analysis (FCA)-based model [28]. Following that work, a highly-scalable approach was utilized to determine how well SNOMED conformed to a lattice structure and to suggest possible content extensions [29].

Lexical information (specifically, term substrings) was used to detect potential classification omissions [30]. In other work, lexical analysis of SNOMED concepts' textual descriptions has yielded a large collection of underspecified concepts and possibilities for refining SNOMED's content [31]. Another lexical approach has identified a variety of inconsistencies between SNOMED terms and the underlying logical modeling of seemingly similar concepts [32]. Inconsistent usage of the words “and” and “or” in SNOMED terms has been studied [33].

Ontological and linguistic techniques were utilized to identify duplicates and redundancy [34, 35]. SNOMED has been analyzed to determine how well its hierarchical relations adhere to four basic ontological principles [36, 37]. Since SNOMED is based on a description logic (DL) formalism, it is amenable to algorithms developed in the context of DL representations for the detection of terminological inconsistencies [38] and synonymy [39]. The impact of SNOMED revisions was assessed by investigating the manual mappings between a proprietary interface terminology to two versions of SNOMED [40]. A comprehensive review of auditing methodologies used for SNOMED are presented in [41] along with a useful general glossary pertaining to auditing.

In general, the typically limited availability of auditing resources makes it imperative to develop systematic techniques that focus efforts on concepts or groups of concepts that are likely to have higher rates of errors. In this way, a better return, measured in the number of errors found, can be expected for a given amount of auditing work.

Many important terminologies and terminological systems, aside from SNOMED and the others mentioned above, have been the focus of systematic auditing regimens. In fact, a special issue of JBI [14] has been devoted exclusively to terminology auditing methodologies. In [41] in that issue, a framework was introduced to help classify the large body of disparate techniques based on various criteria. For example, distinctions were made based on the kind of terminology attribute that was the focus of the audit, e.g., terms and concepts vs. semantic classification. Moreover, the methodologies were categorized according to their uses of various knowledge and their levels of automation in the identification of problems. According to the classification, the methodologies being presented in this dissertation can be described as “automated systematic.”

Some of the methodologies surveyed in [41] that were designated automated systematic involved some kind of rule specification. For example, the work in [42] used rules to assess certain uniqueness constraints in Read Codes. In the context of the UMLS, a search for concept redundancy was aided by constraints on semantic types [43]. The algorithm [44] for finding all redundant semantic-type assignments is based on a rule for the UMLS Semantic Network [45]. Concept redundancy was also addressed in SNOMED with the use of rules based on a mapping to LinKBase, a medical ontology [35]. A number of automated systematic methods have exploited DL representations of terminologies. The methodology of [39] is such an example.

The methodologies presented in this dissertation do not utilize any DL classifier functions or any features of SNOMED’s underlying DL framework, except for its systematic definition of relationships and their inheritance via the IS-A hierarchy. Instead, a classification of a collection of complex concepts is made and multiple abstraction networks are defined on top of that collection to guide the auditing efforts.

1.2.4 Previous Work on MED and NCIt

In previous work [18, 19] on the MED [17], an abstraction network called a *schema* was introduced. In this dissertation, the fundamental partition techniques are extended to apply to a SNOMED hierarchy, where the resulting taxonomies prove to be a necessary alternative to the schema for MED. As an example, the schema could not accommodate the situation where the same relationship is introduced at multiple, independent points in the terminology's hierarchy. A schema can only accommodate a given relationship introduction at a unique concept [19]. The taxonomy remedies this deficiency (see Section 2.2). The natural progression from the schema to the area taxonomy and on to the augmented partial-area taxonomy is described.

A similar structural abstraction network was presented in [46] in the context of work on auditing the NCI Thesaurus (NCIt) [47]. The partition technique was applied to the NCIt's small Biological Process hierarchy, which consisted of 589 concepts and had seven relationships defined for these concepts. As it happened, the Biological Process hierarchy was effectively a tree structure, where each concept had just one parent. (In fact, only four concepts had more than one parent, and following the feedback of this study the hierarchy was reorganized into a strict tree structure [46].) The tree-structured hierarchy did not require the full scope of taxonomic development that a directed acyclic graph (DAG) terminology does, as is manifested in this dissertation. It was natural, in fact, to proceed from the easier to the harder and first tackle the tree-structured case and only then extend the methodologies to the DAG case, as found in SNOMED.

1.3 Dissertation Overview

Based on analyses of the SNOMED hierarchy's attribute relationships and their patterns of inheritance, this research explores the automated techniques to devise high-level abstraction networks (called *taxonomies*), that facilitate terminology orientation and comprehension. A number of systematic auditing regimens are formulated based on these taxonomies. The effectiveness of the so-called taxonomy-based auditing is demonstrated in multiple hierarchies of SNOMED. This dissertation is organized as follows:

Chapter 2 presents the structural auditing methodologies based on partitioning and abstraction. Automated techniques are developed for partitioning SNOMED into smaller groups of concepts. From the partition, two different abstraction networks, the *area taxonomy* and *partial-area taxonomy* are derived. Multiple systematic auditing methodologies utilizing the taxonomies are presented, and the results garnered from applications of the auditing regimens to SNOMED are used to investigate the concentration of errors among certain types of concept groups.

Chapter 3 further extends the taxonomy paradigm to deal with particularly complex portions of a SNOMED hierarchy, where *overlapping concepts* reside. A new abstraction network, called the *disjoint partial-area taxonomy*, is introduced as a refinement to the partial-area taxonomy, which provides a better high-level view of the tangled portion of a hierarchy, and facilitates orientation and assessment of SNOMED's content. The techniques are demonstrated using the Specimen hierarchy.

Chapter 4 introduces an systematic auditing regimen based on the disjoint partial-area taxonomy presented in Chapter 3. The methodology constitutes a systematic review of the overlapping concepts as determined by their hierarchical ordering within the disjoint partial-area taxonomy. A thorough analysis of errors that are found as a result of auditing the overlapping concepts shows a need for enhancements to the partial-area taxonomy in order to capture a partition into disjoint sets having uniform semantics.

Chapter 5 outlines the future direction of this research.

The research work described in this dissertation has been presented in a number of papers [48-51, 70].

CHAPTER 2

STRUCTURAL METHODOLOGIES FOR AUDITING SNOMED

2.1 Introduction

In this chapter, two high-level abstraction networks are devised based on analyses of a SNOMED hierarchy's attribute relationships and their patterns of inheritance. First a hierarchy's concepts were partitioned into groups, called *areas*, according to their specific attribute relationships. From this partition, an abstraction network, referred to as the *area taxonomy*, affording a summary view of the distribution of the attribute relationships was constructed. Further refinement of areas led to another abstraction network, the *partial-area taxonomy*, which conveyed information about sub-area hierarchical arrangements. In addition to their support for orientation to and comprehension of a SNOMED hierarchy, the two networks have served as the foundation of the formulation of structural methodologies for auditing SNOMED hierarchies.

Multiple auditing regimens that make use of the taxonomies are put forward. The first of these detects errors that have manifested themselves as structural irregularities at the abstract level in the area taxonomy. The second investigates irregularities occurring within the partial-area taxonomy. The third methodology, group-based auditing, is also supported by the partial-area taxonomy, where sets of purportedly similar concepts are reviewed together as a group.

The partitioning, abstraction, and auditing methodologies are demonstrated on the Specimen hierarchy. Errors discovered during the auditing process are reported and analyzed to assess several hypotheses about concentration of errors within various parts of the terminology. Using results garnered from applications of the auditing regimens to SNOMED, an investigation into the concentration of errors among such groups was carried out. Three hypotheses pertaining to the error distributions are put forth. The results support the fact that certain groups presented by the taxonomies show higher error percentages as compared to other groups. This knowledge will help direct auditing efforts to increase their impact.

2.2 Methods

The partitioning methodology presented in this chapter focuses primarily on the sets of relationships exhibited by various concepts. In particular, the similarity and disparity of such sets were used as the basis for partitioning of the terminology. Relationships are given primacy because of their overall definitional importance in terminologies. The reasoning underlying this approach is that dividing with respect to relationships along structural lines yields groups which are also likely to be semantically uniform.

Furthermore, the author seeks a partition into groups of concepts that are semantically cohesive, as defined in terms of having a unique root concept. This provides a second dimension of division and results in two levels of partition granularity. From the various partitions, abstraction networks called *area taxonomies* are derived automatically.

Ordinarily, a concept's relationships are inherited from its parent concepts via the IS-As. However, for each kind of relationship, there is always a top concept in the hierarchy at which it first appears. Such concepts are defined as *introducing concepts*. Unlike the MED [17], SNOMED does not exhibit uniqueness of relationship introduction. As a consequence, the analysis becomes more complex.

2.2.1 Areas and Schemas

The first phase of partitioning focuses on the distribution pattern of relationships in the terminology and is based on the notion of *area*. In the following, *structure of a concept* is used to denote a concept's complete set of relationships.

An *area* is a collection of all concepts with the exact same structure. It can be seen that all areas are disjoint since a concept will belong to one and only one of them. Hence, the areas of a terminology form a partition. S*tructure* with respect to an area is defined to be the structure of its constituent concepts.

A concept is a *root* of its area if all its parent(s) are not in the area. (As a special case, a concept without parents is defined to be a root.) That is, a root is characterized by having parents with different structures. As a consequence of the fact that an introducing concept is the first point at which a given relationship appears, such a concept will be a root of its area. A root is a generalization of all its descendants in an area and thus conveys the overarching semantics of the set.

An area may have one or more roots. Consider the simpler case of a singly rooted area first. In such a case, the root concept neatly conveys the prevailing semantics of the whole area. For this reason, such an area is named after its root.

Look at an abstract example to illustrate these ideas for singly rooted areas. Figure 2.1 shows a terminology fragment with five introducing concepts, A through E, and some other unlabeled concepts that do not introduce any new relationships. All concepts are drawn as rounded rectangles. The unlabeled, thick arrows stand for IS-As among concepts. Other labeled arrows represent the relationships between the two concepts. For example, the arrow from *A* to *C* labeled r_1 means that *A* has a relationship r_1 with *C*. Concepts *A* through *E* introduce the relationships r_1 , r_2 , r_3 , r_3' (the converse of r_3), and r_4, respectively. Note that the children and grandchildren of *A* all exhibit the relationship r_1 (and only that relationship in this terminology fragment) due to inheritance. Therefore, all these are grouped into an area called *A*, after the root, drawn as a box enclosing its constituent concepts. While concept *B* (a great grandchild of *A*) also inherits r_1, it introduces the relationship r_2. Hence, *B* and its descendants exhibit the two relationships r_1 and r_2 and are grouped in area *B*. Similarly, there are the areas *C*, *D*, and *E*.

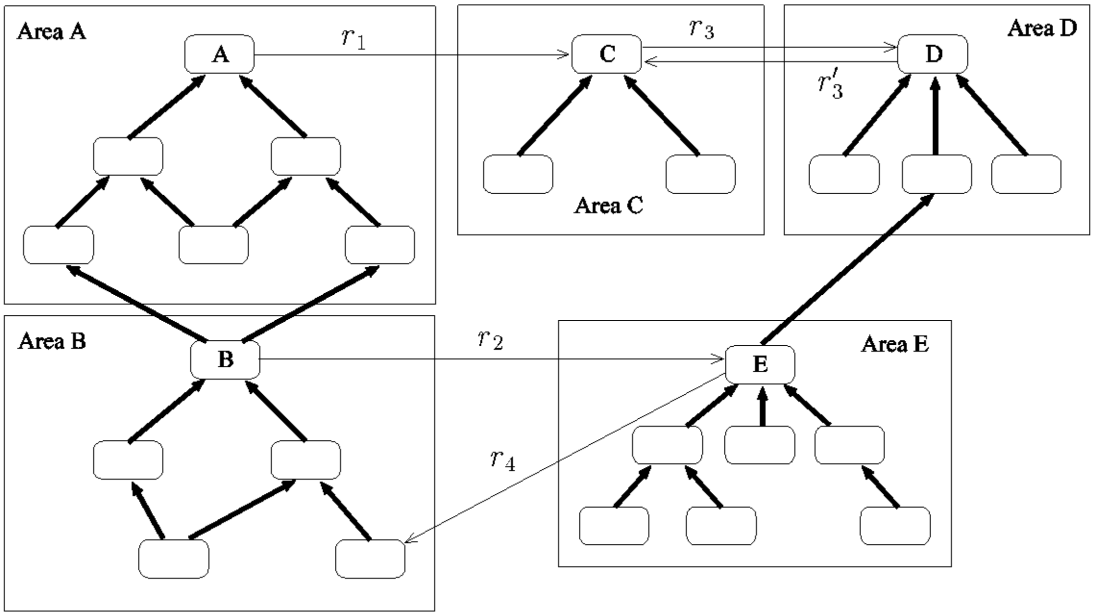


Figure 2.1 Five introducing concepts and associated areas.

Following the analysis of the MED [18, 19], an abstraction network, called an *area schema*, can be automatically derived from the partition into areas as follows. For each area in the partition, a single corresponding node—labeled with the area's name—is defined in the schema. For conciseness, the node in the schema is referred to as an area, too. One area *B* is defined as a *child-of* of another area *A*—and is connected to it via an unlabeled, thick arrow—if the root of area *B* IS-A some concept (not necessarily the root) in area *A*. A relationship *r* is directed from area *X* to area *Y* if the root of area *X* introduces or inherits a relationship *r* whose target is some concept (not necessarily the root) in area *Y*.

Note that for the definitions of both kinds of relationships, the target concept is not required to be the root of its area. Only the source concept of the relationship is required to be the root. This guarantees that all concepts of the area share the relationships of the root (and thus the area) whether those relationships are introduced at the root or inherited from the parent area. The inheritance is enabled at the source of the relationship. The target is inherited with the relationship kind.

Overall, the schema abstractly displays the relationships exhibited by the various areas of similar concepts. It differentiates among the various kinds of concepts based on their differing structures. In particular, the semantics of one group of concepts is clearly distinguished from that of another group if each group exhibits a different structure. The naming convention for nodes makes each introducing concept a focal point. This is warranted because such a concept is where new semantics is introduced, paving the way for the spread of the new knowledge in the portion of the hierarchy below it. Hence, in the area schema, the name of an area expresses the semantics of its concepts, and its structure expresses the structure of its concepts. Thus, the area schema captures both the structure and semantics of a terminology in a compact and abstract way.

Figure 2.2 shows the area schema derived from Figure 2.1, consisting of five areas, two *child-of* relationships, and five relationships. As can be seen, this area schema is a compact representation of the prevailing relationship pattern of the terminology fragment.

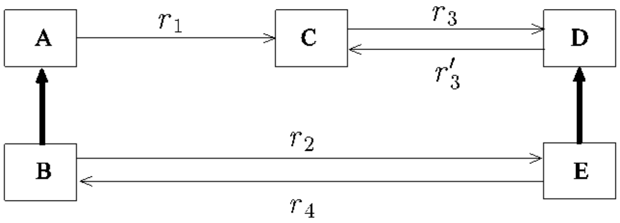


Figure 2.2 Area schema derived from partition in Figure 2.1.

2.2.2 Multi-Rooted Areas

An underlying assumption in the development of the area schema of the MED, guaranteeing singly rooted areas, was that each kind of relationship was introduced at a unique concept in the terminology. However, such a unique introduction point is not a natural requirement for a terminology. SNOMED and other terminologies do not adhere to this.

Under the condition of unique introduction points, all areas are guaranteed to be singly rooted. Multiple introduction points for a given relationship imply that an area can have multiple roots. While the partition of concepts into areas with multiple roots is straightforward, complications do arise with respect to the area schemas. For example, consider Figure 2.3, where there are five areas. The interesting one is on the lower left side and contains two roots, *X* and *Y*, and their respective children. The concept *X* introduces the relationship *r* directed at concept *W*, which happens to be the unique root of its area. The concept *Y* also introduces *r*, which in this case is directed at *Z*, also the unique root of its area. Since *X*, *Y*, and their children all exhibit *r*, they are placed together in an area, as shown in Figure 2.3. Meanwhile, the ancestor *A* of *X* and *Y* introduces the relationship r_1 directed at *B*, the parent of *W* and *Z*. *B* itself introduces no relationship. In addition, *Z* introduces the relationship r_2 targeted at *W*, while the converse relationship r_2' points from *W* to *Z*.

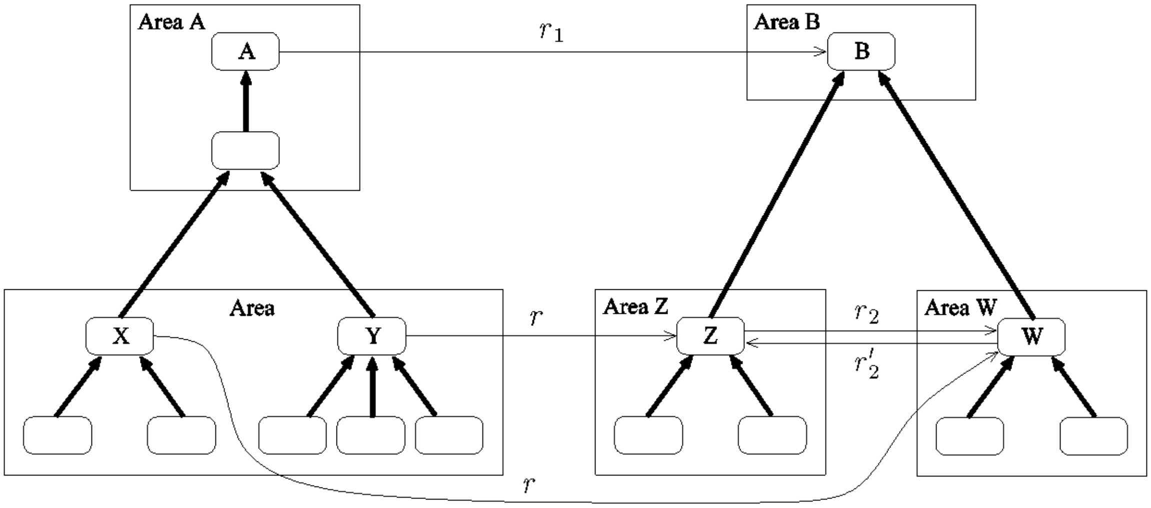


Figure 2.3 A multi-rooted area (with roots *X* and *Y*).

There are problems with an area schema for this configuration. First, what does one call this area rooted at *X* and *Y*? None of the two roots is a generalization of all concepts of the area and thus none of them is appropriate as a name of the area. A second problem is that if the relationship *r* head in multiple directions from this area to areas *W* and *Z*, it conveys the knowledge that for each concept of the area there are two relationships *r*, one to a concept of the area *W* and one to a concept of the area *Z*. But this configuration is not applicable to any concepts in this area. Hence, there is no natural area schema for the terminology fragment of Figure 2.3.

2.2.3 Area Taxonomy

Due to the above problems, an alternative abstract view, called an *area taxonomy*, is introduced. The term “taxonomy” typically denotes a terminology’s entire set of concepts and the IS-As connecting them [52, 53]. The non-IS-A relationships are not included. Similarly, an area taxonomy graphically consists of only the area nodes and hierarchical *child-of* relationships (defined as in the area schema) connecting them. Note that an area taxonomy is acyclic, since a cycle in the area taxonomy will imply a cycle of IS-A in the underlying hierarchy, which is impossible due to the hierarchical nature of IS-As. Relationship arrows other than those for *child-of* are not defined as part of the area taxonomy. The only information pertaining to such relationships is maintained inside an area node in textual form. As a matter of fact, the set of relationships defined for an area node (i.e., its structure) is used as its name to overcome the above area's naming problem for multi-rooted areas. The targets of the relationships are not represented in any way. The area taxonomy ignores the targets of relationships, and instead concentrates on the relationships' names. Hence, it avoids the above two problems that prevented the author from defining an area schema.

Figure 2.4 shows the area taxonomy for the terminology fragment of Figure 2.3, where the rectangles are area nodes and the solid arrows stand for *child-of* relationships between area nodes. The area rooted at *A* in Figure 2.3 is named $\{r_1*\}$, the relationship it introduces. In the text, an area will be denoted by listing its relationship(s) in a pair of braces. The "*" indicates that r_1 is introduced in this particular area. The area with the two roots *X* and *Y* is named after its relationships *r* and r_1. In particular, its name $\{r_1, r*\}$ indicates that this area inherits (via its roots) the relationship r_1 and introduces the new relationship *r*, as again denoted by the "*". The area rooted at *B* in Figure 2.3 exhibits no relationships, so it is named Ø, the symbol for the empty set.

The area taxonomy succeeds in providing a compact, abstract, structural view of a terminology. That is, an area contains all the concepts of the terminology sharing the same structure, and this structure is used to name the area. However, the area taxonomy fails to provide semantic uniformity, as illustrated by the concepts *X* and *Y* in the multi-rooted area of Figure 2.3. Their different semantics is manifested by having the targets for the common relationship *r* in two different areas, *W* and *Z*, and by the lack of one concept as a generalization of all concepts in this multi-rooted area.

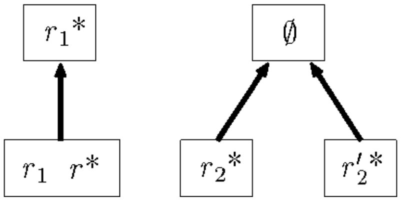


Figure 2.4 Area taxonomy of Figure 2.3.

To illustrate the definitions and further demonstrate the details of the area taxonomy in the context of SNOMED, Figure 2.5 shows an excerpt of the area taxonomy for the Specimen hierarchy. While Figure 2.3 and Figure 2.4 followed a graph model where each concept (area) was displayed as a node in a semantic network, Figure 2.5 simply lists some of the concepts in their area boxes. (The ellipsis "..." indicates the omission of other concepts.) An area is named by its list of relationships enclosed in braces, e.g., {*specimen substance*, *specimen procedure**, *specimen source morphology**} in the lower left of Figure 2.5. A relationship may be marked by a "*" indicating that it is introduced at the particular area. (The "+" marking will be explained in Section 2.2.5.) Thick arrows are *child-of* relationships between areas.

The concept indentation within an area box indicates IS-As. IS-As across areas are drawn as thin arrows. These concept-to-concept arrows are not part of an area's definition, but are included to illustrate the IS-As that underlie the area's *child-of* relationships.

There are over 50 concepts in {*specimen substance**}, all of which share that single relationship. Similarly, all four concepts in {*specimen substance+*, *specimen source morphology*} have the same structure comprising these two relationships. In fact, each area has a structurally uniform set of concepts. The *child-of* from {*specimen source morphology**} to the top-level area Ø is due to the IS-A from the root *Lesion sample* to *General biological sample*. The IS-As from *Liquid material specimen* to *Inanimate samples and substances*, and from *Fluid sample* to *General biological sample* are responsible for the *child-of* from {*specimen substance**} to Ø.

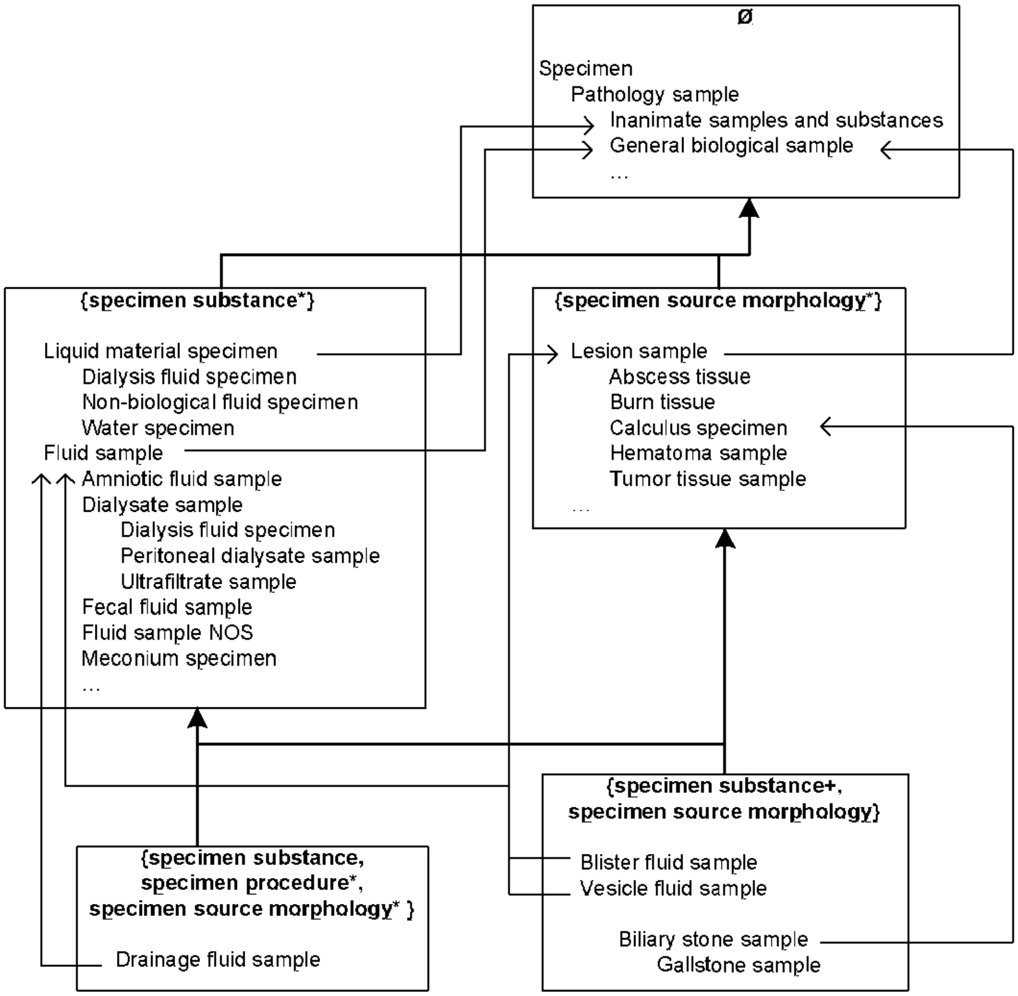


Figure 2.5 Excerpt of the Area Taxonomy for SNOMED's Specimen hierarchy.

Since an area taxonomy is a high-level abstraction of the actual hierarchy, some information is naturally not displayed. For instance, in Figure 2.4, there is no indication of whether a particular area is multi-rooted or not. More specific information is shown in the next level taxonomy.

2.2.4 Partial-areas and Partial-area Taxonomy

A natural solution to the semantic problems in the area taxonomy of Figure 2.4 is to divide the area {r_1, r*} into two constituent parts. Even though the roots *X* and *Y* introduce the same kind of relationship *r*, they each represent a unique semantics given that the targets of their *r* relationships are in different areas. Furthermore, each of the two roots, being a generalization of its descendants, captures their overarching semantics. Thus, *X* and its descendants in the area {r_1, r*} can be seen as a unique semantic grouping. The same is true of *Y* and its descendants in this area. Such a grouping is defined as a *partial-area*. While such a multi-rooted area is named after its relationship(s), each partial-area can be named after its unique root. The root *X* and its descendants form one partial-area *X*, while *Y* and its descendants form another partial-area *Y*. The area {r_1, r*} contains both partial-areas *X* and *Y*.

It is important to note that while the partial-areas form a semantic division of an area, they do not necessarily constitute a partition of the area. In particular, a concept, say, O in {r_1, r*} might be a descendant of both *X* and *Y*. In such a case, *O* would be in both partial-areas *X* and *Y*. Formally, the collection of partial-areas of an area is thus a cover [54] and not a partition.

This second level of division of areas into partial-areas induces a second-level *partial-area taxonomy*. The partial-areas themselves are defined as nodes, and each area is displayed as a collection of partial-area nodes within an area node that is named after the relationship(s). In a partial-area taxonomy, a dashed (instead of solid) rectangle is used to stand for an area, indicating that it comprises partial-areas. For consistency, the notion of partial-area for singly rooted areas is defined as well, although in such a case it contains all (not part of) the concepts in the area.

The hierarchical *child-of* relationship in the partial-area taxonomy is defined similarly to the one in an area taxonomy. That is, if there is an IS-A from the root of a partial-area P_1 to a concept (not necessarily a root) of a partial-area P_2, then in the partial-area taxonomy there is a child-of hierarchical relationship from node P_1 to node P_2. Note that this IS-A needs to be from the root of the partial-area P_1 to guarantee that each of the concepts of P_1 is a descendent of the root of P_2 and inherits its relationships as symbolized by P_1 *child-of* P_2. However, this purpose is achieved even if this IS-A's target is any concept of P_2, as such a concept itself is a descendent of the root of P_2 and inherits its relationships. In the case where all the partial-areas of an area $\{Q\}$ are *child-of* the same partial-area P, one may, in order to prevent clutter, draw one hierarchical arrow from the boundary of $\{Q\}$ to P. This also pertains, as a special case, to areas with a single partial-area.

Figure 2.6 shows the partial-area taxonomy for the terminology fragment of Figure 2.3, including five different areas (the same as in Figure 2.4). Partial-areas, named after their roots, are arrayed inside the area nodes. One area $\{r_1, r^*\}$, contains two partial-areas X and Y. All other areas are singly rooted and thus contain only one partial-area: A, B, Z, and W, respectively. The child-of relationship from the area node $\{r_1, r^*\}$, to the partial-area node A indicates that both partial-areas X and Y are *child-of* the partial-area A.

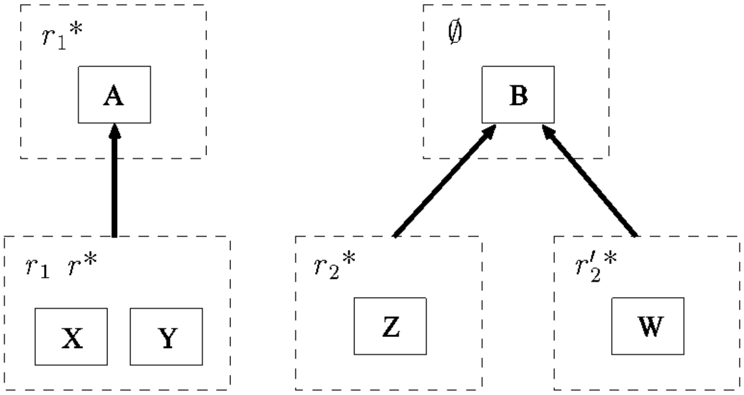


Figure 2.6 Partial-area taxonomy of Figure 2.3.

Note that each partial-area is singly rooted. As discussed in Section 2.2.1, it is paramount that the units of a division be singly-rooted if they are to yield nodes which make a view readily comprehensible. The single root of a partial-area provides one uppermost generalized concept of which all other concepts in the group are descendants. The comprehensibility of a partial-area taxonomy stems from the fact that each one of the concepts in the partial-area is a specialization of the unique root. Due to this, the root functions as an effective designation for an aspect of the semantics: all things in the group are “specializations of the root.” The root itself can be a representative of the entire collection, capturing its general category, and thus in the partial-area taxonomy the corresponding partial-area is named after the root.

Similar to the area schema discussed in Section 2.2.1, one may define relationships among partial-areas and create the *partial-area schema*. However, the choice is made to avoid those relationships and prefer the framework of a partial-area taxonomy. There are several reasons for this choice. The first is that for the purpose of auditing, the partial-area taxonomy is sufficient. The other reason is that a potential partial-area schema will be so overwhelming in its size and complexity that it will not properly promote comprehension of a terminology. In Section 2.3, it is seen that the number of partial-areas for a sample hierarchy of SNOMED is an order of magnitude higher than the number of its areas. Furthermore, the target partial-areas of relationships of a partial-area of one SNOMED hierarchy are typically in another hierarchy. Thus, a partial-area schema will not be constrained within one hierarchy. Thus it will be very difficult to graphically display a partial-area schema and comprehend all its parts. By keeping only the names of the relationships listed once in an area node and not repeated in its multiple partial-area nodes, a much more compact view of the relationships of the partial-areas is provided. Since all partial-areas in an area share the same structure, there is no need to make the structure part of the display of each partial-area. At the same time, for the purpose of auditing, displaying just the names of the relationships of a partial-area without the targets will be sufficient for highlighting most of the irregular or missing concepts of a partial-area. Thus, the decision to use a partial-area taxonomy rather than a partial-area schema seems to be both practical and functional for the purpose of auditing. When needed, an auditor can review the targets of relationships of concepts by accessing the terminology itself.

2.2.5 Regions

Another complication that can arise due to multiple introduction points for the same relationship is demonstrated by the terminology fragment in Figure 2.7. It is seen that the concept H introduces the relationship r_6, while it inherits the relationship r_5 from its parent *F*. As such, *H* is the root of an area, which by the convention in Section 3.1.3 would be denoted {r_5, r_6*}. (To simplify the discussion, assume that r_5 is the only relationship exhibited by *F*. Thus, *F*'s area is {r_5}.) However, this name is not accurate in this context. The concept I also has the relationships r_5 and r_6, but it introduces r_5

while inheriting r_6 from *G*. Therefore, with respect to the root *I*, the area should be named {r_5*, r_6}.

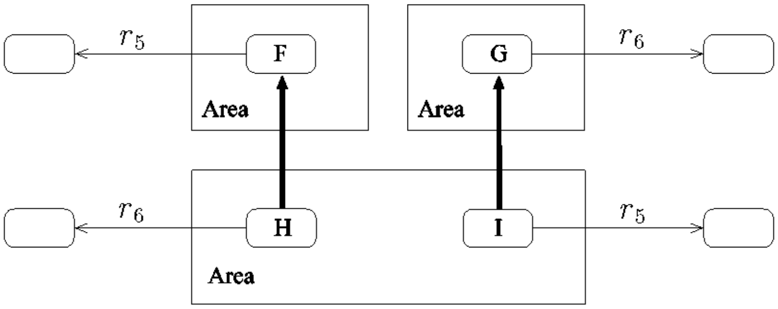


Figure 2.7 Two patterns of relationship obtainment.

There are also problems concerning the *child-of* relationships of this area. The IS-A between H and F induces a *child-of* from H's area to {r_5*}. A similar situation exists regarding the concepts *I* and *G,* with a child-of pointing to {r_6*}. However, the area-taxonomy abstraction in this case gives an inaccurate picture of the status at the concept level. One would infer that all concepts in the area rooted at *H* and *I* would have ancestors in both areas {r_5*} and {r_6*}. But that is not even true for *I* and *H*.

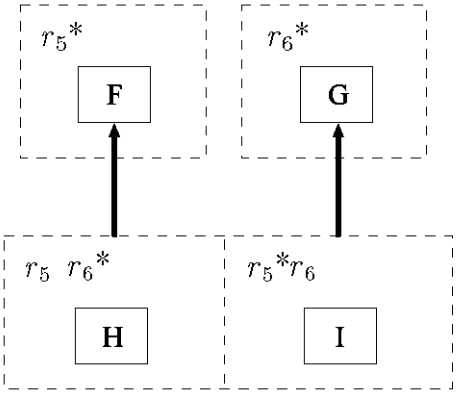


Figure 2.8 Partial-area taxonomy including regions.

To deal with these issues, the partial-area taxonomy is augmented with a division of the problematic area into separate *obtainment-pattern regions* (just *regions* for short). Each region is distinguished by the pattern in which its relationships are introduced and/or inherited, and each is named as if it were a separate area. But graphically all regions of a single area are drawn within the same box, with boundaries between regions drawn as dashed lines. Moreover, for an area with multiple regions, the *child-of*'s are directly from the regions instead of from the area as a whole. An example can be seen in Figure 2.8.

Note that the combination of two relationships, such as r_5 and r_6, leads to the possibility of four different regions. In Figure 2.9, all four possible patterns of relationship obtainment with respect to r_5 and r6 are illustrated. The two additional patterns yield the other two possible regions: {r_5*, r_6*} and {r_5, r_6}. The former is a *strict introduction region*. The latter is a *strict inheritance region*, previously referred to as an *intersection area* [18, 9]. Such strict inheritance regions play an important role in the auditing methodology, as will be discussed below. If a region is neither a strict introduction region nor a strict inheritance region, such as the two regions in Figure 2.8, it is referred to as a *mixed region*. The partial-area taxonomy for Figure 2.9 can be seen in Figure 2.10. Notice that the region {r_5, r_6} is a child of two areas. Strict inheritance regions always have multiple parents. They are also distinguished by the absence of "*" from their names.

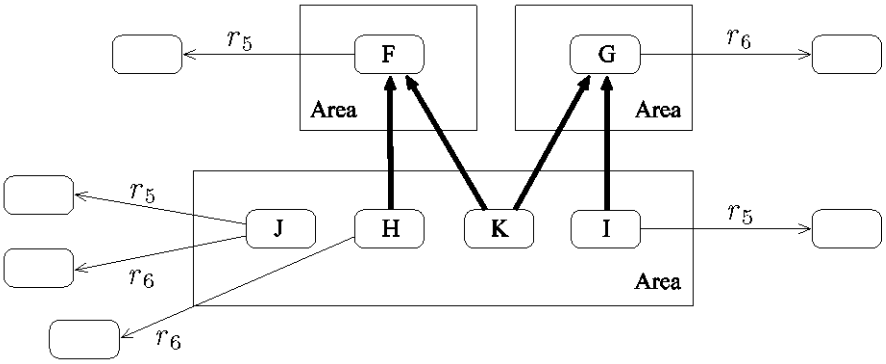


Figure 2.9 All four types of relationship obtainment.

It will be noted that in an area taxonomy areas do not display down to the level of regions. However, when an area exhibits multiple patterns of obtainment with respect to a given relationship, say, *r*, then *r*+ is used in its name. For example, Figure 2.11 shows the area taxonomy of Figure 2.9, where the area involving r_5 and r_6 is marked as {r_5+, r_6+}. As discussed below, this notation is useful in the auditing process.

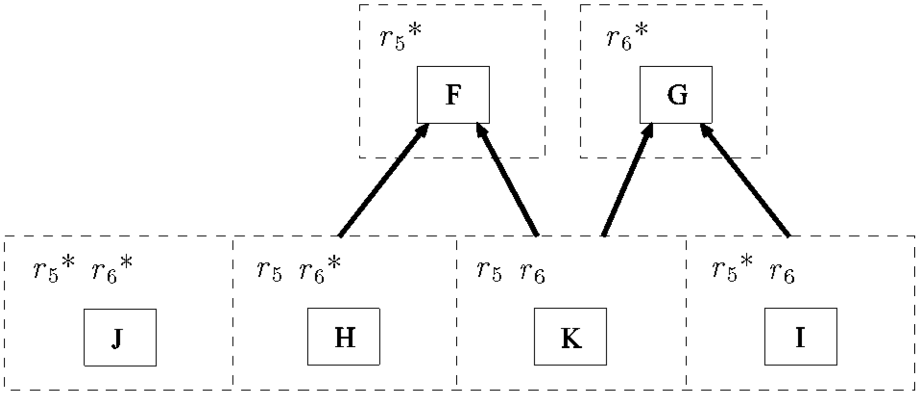


Figure 2.10 Partial-area taxonomy for Figure 2.9.

For convenience, in the following discussion, areas containing only a strict inheritance region will be referred as "strict inheritance areas," while partial-areas of strict inheritance regions are referred as "strict inheritance partial-areas."

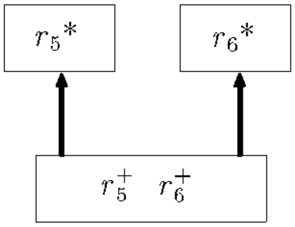


Figure 2.11 Area taxonomy for Figure 2.9.

Figure 2.12 presents the partial-area taxonomy excerpt corresponding to the area taxonomy excerpt of Figure 2.5. In Figure 2.12, the partial-areas appear as solid-line boxes inside their respective areas, now drawn as dashed-line boxes. Inside the box of a partial-area, its name, derived from its unique root, along with its number of concepts (in parentheses) is listed. For example, the area Ø has only one partial-area Specimen containing 30 concepts.

The thick arrows stand for the *child-of* relationships between partial-areas. For example, the partial-area *Fluid sample (9)* is a *child-of Specimen (30)*. In the case of multiple obtainment patterns within one area (indicated by a "+" following the appropriate relationships in the area's name), the area is divided into several regions, each with a disambiguated name. For example, the area {*specimen source morphology*, *specimen substance*+} in Figure 2.5 is divided into two regions in the partial-area taxonomy of Figure 2.12: {*specimen source morphology*, *specimen substance*} and {*specimen source morphology, specimen substance**}. The strict inheritance region on the left contains two partial-areas, *Blister fluid sample (1)* and *Vesicle fluid sample (1)*, both of which do not introduce any relationship. Instead, both inherit *specimen substance* from *Fluid sample* and *specimen morphology* from *Lesion sample*, respectively. The partial-area *Biliary stone sample (2)* in the right region introduces the relationship *specimen substance* while inheriting *specimen source morphology* from *Lesion sample*. Thus, an individual region exhibits a unique obtainment pattern and has partial-areas whose *child-of*'s capture their roots' parentage in other areas' partial-areas.

2.2.6 Auditing Methodologies

The concept groupings and the taxonomy diagrams they induce can serve as the basis for efficient auditing by highlighting irregularities in the terminology. The two levels of taxonomy offer the auditor opportunities to detect irregularities of two kinds, structural and semantic, respectively.

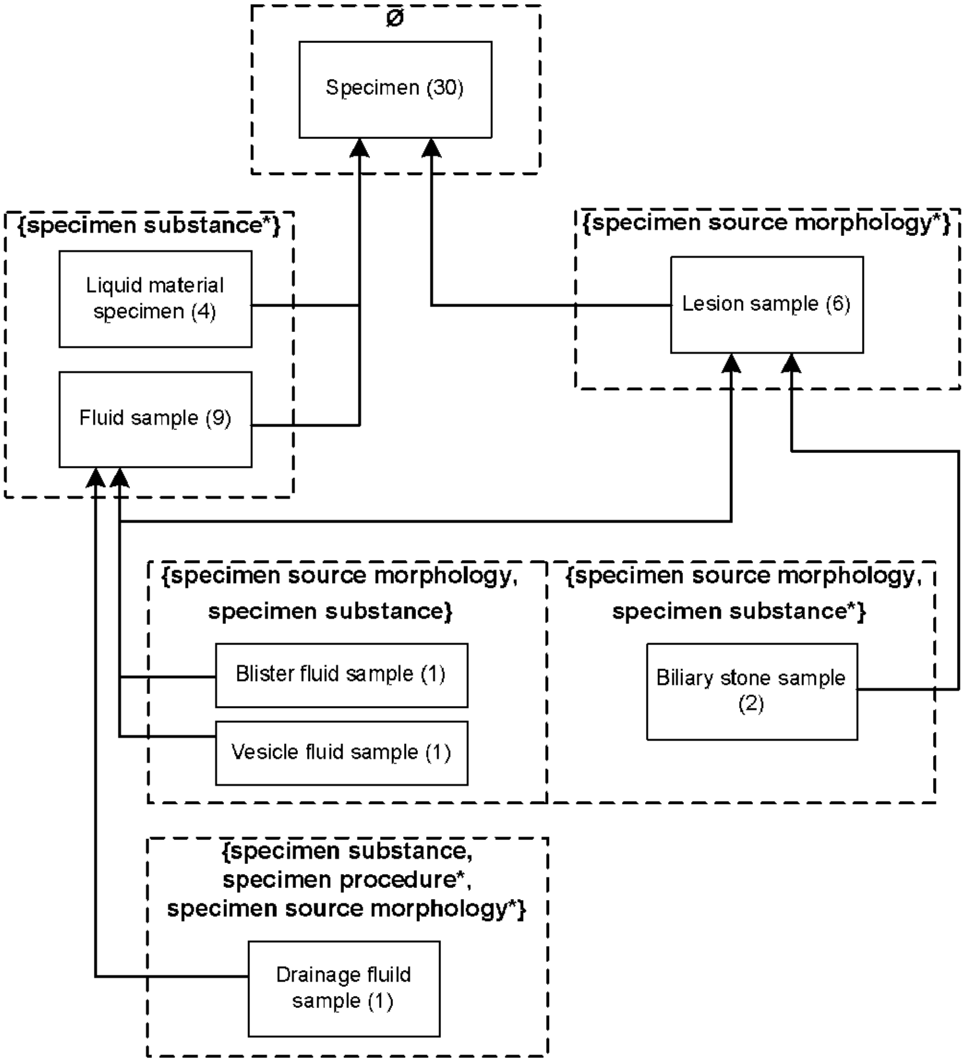

Ø
Specimen (30)
{specimen substance*}
Liquid material specimen (4)
Fluid sample (9)
{specimen source morphology*}
Lesion sample (6)
{specimen source morphology, specimen substance}
Blister fluid sample (1)
Vesicle fluid sample (1)
{specimen source morphology, specimen substance*}
Biliary stone sample (2)
{specimen substance, specimen procedure*, specimen source morphology*}
Drainage fluild sample (1)

Figure 2.12 Excerpt of partial-area taxonomy for the Specimen hierarchy.

2.2.6.1 Detecting Structural Irregularities in the Area Taxonomy. In the area taxonomy, one could detect structural or hierarchical irregularities on the abstract level that may indicate errors on the concrete level. Generally, areas are arranged in levels according to their numbers of relationships. The number of levels depends on the total number of relationships defined for a hierarchy and the actual combinations. These relationships may combine with one another in any form. Combinatorially, n relationships may have up to 2^n different combinations, with n combinations on the first level, $\binom{n}{2}$ on the second level, etc. Comparing the actual number of areas on the lower levels with the theoretical bound can help to uncover potential errors. Missing relationship combinations in the levels with fewer relationships may be due to errors that occurred in the editing process.

Areas on the first level are usually expected to have children in the area taxonomy, because relationships are presumably introduced in the lower levels and are inherited all the way through the hierarchy. Therefore, a first-level area $\{r^*\}$ without any children is a noticeable irregularity, especially when the particular relationship r appears in higher levels of the hierarchy combined with other relationships. A natural question is: is this introduction pattern without further inheritance a reasonable one, and why does it exist? Similarly, a first-level area with very few children (e.g., one child) in the second level compared with other such areas may indicate an irregularity.

It is not expected to encounter many concepts with a large number of relationships since such situations typically denote very complex concepts. If they were to be found, they would be at the higher levels of the area taxonomy. Of special interest in auditing are areas with a large number of relationships but very few concepts, since those concepts would have a complex and uncommon structure.

2.2.6.2 Detecting Irregularities in the Partial-area Taxonomy. The area taxonomy itself is not sufficient to answer these questions because it only contains structural information. This is where the partial-area taxonomy with its semantic knowledge comes in to support the auditing process. It presents a "close-up" abstraction of the concept hierarchy, including information on regions and partial-areas, identifying groups of concepts of uniform structure (relationships) and semantics (a unique generalizing root concept).

Areas with Few Small Partial-areas. An area taxonomy also conveys the number of partial-areas each area has. Using the area taxonomy, one can concentrate on areas with small numbers of partial-areas. The partial-area taxonomy would be further checked to see whether those few partial-areas have a small number of concepts. In such a case, a partial-area having a small number of concepts have been identified with this two-step process, whose combination of relationships (from its area) and its semantics (represented by its root) both occur infrequently. A domain expert would review such a small group of concepts in the context provided by the two taxonomies.

Small Partial-areas with Many Relationships. As mentioned in Section 2.2.6.1, it is recommended that an expert review the partial-areas with a large number of relationships in the higher levels of the partial-area taxonomy. Special attention should be given to such partial-areas with only a few concepts. As mentioned before, the concepts of a small partial-area with an infrequently occurring combination of relationships are highly suspicious.

Strict Inheritance Small Partial-areas. Multiple obtainment patterns (denoted using "+" notation) induce more than one region in an area. When looking into these regions, strict inheritance regions are of special interest in the auditing process. As a matter of fact, the experience [20] in auditing the MED has shown that hunting for errors among strict inheritance regions (referred to in [20] as "intersection areas") can be extremely fruitful. Concepts in strict inheritance regions are more complex, as manifested not only by their compound nature but also by the multiple inheritance of relationships from different parents. Thus, a higher likelihood of errors is expected in strict inheritance regions than in other regions, especially when such a region contains only a few partial-areas of small size. It is expected to have errors such as misclassifications, redundancies, omissions of concepts and relationships, incorrect synonyms, incorrect relationships and relationship targets, incomplete modeling, and modeling inconsistencies.

Compact View Irregularities. The partial-area taxonomy provides a concept-oriented compact view of the content of an area. For example, the area {*specimen substance*} with its 51 concepts is summarized by the partial-area taxonomy as just nine partial-areas whose names indicate what kind of concepts are found in each. This compact view helps the auditor detect irregularities such as duplicate concepts and missing concepts. Such irregularities may be found strictly at the partial-area level or in conjunction with the concept level. An example of a concept duplication observed strictly on the partial-area level is the existence of the two partial-areas *Specimen from ear* and *Ear sample* in the area {*specimen source topography**} (Figure 2.16). Clearly, their roots are redundant.

An example of a missing-concept irregularity observed in conjunction with the concept level occurs with the partial-area S*urgical excision sample* which has only two concepts (Figure 2.15). There are certainly more kinds of surgical excisions that should exist in this partial-area besides the child concept *Specimen obtained by radical excision*. In this case, the partial-area's number of concepts alerted the author to these omissions.

2.2.6.3 Group-based Auditing. The current systematic quality-assurance methods used by the SNOMED editorial staff employ several different tools, notably, Apelon's TDE [55], the CliniClue browser [56], Protégé [57] and IHTSDO Workbench [3]. Most of the editing work is done using the TDE “tree editor” display that focuses on the relationships of one concept. This display shows the children of a concept, along with its defining relationships. When displaying multiple concepts and their interrelationships, the various tools currently employed all display a single folder-type view of a hierarchy, or minor variations such as the TDE’s “concept walker.” The concept walker displays the parents of a concept, as well as its children. Each of these can be expanded to display indented hierarchy views of the corresponding ancestor and descendant hierarchies. The IHTSDO Workbench includes a set of tools that allow users to author terminology, map terminology to other code sets, undertake workflow and process automation, and search, browse or classify terminology.

The efficient auditing methodologies previously developed [9, 20, 58] for large terminologies are based on partitions/divisions and their derived associated abstractions, which distill large networks of concepts down to more manageably sized networks. This distilling process divides the terminology into small groups of “similar concepts,” as defined by a variety of criteria. In turn, reviewing such groups directs auditors toward identifying concepts that are clearly different from others in the group—though they were presumed to be similar—and are thus potentially in error in some way. Forming smaller groups of structurally and semantically similar concepts also enables the identification of “missing” concepts, those which would naturally be expected to belong to a group but are currently absent. Such situations could arise because the concepts were omitted from the terminology originally (perhaps by mistake), or were misclassified or misplaced in the IS-A hierarchy. As such, one can characterize these auditing methodologies as “group-based” auditing as opposed to the standard “concept-based” approaches.

An alternative approach to the interfaces currently used by the SNOMED editorial staff is being presented here. According to the paradigm of area and partial-area taxonomies, concepts are first grouped according to similar structure, and then as a secondary criterion, are grouped as descendants of a root concept. That is, concepts are grouped by areas and partial-areas. Group-based auditing is organized around these groups instead of around individual concepts. Of course, specific concepts are the ultimate targets of auditing, but this particular approach offers a unique path for arriving at them.

The author believes that reviewing the concepts of a partial-area as a group provides a context that helps in detecting errors that would not be exposed when each concept is reviewed separately. Besides the error of missing concepts, other kinds of errors that are expected to find in terminologies while reviewing uniform groups of concepts include: redundant concepts, incorrect IS-A arrangements, erroneous relationship configurations, and modeling errors.

It will be noted that a partial-area taxonomy provides an effective basis for group-based auditing. Moreover, the current auditing methodology fits the characterization of group-based auditing even more so than those methodologies developed previously. While the identified groups in [20] were structurally similar and those in [9, 58] were semantically similar, a partial-area is a group of concepts of both structural and semantic uniformity, and thus is an ideal unit for group-based auditing.

2.2.7 Hypotheses

Based on specific concept groups presented automatically by these partitioning and abstraction methodologies, a few auditing regimens have been put forward, which proved effective. In particular, three regimens focused respectively on two kinds of regions and small-sized partial-areas are applied to a top-level hierarchy of SNOMED. It is noted that the auditing and the subsequent analysis carried out here are based on the inferred (distributed) view of the terminology, i.e., the results after the DL classifier has computed all entailed subsumption relationships.

For the sake of comparison, all the concepts in the chosen SNOMED hierarchy are reviewed for errors. Based on the overall outcomes of these efforts, the validity of the following three hypotheses pertaining to the efficacy of the auditing regimens described here is investigated.

Hypothesis 2.1: There is a higher likelihood for the existence of concept errors in strict inheritance regions than in strict introduction regions or mixed regions. ■

Hypothesis 2.2: There is a higher likelihood for the existence of concept errors in mixed regions than in strict introduction regions. ■

The idea underlying these two hypotheses has to do with hierarchical complexity accumulated in the inheritance process. When a relationship is inherited, it comes down through a path of ancestors who contribute—in addition to the relationship—their accumulated definitional knowledge to the descendant.

Typically, at each level, a constraint or limiting scope is added. Such additional knowledge is sometimes manifested as a more detailed concept name. For example, consider the path from *Specimen to Cyst tissue* (Figure 2.13). It goes through the concepts

Lesion sample (introducing *specimen source morphology*) and *Specimen from cyst*. Naturally, each concept along the path is more specialized than its parent. The specialized knowledge accumulated along the path is referred to as the hierarchical complexity.

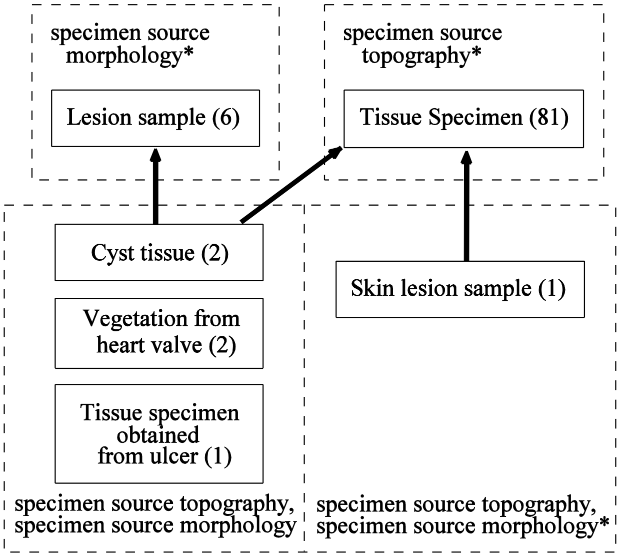

specimen source morphology*
Lesion sample (6)
specimen source topography*
Tissue Specimen (81)
Cyst tissue (2)
Vegetation from heart valve (2)
Tissue specimen obtained from ulcer (1)
Skin lesion sample (1)
specimen source topography, specimen source morphology
specimen source topography, specimen source morphology*

Figure 2.13 Excerpt of partial-area taxonomy showing three areas, four regions, and six partial-areas.

When a concept inherits a relationship, the path has to go through an area where that relationship is introduced. Traversing an area may mean visiting several concepts (e.g., two from the Lesion sample partial-area above). If a concept introduces a relationship instead, then a sub-path going through an area for the sake of picking up the relationship can be avoided, making the overall path shorter. For example, *Hematological sample*, the root of the only partial-area (of 26 concepts) in the strict introduction region of the area {*specimen substance, specimen procedure*} (the rightmost area on level 3 in Figure 2.14). That concept has just one parent Specimen, belonging to the area Ø, and introduces its own two relationships without gaining hierarchical complexity. In general, an inherited relationship implies more hierarchical complexity than an introduced relationship.

A strict inheritance region implies more paths, each of which must travel through areas where inherited relationships are introduced and collected from. This in turn implies that concepts in such a region will, in general, have more ancestors and more hierarchical complexity. The case of strict introduction is of lower hierarchical complexity due to the fact that no extra path is needed to deliver the relationship. A mixed region has an intermediate hierarchical complexity as it inherits some relationships (via an ancestor path) but introduces others (without going through extra areas). The underlying assumption and motivation for the two hypotheses is that concepts with higher hierarchical complexity are more prone to modeling errors.

An example of this can be found in the context of the partial-areas in Figure 2.13. The concept *Skin lesion sample*, the root of its partial-area in the region {*specimen source topography*, *specimen source morphology**}, has a single parent *Skin tissue specimen*, residing in the partial-area *Tissue specimen*, from which it inherits *specimen source topography*. *Skin lesion sample* explicitly introduces *specimen source morphology*, providing further hierarchical complexity. The concept *Cyst tissue* in the neighboring region {*specimen source topography, specimen source morphology*} inherits those two relationships respectively from its parents *Tissue specimen* (the root of its partial-area) and *Specimen from cyst* (in the partial-area *Lesion sample*). Two ancestor paths through these two parents lead to *Cyst tissue*. The one through the latter parent was described above. Therefore, *Skin lesion sample* obtains its relationships in a simpler hierarchical configuration than that needed for *Cyst tissue* and is thus less complex.

Hypothesis 2.3: There is a higher likelihood for the existence of concept errors in small partial-areas than in large partial-areas. ■

This hypothesis indicates the expectation that a small group of concepts similar in their structure and semantics is less likely to be properly modeled and have proper classifications than a similarly constituted *large* group with a common structure and semantics. That is, the high incidence of a combination of a structure and semantics supports its feasibility, while a rarely seen combination raises questions about whether it is the correct structure and root for its few elements. (Note that a similar hypothesis was proposed and verified [46] in the context of the NCIt [47].)

Bootstrap [59] was used to assess the statistical significance of the hypotheses while accounting for the clustering of concepts within partial-areas.

2.3 Results

The techniques presented in Section 2.2 will be demonstrated on an excerpt of SNOMED, specifically, the Specimen hierarchy. The hierarchy contains 1,056 concepts (as of the January 2004 release), and it gives a good illustration of the benefits of the methodology.

2.3.1 Area and Partial-area Taxonomies for the Specimen Hierarchy

There are five relationships defined for concepts of the Specimen hierarchy: *specimen substance*, *specimen source identity*, *specimen source topography*, *specimen source morphology*, and *specimen procedure*. The area taxonomy derived for this hierarchy contains 19 areas, each named after its relationships, with the number of its partial-areas appearing in parentheses (Figure 2.14). For example, the area {*specimen substance*∗} has nine partial-areas. The areas in Figure 2.14 are displayed in color-coded levels according to the number of relationships defined for each. Note that the rightmost area {*specimen* *substance* ∗, *specimen procedure*∗} on level 2 (area Ø is on level 0) is an area consisting of one strict introduction region. Another area where two relationships are introduced together is {*specimen substance, specimen procedure*∗, *specimen source morphology*∗} (leftmost on level 3), where only specimen substance is inherited from its parent area on level 1.

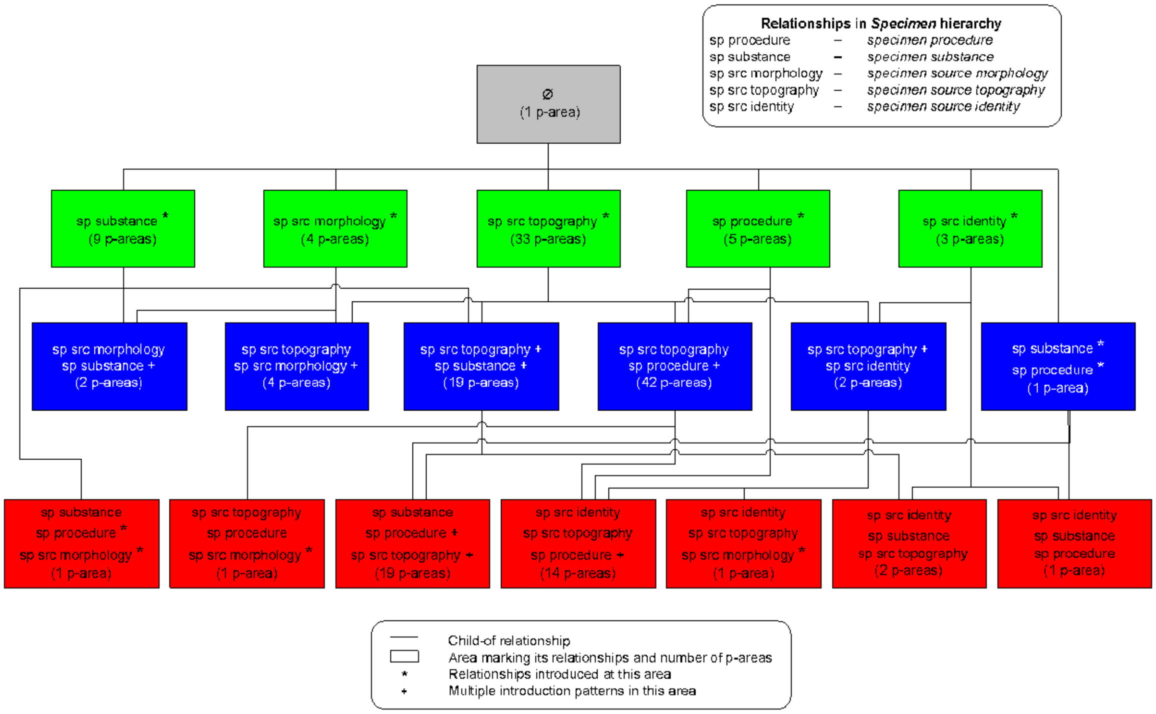

Relationships in *Specimen* hierarchy
sp procedure – *specimen procedure*
sp substance – *specimen substance*
sp src morphology – *specimen source morphology*
sp src topography – *specimen source topography*
sp src identity – *specimen source identity*
Ø (1 p-area)
sp substance * (9 p-areas)
sp src morphology * (4 p-areas)
sp src topography * (33 p-areas)
sp procedure * (5 p-areas)
sp src identity * (3 p-areas)
sp src morphology sp substance + (2 p-areas)
sp src topography sp src morphology + (4 p-areas)
sp src topography + sp substance + (19 p-areas)
sp src topography sp procedure + (42 p-areas)
sp src topography + sp src identity (2 p-areas)
sp substance * sp procedure * (1 p-area)
sp substance sp procedure * sp src morphology * (1 p-area)
sp src topography sp procedure sp src morphology * (1 p-area)
sp substance sp procedure + sp src topography + (19 p-areas)
sp src identity sp src topography sp procedure + (14 p-areas)
sp src identity sp src topography sp src morphology * (1 p-area)
sp src identity sp substance sp src topography (2 p-areas)
sp src identity sp substance sp procedure (1 p-area)
Child-of relationship
Area marking its relationships and number of p-areas
* Relationships introduced at this area
+ Multiple introduction patterns in this area

Figure 2.14 Area taxonomy for the Specimen hierarchy of SNOMED.

Among these 19 areas, seven have multiple patterns of relationship obtainment. For instance, {*specimen source topography*+, *specimen substance*+} contains 19 partial-areas of three different obtainment patterns. Detailed information about the obtainment patterns is shown in the partial-area taxonomy that will be discussed below.

The partial-area taxonomy of the Specimen hierarchy is shown in a sequence of three figures, Figure 2.15-2.17. Due to the extent of some areas, some partial-areas have been omitted from Figure 2.15. They can be found in later figures. In particular, {*specimen source topography**} consisting of 33 partial-areas is fully displayed in Figure 2.16. Similarly, {*specimen source topography*, *specimen procedure*+} with 42 partial-areas is fully displayed in Figure 2.17. The number of concepts in a partial-area appears in parentheses. For example, among the nine partial-areas of {*specimen substance**}, the partial-area Body fluid specimen contains eight concepts.

The partial-area taxonomy also presents the regions of the Specimen hierarchy's areas. An example with a complex obtainment pattern is shown in Figure 2.16. The area {*specimen source topography*+, *specimen substance*+} contains 19 partial-areas divided into three regions. Among them, two partial-areas, Tears specimen and *Peritoneal fluid specimen*, inherit from {*specimen substance**} and introduce the other relationship *specimen source topography*. (A partial-area is considered introducing a relationship when its root does.) Another seven partial-areas, including *Breast fluid sample* and *Urological fluid sample*, have the opposite inheritance pattern: they introduce *specimen substance* while inheriting from {*specimen source topography**}. The other ten partial-areas, e.g., *Sweat specimen* and *Saliva specimen*, are in a strict inheritance region. Another such complex area is {*specimen substance*, *specimen procedure*+, *specimen source topography*+} (Figure 2.16), which also contains three regions.

Special attention is given to the strict inheritance regions because of their special importance to the auditing methodologies. The partitioning of the Specimen hierarchy yields nine strict inheritance regions, containing 27 partial-areas and 83 concepts altogether.

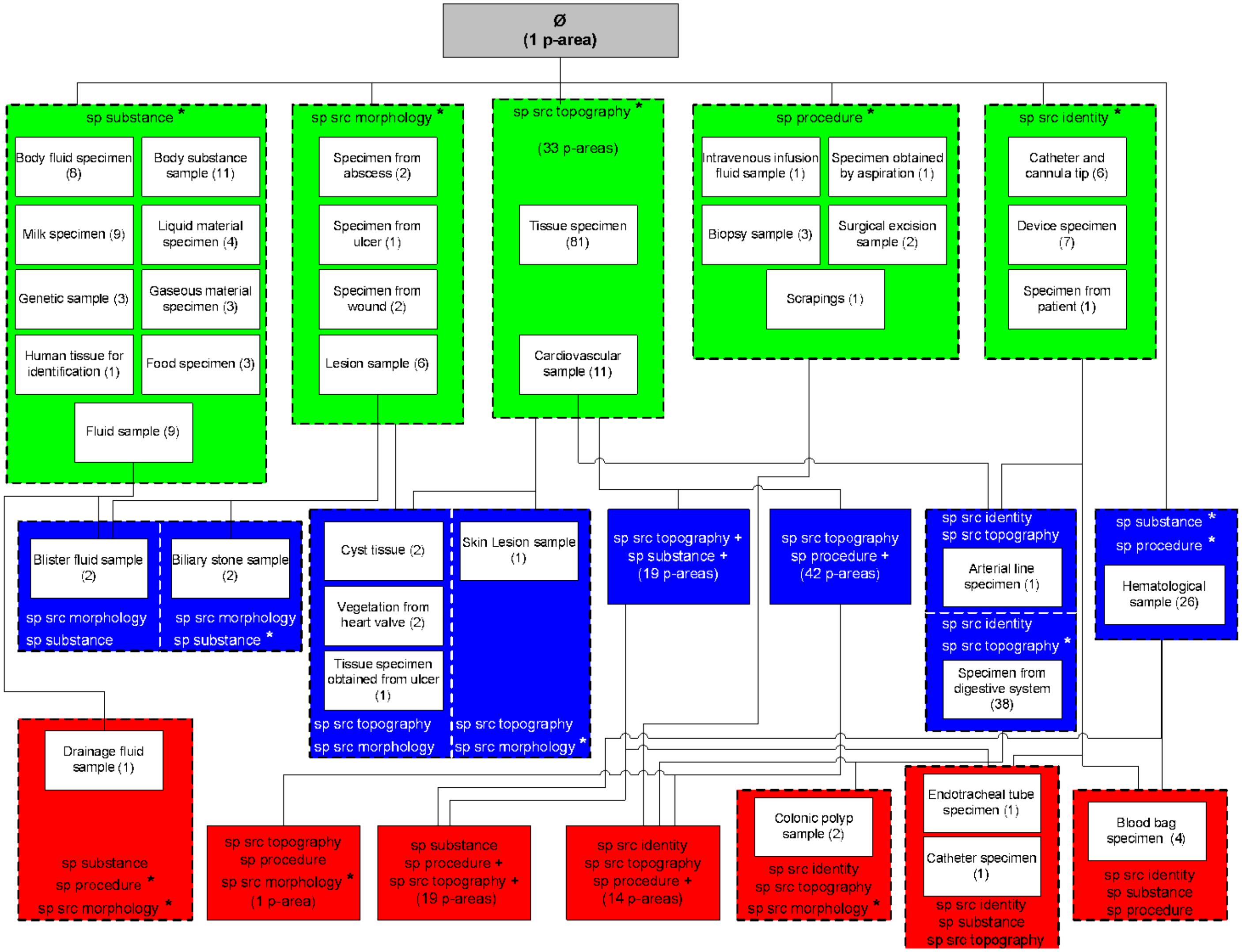


Figure 2.15 Partial-area taxonomy for the Specimen hierarchy (incomplete).

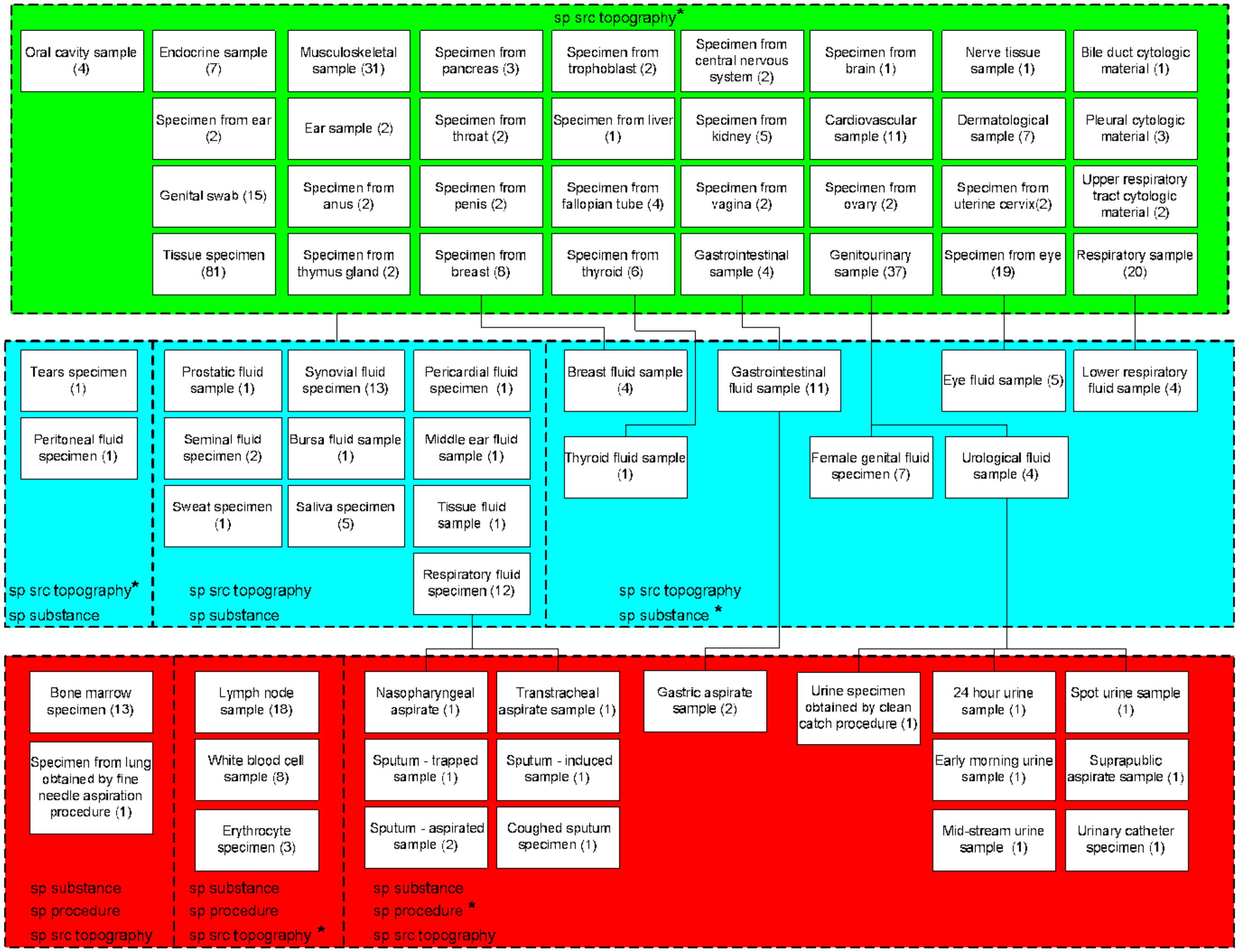

sp src topography*
Oral cavity sample (4)
Endocrine sample (7)
Musculoskeletal sample (31)
Specimen from pancreas (3)
Specimen from trophoblast (2)
Specimen from central nervous system (2)
Specimen from brain (1)
Nerve tissue sample (1)
Bile duct cytologic material (1)
Specimen from ear (2)
Ear sample (2)
Specimen from throat (2)
Specimen from liver (1)
Specimen from kidney (5)
Cardiovascular sample (11)
Dermatological sample (7)
Pleural cytologic material (3)
Genital swab (15)
Specimen from anus (2)
Specimen from penis (2)
Specimen from fallopian tube (4)
Specimen from vagina (2)
Specimen from ovary (2)
Specimen from uterine cervix(2)
Upper respiratory tract cytologic material (2)
Tissue specimen (81)
Specimen from thymus gland (2)
Specimen from breast (8)
Specimen from thyroid (6)
Gastrointestinal sample (4)
Genitourinary sample (37)
Specimen from eye (19)
Respiratory sample (20)
Tears specimen (1)
Peritoneal fluid specimen (1)
sp src topography*
sp substance
Prostatic fluid sample (1)
Synovial fluid specimen (13)
Pericardial fluid specimen (1)
Seminal fluid specimen (2)
Bursa fluid sample (1)
Middle ear fluid sample (1)
Sweat specimen (1)
Saliva specimen (5)
Tissue fluid sample (1)
Respiratory fluid specimen (12)
sp src topography
sp substance
Breast fluid sample (4)
Gastrointestinal fluid sample (11)
Eye fluid sample (5)
Lower respiratory fluid sample (4)
Thyroid fluid sample (1)
Female genital fluid specimen (7)
Urological fluid sample (4)
sp src topography
sp substance *
Bone marrow specimen (13)
Specimen from lung obtained by fine needle aspiration procedure (1)
sp substance
sp procedure
sp src topography
Lymph node sample (18)
White blood cell sample (8)
Erythrocyte specimen (3)
sp substance
sp procedure
sp src topography *
Nasopharyngeal aspirate (1)
Transtracheal aspirate sample (1)
Sputum - trapped sample (1)
Sputum - induced sample (1)
Sputum - aspirated sample (2)
Coughed sputum specimen (1)
Gastric aspirate sample (2)
Urine specimen obtained by clean catch procedure (1)
24 hour urine sample (1)
Spot urine sample (1)
Early morning urine sample (1)
Suprapublic aspirate sample (1)
Mid-stream urine sample (1)
Urinary catheter specimen (1)
sp substance
sp procedure *
sp src topography

Figure 2.16 Excerpt of the partial-area taxonomy for the Specimen hierarchy.

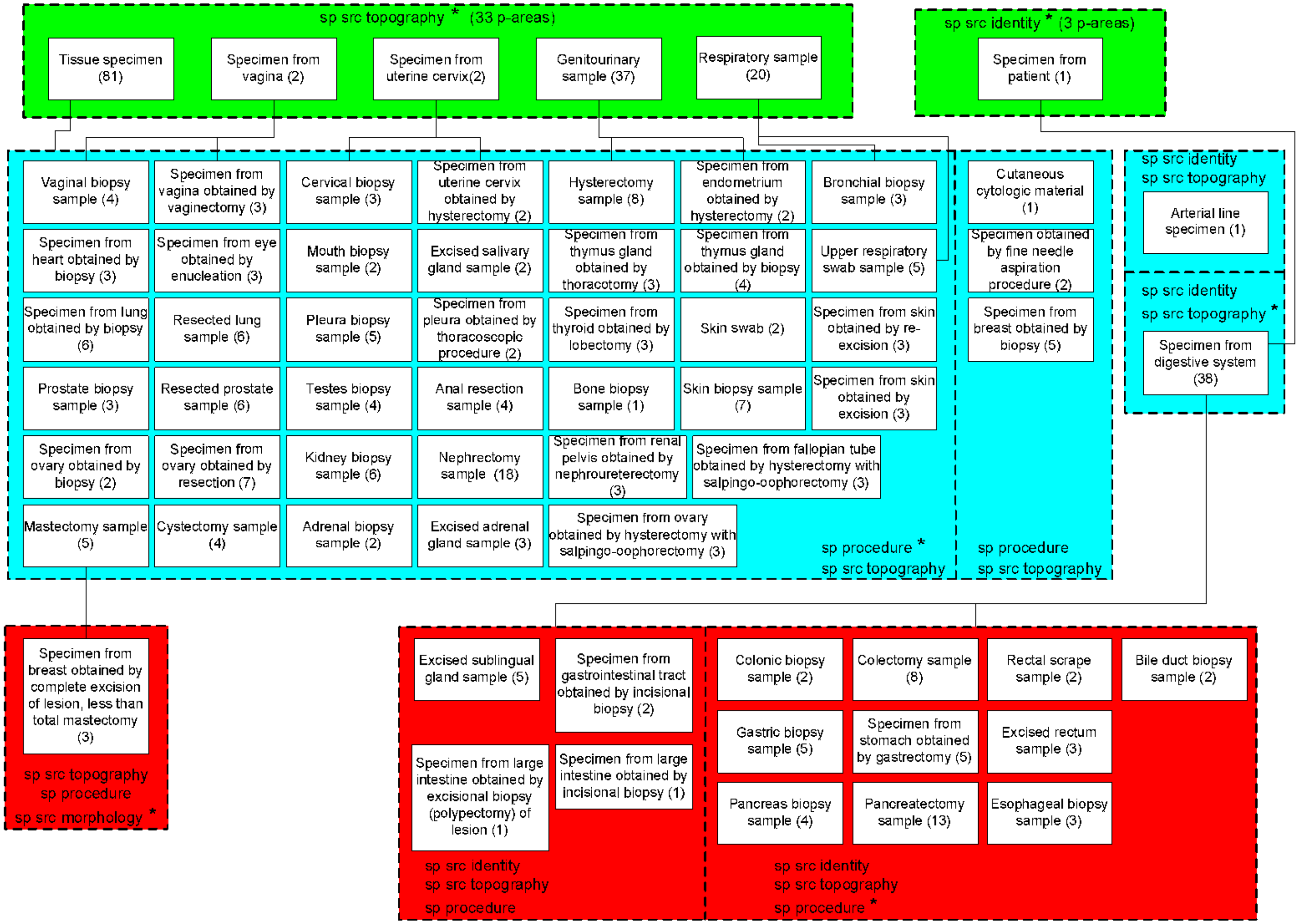

sp src topography * (33 p-areas)
Tissue specimen (81)
Specimen from vagina (2)
Specimen from uterine cervix(2)
Genitourinary sample (37)
Respiratory sample (20)
sp src identity * (3 p-areas)
Specimen from patient (1)
Vaginal biopsy sample (4)
Specimen from vagina obtained by vaginectomy (3)
Cervical biopsy sample (3)
Specimen from uterine cervix obtained by hysterectomy (2)
Hysterectomy sample (8)
Specimen from endometrium obtained by hysterectomy (2)
Bronchial biopsy sample (3)
Specimen from heart obtained by biopsy (3)
Specimen from eye obtained by enucleation (3)
Mouth biopsy sample (2)
Excised salivary gland sample (2)
Specimen from thymus gland obtained by thoracotomy (3)
Specimen from thymus gland obtained by biopsy (4)
Upper respiratory swab sample (5)
Specimen from lung obtained by biopsy (6)
Resected lung sample (6)
Pleura biopsy sample (5)
Specimen from pleura obtained by thoracoscopic procedure (2)
Specimen from thyroid obtained by lobectomy (3)
Skin swab (2)
Specimen from skin obtained by re-excision (3)
Prostate biopsy sample (3)
Resected prostate sample (6)
Testes biopsy sample (4)
Anal resection sample (4)
Bone biopsy sample (1)
Skin biopsy sample (7)
Specimen from skin obtained by excision (3)
Specimen from ovary obtained by biopsy (2)
Specimen from ovary obtained by resection (7)
Kidney biopsy sample (6)
Nephrectomy sample (18)
Specimen from renal pelvis obtained by nephroureterectomy (3)
Specimen from fallopian tube obtained by hysterectomy with salpingo-oophorectomy (3)
Mastectomy sample (5)
Cystectomy sample (4)
Adrenal biopsy sample (2)
Excised adrenal gland sample (3)
Specimen from ovary obtained by hysterectomy with salpingo-oophorectomy (3)
sp procedure *
sp src topography
Cutaneous cytologic material (1)
Specimen obtained by fine needle aspiration procedure (2)
Specimen from breast obtained by biopsy (5)
sp procedure
sp src topography
sp src identity
sp src topography
Arterial line specimen (1)
sp src identity
sp src topography *
Specimen from digestive system (38)
Specimen from breast obtained by complete excision of lesion, less than total mastectomy (3)
sp src topography
sp procedure
sp src morphology *
Excised sublingual gland sample (5)
Specimen from gastrointestinal tract obtained by incisional biopsy (2)
Specimen from large intestine obtained by excisional biopsy (polypectomy) of lesion (1)
Specimen from large intestine obtained by incisional biopsy (1)
sp src identity
sp src topography
sp procedure
Colonic biopsy sample (2)
Colectomy sample (8)
Rectal scrape sample (2)
Bile duct biopsy sample (2)
Gastric biopsy sample (5)
Specimen from stomach obtained by gastrectomy (5)
Excised rectum sample (3)
Pancreas biopsy sample (4)
Pancreatectomy sample (13)
Esophageal biopsy sample (3)
sp src identity
sp src topography
sp procedure *

Figure 2.17 A second excerpt of the partial-area taxonomy for the Specimen hierarchy.

2.3.2 Auditing Using Taxonomies

With the area taxonomy and partial-area taxonomy in place, this section is dedicated to demonstrate how to utilize them to uncover errors of various kinds.

2.3.2.1 Structural Irregularities in Specimen Area Taxonomy. Consider the area taxonomy (Figure 2.14) of the Specimen hierarchy of SNOMED. Theoretically, with five relationships, the child-of hierarchy could be as deep as five levels. The actual taxonomy turns out to have areas with at most three relationships. That is, the most complex specimen concepts have no more than three relationships.

All five relationships are represented in the first-level areas, and each area has children on level 2. Among the five first-level areas, only two, {*specimen procedure**} and {*specimen source identity**}, have just a single child on level 2. According to the methodology, those situations are suspicious and need to be investigated. The other three first-level areas, {*specimen source topography**}, {*specimen substance**}, and {*specimen source morphology**}, have more children on level 2.

The sole child of {*specimen source identity**} on level 2 is {*specimen source topography*+ , *specimen source identity*}, containing two partial-areas. The partial-area taxonomy shows two regions (Figure 2.15). The only partial-area, *Specimen from digestive system*, in the region {*specimen source identity*, *specimen source topography**} contains 38 concepts denoting specimens from different parts of the digestive system, such as *Specimen from stomach*, *Tissue specimen from liver*, etc. While the introduction of *specimen source topography* is totally legitimate, the fact that it is a child of the partial-area *Specimen from patient*, from which it inherits *specimen source identity*, is wrong. The root concept *Specimen from digestive system* should rather be a child of

Specimen. Instead, *Specimen from patient* should have six other new concepts as its children, e.g., *Blood bag specimen from patient*, *Leucocyte specimen from patient*, and *Serum specimen from patient.* Thus, this structural irregularity leads to the discovery of a modeling error. It will be noted that following this study, this error has been corrected in the Jan. ’05 release of SNOMED by removing the *specimen source identity* relationship from the root concept *Specimen from digestive system*. Thus, the partial-area moves accordingly to the area {*specimen source topography**}. Furthermore, six new concepts were added as children of *Specimen from patient*.

The only child of {*specimen procedure**} on level 2 is {*specimen source topography*, *specimen procedure*+}, with 42 partial-areas. The partial-area taxonomy shows two regions (Figure 2.17). One region {*specimen procedure**, *specimen source topography*} of 39 partial-areas introduces *specimen procedure* rather than inheriting it directly from the first level. A natural question is: why is this region not a child of {*specimen procedure**} instead?

The 39 partial-areas in this region are further reviewed. Each introduces the relationship *specimen procedure* connecting it with one of the three following procedures: biopsy, excision or resection, and swab (although the actual terms may be different when the procedure is applied to different body parts, e.g., the excision of breast is *Mastectomy*). Due to the difference in the names of the procedures, these subsumptions were probably not realized in the editing stage. Two of these three procedures appear at {*specimen procedure**} in the partial-area taxonomy excerpt in Figure 2.15. The concept *Swab* is in the hierarchy residing at Ø on level 0 and not in {*specimen procedure**} because of a missing-relationship error. Adding this relationship, *Swab* will move to the {*specimen procedure**} area. Therefore, in addition to their current parent partial-areas in {*specimen source topography**}, these 39 partial-areas should be children of one of the corresponding partial-areas in {*specimen procedure**}. For example, *Skin biopsy sample* should be a child of *Biopsy sample* in addition to *Tissue specimen*. Likewise, *Excised salivary gland sample* and *Resected lung sample* should have another parent, *Surgical excision sample*.

As a matter of fact, the Jan. ’05 release of SNOMED confirmed these findings: 37 out of 39 partial-areas appearing in {*specimen source topography*, *specimen procedure**} (Figure 2.17) have been corrected to include one more parent partial-area depicting the procedures. Although the SNOMED editorial team uncovered the errors using other editing tools, they serve to show the effectiveness of the auditing methodology presented here. The only two partial-areas left, *Specimen from pleura obtained by thoracoscopic procedure* and *Specimen from thymus gland obtained by thoracotomy,* do not correspond to any specific procedure in {*specimen procedure**}, and thus will remain in this region. After this correction, 37 out of 39 partial-areas move to the strict inheritance region {*specimen procedure*, *specimen source topography*}, joining three other partial-areas that were there before. As such, the irregularity of two first-level areas having just one child on level 2 led to the discovery of these errors.

2.3.2.2 Irregularities in the Specimen Partial-area Taxonomy.

Areas with Just a Few Small Partial-areas. Special attention is also given to areas/regions with small numbers of partial-areas. There are ten regions having only one partial-area in the partial-area taxonomy (Figures 2.15-2.17). One problematic partial-area, *Specimen from digestive system*, with 38 concepts has been previously identified by its structural irregularity. Among these ten regions, seven of them (four on the second level and three on the third) consist of a single partial-area with three concepts or less. These "small" partial-areas are deemed highly suspicious according to the auditing guidelines. In fact, after review of the partial-area taxonomy and the actual concepts, three such partial-areas having confirmed errors were found. For example, the partial-area *Skin lesion sample* in the region {*specimen source topography*, *specimen source morphology**} (Figure 2.15) has only one concept. In addition to its current parent partial-area, *Tissue specimen*, it should also have the parent *Lesion sample* in the area {*specimen source morphology**}.

Thus, this region disappears and the partial-area joins the three other partial-areas in the strict inheritance region of the same area. Another example is the partial-area rooted at *Biliary stone sample* with two concepts (Figure 2.15), which should not inherit *specimen source morphology* from *Lesion sample*. In this case, the partial-area moves to the area {*specimen substance**}, and the region {*specimen source morphology*, *specimen substance**} disappears as a result of the removal of the relationship *specimen source morphology*.

Small Partial-areas with Many Relationships. The relationship combinations get more complex on the third level. A review of the third-level partial-areas reveals more errors. One area, {*specimen source identity*, *specimen source topography*, *specimen source morphology**}, contains only one partial-area, *Colonic polyp sample* (Figure 2.15), which includes only two concepts. It is obvious that the relationship *specimen source identity* is irrelevant in this context. As being pointed out previously, this area's parent {specimen source identity, *specimen source topography**} inherits an incorrect relationship *specimen source identity*, and this error propagates via the subsumption hierarchy to its descendants. In fact, another third-level area, {*specimen source identity*, *specimen source topography*, *specimen procedure*+}, has the same error due to this problematic parent. After removing the incorrect relationship *specimen source identity*, these two areas disappear and their partial-areas move accordingly to some second-level areas.

Small, Strict Inheritance Partial-areas. The auditing methodology pays special attention to concepts of strict inheritance regions and especially to their small partial-areas. The root concept of the partial-area *Specimen obtained by fine needle aspiration procedure* (Figure 2.17) in the strict inheritance region {*specimen procedure*, *specimen source topography*} has only one child, *Fine needle aspirate of thyroid, cytologic material*. This is thus a small partial-area of a strict inheritance region with few partial-areas. Other specimens obtained by the same procedure are missing from SNOMED, demonstrating the incompleteness of the modeling. This is another example where the compact view of the partial-area taxonomy exposes irregularities on the concept level.

All concepts of such small partial-areas warrant close inspections, not just the roots. For example, the partial-area *Specimen from gastrointestinal tract obtained by incisional biopsy* in the strict inheritance region {*specimen source identity, specimen source topography, specimen procedure*} (Figure 2.17) has only two concepts. Its child concept *Specimen from stomach obtained by incisional biopsy* has a relationship *specimen source topography* connecting it with the wrong target, *Large intestinal structure*. Another error was revealed when reviewing the singleton partial-area *Specimen* *from lung obtained by fine needle aspiration procedure* in the strict inheritance region {*specimen procedure*, *specimen source topography*, *specimen substance*} (Figure 2.16). The root concept should have *Specimen obtained by fine needle aspiration procedure* as a parent instead of *Specimen from lung obtained by biopsy*.

Respiratory fluid specimen in the strict inheritance region {specimen source topography, specimen substance} (Figure 2.16) has *Upper respiratory sample* as one of its parents. Apparently, *Respiratory fluid specimen* could be from either the upper or lower respiratory tracts. The fact that all its children are fluid samples from *upper* respiratory tract (Figure 2.19) made the auditor wonder whether the correct concept here should be *Upper respiratory fluid sample,* which was mistakenly defined as a synonym of *Respiratory fluid specimen* in SNOMED.

Compact View Irregularities. As mentioned in Section 2.2.6, the compact view of the concepts in an area provided by a partial-area taxonomy can help expose irregularities. For example, a partial-area *Female genital fluid specimen* is in the region {*specimen source topography, specimen substance**} (Figure 2.16), but its potential counterpart *Male genital fluid specimen* is missing from SNOMED. Such an omission is observed due to the view of just seven partial-areas in the region containing 36 concepts.

Furthermore, the review of these seven partial-areas reveals that all consist of body fluid sample concepts, and their roots, including *Breast fluid sample* and *Urological fluid sample*, should therefore have IS-As to *Body fluid specimen,* the root (and name) of a partial-area observed in the review of {*specimen substance**} (Figure 2.15). Due to these new IS-As, the relationship specimen substance of all these partial-areas will be inherited rather than introduced, and the whole region will disappear because its partial-areas will move to the strict inheritance region {*specimen source topography, specimen substance*}.

Moreover, when the above mentioned area {*specimen substance**} is reviewed in the context of the partial-area taxonomy, it is observed that *Body fluid specimen* itself should be a child of both *Fluid sample,* the root of its partial-area in {*specimen substance**} (Figure 2.15), and *Body substance sample,* a root of another partial-area in that same area. But when one tries to add *Body fluid specimen* as a child of *Body substance sample*, it becomes apparent that there is already a child *Body fluid sample*. This is an example of two identical concepts, one of which should be a synonym of the other, instead. The reason for such an error is that the term "specimen" was used previously in SNOMED RT, and "sample" was used in CTV3. Such redundancy errors occurred as a result of the integration process.

The incorrect subsumption relationships among *Fluid sample*, *Body fluid sample*, and *Body fluid specimen* lead to other errors in the strict inheritance regions. For example, there are redundant IS-A links if concepts have both *Fluid sample* and *Body fluid sample/specimen* as their parents. The roots *Saliva specimen*, *Sweat specimen*, and *Seminal fluid specimen* of their respective small partial-areas in the strict inheritance region {*specimen source topography*, *specimen substance*} (Figure 2.16) should not be, as a consequence, children of *Fluid sample*, just of *Body fluid specimen*. When reviewing some larger partial-areas in that region, the author found some other roots, such as *Respiratory fluid specimen* and *Saliva specimen*, that should also not have *Fluid sample* as a parent.

2.3.3 Group-based Auditing

To demonstrate such group-based auditing, consider the part of the Specimen hierarchy that includes *Body fluid specimen* and all its 70 descendants. In the structural analysis displayed by the partial-area taxonomy (Figure 2.15), the partial-area rooted at *Body fluid specimen* contains only eight concepts. Hence, the other 63 descendants have a different structure and thus appear in a different area.

Figure 2.18 and Figure 2.19 present the same 71 concepts in a different way: in an indented format as in the SNOMED CliniClue browser [56]. The concepts are grouped into partial-areas of different regions. Figure 2.19 shows only the concepts that are in the strict inheritance regions; the other descendants of *Body fluid specimen* are shown in Figure 2.18. For completeness, all concepts in every partial-area are shown, but only the descendants of *Body fluid specimen* are shown in black; others are in blue.

The indented hierarchy display of SNOMED CLUE can be used to support review of groups, such as a concept together with all its children (e.g., *Urine specimen* and its nine children), or a concept and all its descendants (e.g., *Sputum specimen*). However, these groups have some deficiencies. Although such a group is cohesive due to its unique root, the structures of its concepts are not necessarily the same. For instance, neither *Catheter specimen* nor *Urinary catheter specimen*, both children of *Urine specimen*, has the same structure as its parent. Furthermore, concepts may have other parents that appear in a different location and are not seen in the tree representation. For example, in addition to the parent *Body fluid specimen*, *Urine specimen* has the parent *Urological fluid sample*, appearing in another part of the Specimen hierarchy. Similarly, *Catheter specimen* has the parent *Device specimen*, in addition to *Urine specimen*.

Figure 2.18 and Figure 2.19 break down the hierarchy into multiple partial-areas and go down to the concept level, thus providing the auditors with a more refined view of the partial-area taxonomy. As mentioned previously, the typically small groups of concepts of a partial-area are uniform both structurally and semantically. In addition, the partial-area taxonomies reflect the multiple parents of a partial-area if they exist (especially for the partial-areas in the strict inheritance regions). Hence, review of concept groups of partial-areas is more promising for the purpose of auditing than review of the indented tree representation.

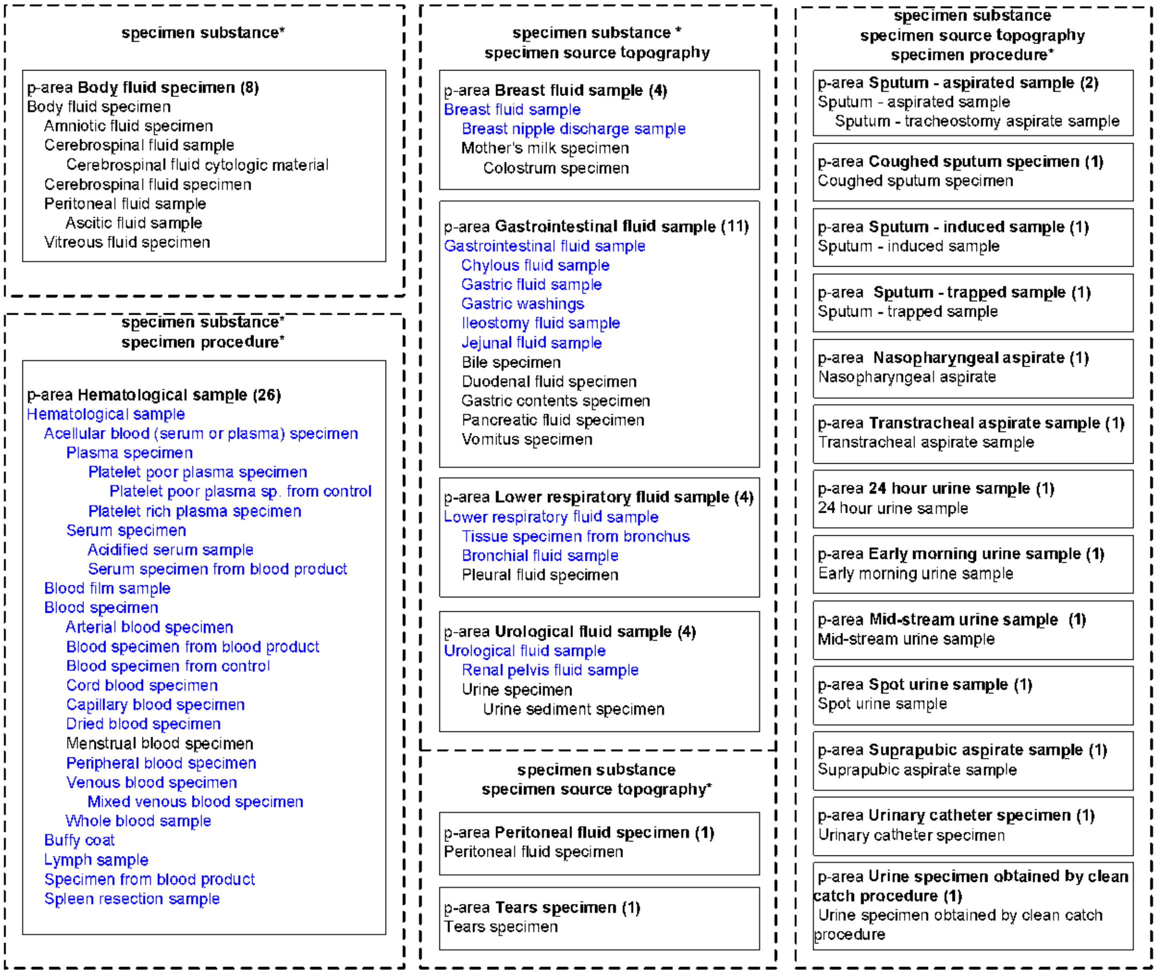

specimen substance*
p-area Body fluid specimen (8)
Body fluid specimen
Amniotic fluid specimen
Cerebrospinal fluid sample
Cerebrospinal fluid cytologic material
Cerebrospinal fluid specimen
Peritoneal fluid sample
Ascitic fluid sample
Vitreous fluid specimen
specimen substance*
specimen procedure*
p-area Hematological sample (26)
Hematological sample
Acellular blood (serum or plasma) specimen
Plasma specimen
Platelet poor plasma specimen
Platelet poor plasma sp. from control
Platelet rich plasma specimen
Serum specimen
Acidified serum sample
Serum specimen from blood product
Blood film sample
Blood specimen
Arterial blood specimen
Blood specimen from blood product
Blood specimen from control
Cord blood specimen
Capillary blood specimen
Dried blood specimen
Menstrual blood specimen
Peripheral blood specimen
Venous blood specimen
Mixed venous blood specimen
Whole blood sample
Buffy coat
Lymph sample
Specimen from blood product
Spleen resection sample
specimen substance *
specimen source topography
p-area Breast fluid sample (4)
Breast fluid sample
Breast nipple discharge sample
Mother's milk specimen
Colostrum specimen
p-area Gastrointestinal fluid sample (11)
Gastrointestinal fluid sample
Chylous fluid sample
Gastric fluid sample
Gastric washings
Ileostomy fluid sample
Jejunal fluid sample
Bile specimen
Duodenal fluid specimen
Gastric contents specimen
Pancreatic fluid specimen
Vomitus specimen
p-area Lower respiratory fluid sample (4)
Lower respiratory fluid sample
Tissue specimen from bronchus
Bronchial fluid sample
Pleural fluid specimen
p-area Urological fluid sample (4)
Urological fluid sample
Renal pelvis fluid sample
Urine specimen
Urine sediment specimen
specimen substance
specimen source topography*
p-area Peritoneal fluid specimen (1)
Peritoneal fluid specimen
p-area Tears specimen (1)
Tears specimen
specimen substance
specimen source topography
specimen procedure*
p-area Sputum - aspirated sample (2)
Sputum - aspirated sample
Sputum - tracheostomy aspirate sample
p-area Coughed sputum specimen (1)
Coughed sputum specimen
p-area Sputum - induced sample (1)
Sputum - induced sample
p-area Sputum - trapped sample (1)
Sputum - trapped sample
p-area Nasopharyngeal aspirate (1)
Nasopharyngeal aspirate
p-area Transtracheal aspirate sample (1)
Transtracheal aspirate sample
p-area 24 hour urine sample (1)
24 hour urine sample
p-area Early morning urine sample (1)
Early morning urine sample
p-area Mid-stream urine sample (1)
Mid-stream urine sample
p-area Spot urine sample (1)
Spot urine sample
p-area Suprapubic aspirate sample (1)
Suprapubic aspirate sample
p-area Urinary catheter specimen (1)
Urinary catheter specimen
p-area Urine specimen obtained by clean catch procedure (1)
Urine specimen obtained by clean catch procedure

Figure 2.18 The partial-areas of strict introduction regions and mixed regions containing the descendants of *Body fluid specimen*.

The errors being reported here were exposed while reviewing such groups. For instance, when the partial-area *Respiratory fluid specimen* in the strict inheritance region {*specimen substance, specimen source topography*} (Figure 2.19) is reviewed, two concepts, *Nasopharyngeal washings* and *Oropharyngeal aspirate*, are found in this partial-area, but a related concept, *Nasopharyngeal aspirate*, which is expected to appear in the same group, was missing. In fact, *Nasopharyngeal aspirate*, a child of *Respiratory fluid specimen*, appears in a separate singleton partial-area in another area {*specimen substance, specimen source topography, specimen procedure**} (Figure 2.18). This leads to the discovery of a "missing relationship" error: seven concepts from this *Respiratory fluid specimen* partial-area, such as *Nasopharyngeal washings*, *Sinus washings*, etc., should have one more relationship, *specimen procedure*, just like two of their siblings *Nasopharyngeal aspirate* and *Transtracheal aspirate sample*.

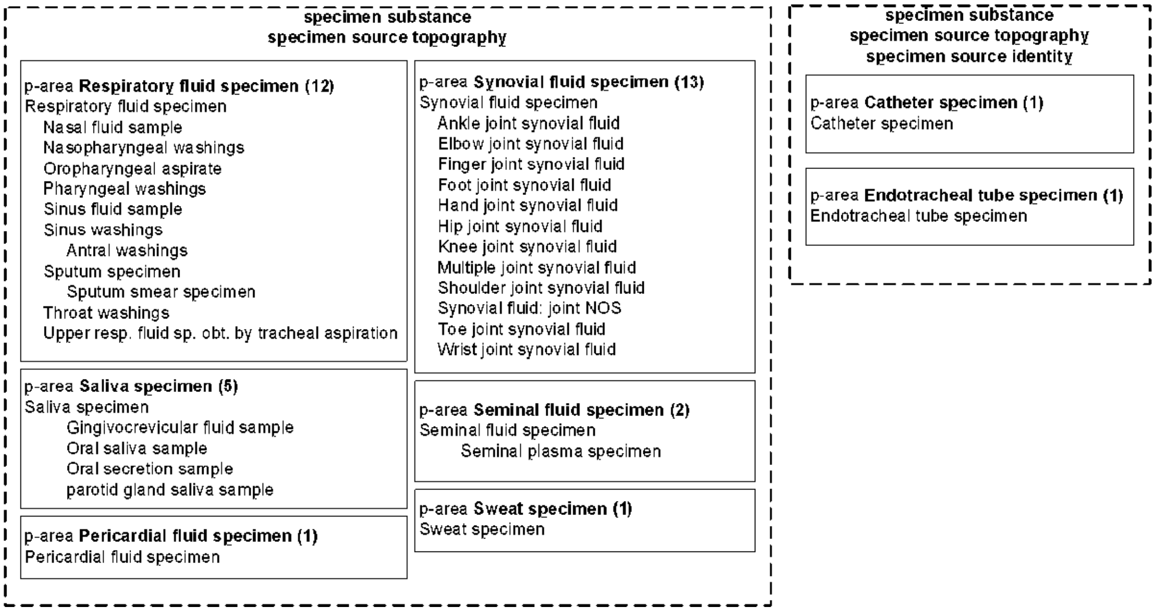

specimen substance
specimen source topography
p-area Respiratory fluid specimen (12)
Respiratory fluid specimen
Nasal fluid sample
Nasopharyngeal washings
Oropharyngeal aspirate
Pharyngeal washings
Sinus fluid sample
Sinus washings
Antral washings
Sputum specimen
Sputum smear specimen
Throat washings
Upper resp. fluid sp. obt. by tracheal aspiration
p-area Saliva specimen (5)
Saliva specimen
Gingivocrevicular fluid sample
Oral saliva sample
Oral secretion sample
parotid gland saliva sample
p-area Pericardial fluid specimen (1)
Pericardial fluid specimen
p-area Synovial fluid specimen (13)
Synovial fluid specimen
Ankle joint synovial fluid
Elbow joint synovial fluid
Finger joint synovial fluid
Foot joint synovial fluid
Hand joint synovial fluid
Hip joint synovial fluid
Knee joint synovial fluid
Multiple joint synovial fluid
Shoulder joint synovial fluid
Synovial fluid: joint NOS
Toe joint synovial fluid
Wrist joint synovial fluid
p-area Seminal fluid specimen (2)
Seminal fluid specimen
Seminal plasma specimen
p-area Sweat specimen (1)
Sweat specimen
specimen substance
specimen source topography
specimen source identity
p-area Catheter specimen (1)
Catheter specimen
p-area Endotracheal tube specimen (1)
Endotracheal tube specimen

Figure 2.19 The partial-areas of strict inheritance regions containing the descendants of *Body fluid specimen*.

Reviewing a group of concepts that is structurally and semantically uniform, as with a partial-area, helps to uncover irregularities. For example, the partial-area *Body fluid specimen* contains the two concepts *Cerebrospinal fluid sample* and *Cerebrospinal fluid specimen*, which are identical. In another example, the partial-area *Peritoneal fluid specimen* is a child of the partial-area *Body fluid specimen*, but the latter contains a concept, *Peritoneal fluid sample*, identical to the root of the former.

When the partial-area *Gastrointestinal fluid sample* is reviewed, it is observed that *Gastric washings* is missing the relationship *specimen procedure*. Such examples demonstrate the power of group-based auditing in exposing irregularities in groups that are supposed to be uniform. Such irregularities may indicate errors that would not otherwise have been detected without the group context.

Altogether 54 errors of different kinds were found using the auditing methodologies reported in this chapter. These errors were reviewed by Dr. Kent A. Spackman, who is the Chief Terminologist of IHTSDO. All but four of the errors were confirmed and corrected in the Jan. ’05 release of SNOMED.

2.3.4 Testing of the Hypotheses

The auditing regimens pertaining to strict inheritance and strict introduction regions and small partial-areas were applied to the Specimen hierarchy of SNOMED, and the resulting error counts with respect to these various groups have been tabulated (Table 2.1 and Table 2.2). For example, within the Specimen hierarchy, there are nine strict inheritance regions encompassing 28 partial-areas and a total of 83 concepts (see the second row of Table 2.1). Among those concepts, 16 errors were discovered, amounting to a percentage of 19.28. The percentages of errors for the other two kinds of regions are:

mixed: 12.60%; and strict introduction: 3.28%. (Note that the first row in Table 2.1 shows the data for the area ∅ whose only region is a special case of a region without any relationships at all.) With respect to the Specimen hierarchy's overall 1,056 concepts, 97 (9.19%) concept errors were found. These figures confirm Hypotheses 1 and 2.

Table 2.1 Errors Across Kinds of Regions

Kind of Region	#	# P-areas	# Concepts	# Errors	% Errors
Ø	1	1	30	2	6.67
Strict Inheritance	9	28	83	16	19.28
Mixed	12	266	516	65	12.60
Strict Introduction	6	157	427	14	3.28
Total:	**28**	**452**	**1,056**	**97**	**9.19**

Table 2.2 Errors Across Ranges of Partial-area Size

P-area Size	# P-areas	# Concepts	# Errors	% Errors
1-7	427	646	69	10.68
8 or more	25	410	28	6.83
Total:	**452**	**1,056**	**97**	**9.19**

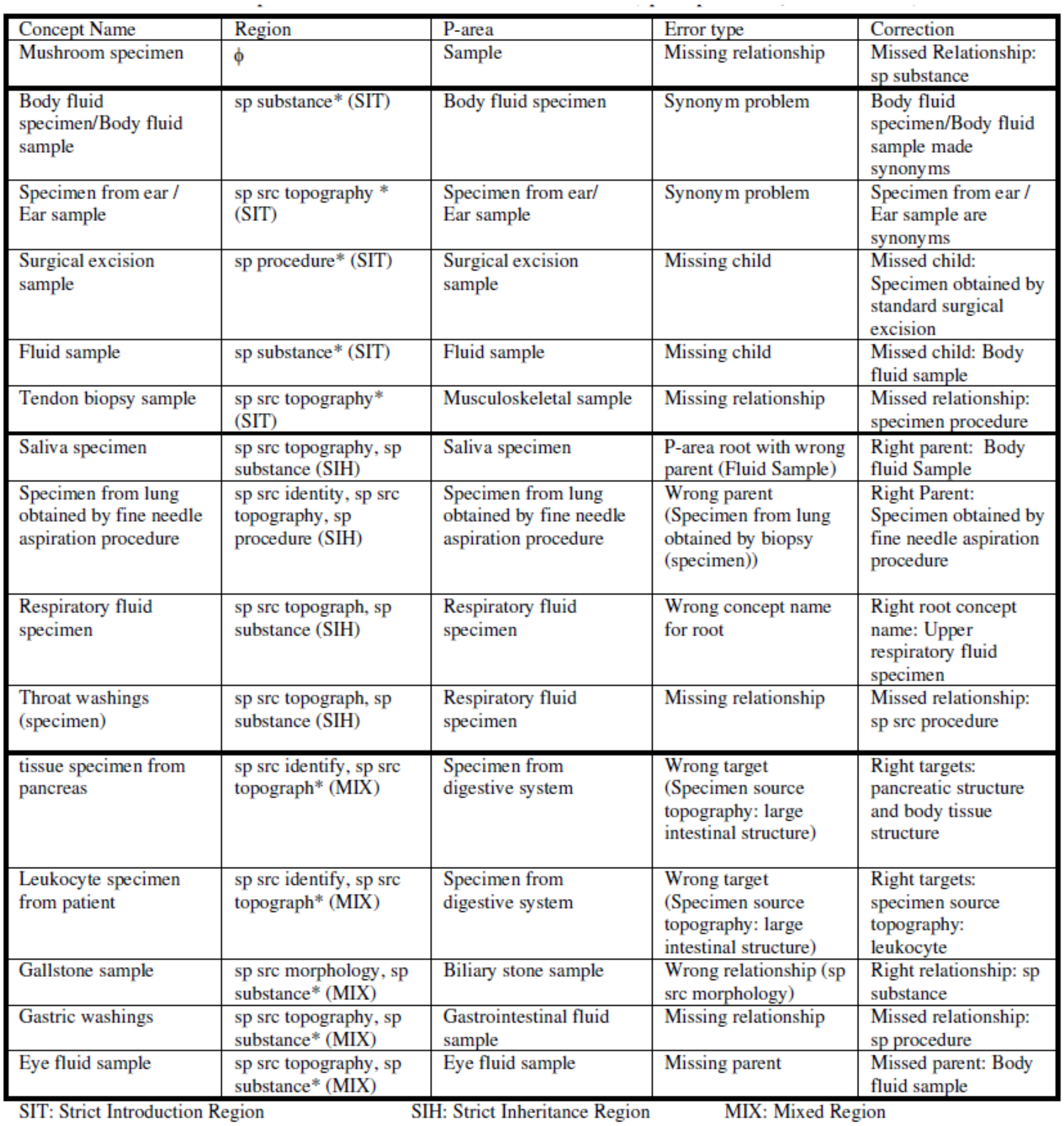
The error totals found in the context of partial-areas of various sizes can be seen in Table 2.2. The table, in fact, breaks the space of partial-areas into two: those with seven or fewer concepts and those with eight or more.

Partial-areas in the former range are deemed to be "small"; those in the latter, large. As can be seen from the table, 10.68% of the concepts in small partial-areas are in error, while the number is only 6.83% for large partial-areas. This result confirms Hypothesis 2.3.

While strict inheritance had a nominally greater error rate than mixed or strict introduction, the differences were not statistically significant, most likely due to the relatively small number of strict inheritance partial-areas. Mixed was greater than strict introduction, and the difference in this case was statistically significant.

The error rate for smaller partial-areas was nominally higher than that for larger partial-areas, but again the difference was not statistically significant, perhaps due to the small number of large partial-areas.

Table 2.3 Sample of Errors Discovered in SNOMED (sp = specimen; src = source)



Concept Name	Region	P-area	Error type	Correction
Mushroom specimen	ϕ	Sample	Missing relationship	Missed Relationship: sp substance
Body fluid specimen/Body fluid sample	sp substance* (SIT)	Body fluid specimen	Synonym problem	Body fluid specimen/Body fluid sample made synonyms
Specimen from ear / Ear sample	sp src topography * (SIT)	Specimen from ear/ Ear sample	Synonym problem	Specimen from ear / Ear sample are synonyms
Surgical excision sample	sp procedure* (SIT)	Surgical excision sample	Missing child	Missed child: Specimen obtained by standard surgical excision
Fluid sample	sp substance* (SIT)	Fluid sample	Missing child	Missed child: Body fluid sample
Tendon biopsy sample	sp src topography* (SIT)	Musculoskeletal sample	Missing relationship	Missed relationship: specimen procedure
Saliva specimen	sp src topography, sp substance (SIH)	Saliva specimen	P-area root with wrong parent (Fluid Sample)	Right parent: Body fluid Sample
Specimen from lung obtained by fine needle aspiration procedure	sp src identity, sp src topography, sp procedure (SIH)	Specimen from lung obtained by fine needle aspiration procedure	Wrong parent (Specimen from lung obtained by biopsy (specimen))	Right Parent: Specimen obtained by fine needle aspiration procedure
Respiratory fluid specimen	sp src topograph, sp substance (SIH)	Respiratory fluid specimen	Wrong concept name for root	Right root concept name: Upper respiratory fluid specimen
Throat washings (specimen)	sp src topograph, sp substance (SIH)	Respiratory fluid specimen	Missing relationship	Missed relationship: sp src procedure
tissue specimen from pancreas	sp src identify, sp src topograph* (MIX)	Specimen from digestive system	Wrong target (Specimen source topography: large intestinal structure)	Right targets: pancreatic structure and body tissue structure
Leukocyte specimen from patient	sp src identify, sp src topograph* (MIX)	Specimen from digestive system	Wrong target (Specimen source topography: large intestinal structure)	Right targets: specimen source topography: leukocyte
Gallstone sample	sp src morphology, sp substance* (MIX)	Biliary stone sample	Wrong relationship (sp src morphology)	Right relationship: sp substance
Gastric washings	sp src topography, sp substance* (MIX)	Gastrointestinal fluid sample	Missing relationship	Missed relationship: sp procedure
Eye fluid sample	sp src topography, sp substance* (MIX)	Eye fluid sample	Missing parent	Missed parent: Body fluid sample

SIT: Strict Introduction Region SIH: Strict Inheritance Region MIX: Mixed Region

Table 2.3 presents a sample of 15 errors discovered with the use of taxonomy auditing regimens in the context of the 2004 release of SNOMED. In each case, the concept's region, partial-area, kind of error, and required correction are listed. The table is subdivided with respect to the different kinds of regions (SIT = strict introduction; SIH = strict inheritance; MIX = mixed). The second row, for example, shows that the concepts *Body fluid sample* and *Body fluid specimen* were found to be independent concepts, when in fact they should be synonyms of each other. Furthermore, the fifth row indicates the discovery of a missing IS-A between the child *Body fluid sample* and the parent *Fluid sample*.

All the errors in Table 2.3 were, again, confirmed by Dr. Kent A. Spackman. Most of the errors have already been corrected as of the 2007 release. The others will be dealt with in the upcoming release.

2.4 Discussion

2.4.1 Interpretation

In summary, auditing using the two-level taxonomies can be very fruitful. The area and partial-area taxonomies provide the auditor abstract views of different granularities, thus prompting the auditor to view the hierarchy first structurally and later semantically. Consequently, the taxonomies help to detect irregularities, which lead to the identification of potential errors.

The development of area and partial-area taxonomies described above is of more than theoretical interest. Maintenance personnel face great challenges when trying to keep a terminology relatively error-free. A thorough understanding of the general structure of a terminology is imperative. On the other hand, an understanding of every last concept in a large terminology is impractical. The taxonomies aptly fulfill this need by providing a high-level abstract view of the terminology. The compact two-level taxonomy enables better navigation and orientation into the content and structure of a terminology.

When a related object-oriented methodology was applied to the MED [19, 20] previously, the schema obtained was 500 times smaller than the original concept network. It thus compactly revealed the gestalt of the terminology and allowed its designers to see it in a brand new perspective. J. J. Cimino, the designer of the MED stated “The schema captures the essence of the MED while ignoring its minutiae.” In addition, the construction of the schema led to the discovery of some errors and inconsistencies that would otherwise have gone undetected.

In the example of the Specimen hierarchy, a similar phenomenon is encountered for the two-level area taxonomy. 19 areas and 164 partial-areas were obtained for a hierarchy of 1056 concepts. Together the two levels, taken in parts provide a compact view of the structure and content of this hierarchy. For example, looking at the partial-area taxonomy in Figure 2.15, one sees several groups of concepts with the same structure of specimen source substance relationship, such as *Body fluid specimen (8)*, *Body substance sample (11)*, *Milk specimen (9)* and *Fluid sample (9)* as well as few smaller groups. Looking at these partial-areas, one obtains a good comprehension for the concepts with such a relationship. The primary partition into areas helps the orientation by providing structurally similar groups of small to medium numbers of partial-areas.

It has often proved to be the case that when new vocabularies are integrated into the UMLS, the developers of that vocabulary have seen opportunities for improvement as a result of the mapping process, e.g., when the Gene Ontology (GO) was integrated into the UMLS [60]. Likewise, the UMLS developers have seen room for improvement and enhancement. When SNOMED was integrated into the UMLS, errors in 800 concepts, about 0.25% of all concepts of SNOMED, were uncovered. In other words, integration of one terminology into another has also a side effect in terms of auditing. However, the percentage of errors found is much lower than when the techniques presented in this chapter were applied to the sample of the Specimen hierarchy.

General quality-assurance techniques employed by SNOMED involve direct inspection of the hierarchies, inspection of the stated and inferred forms of the description logic definitions of individual concepts, and inspection of the hierarchy changes that result from changes in definitions. The focus of the effort is identified by reports of needed corrections that come from multiple parties, including end users of the terminology. In particular, within the Specimen hierarchy, many needed changes were identified as a direct result of feedback from the research described here. Identification of the same errors also occurred independently through inspection of the concepts by the editors. The author does not have specific data that would compare the effort involved in the two different auditing processes.

2.4.2 Limitations

The auditing methodologies presented in this chapter are based on abstraction networks that require systematic inheritance of relationships (via the terminology's IS-A hierarchy) for their derivation. They are, therefore, applicable to a number of terminologies exhibiting this behavior, including: SNOMED; the Veteran Administration’s Enterprise Reference Terminology (ERT) [61]; Kaiser’s Convergent Medical Terminology (CMT) [62] (the preceding two based on SNOMED); NCIt [47]; FMA [63]; RxNorm [64]; MED [17]; and the Vocabulary Server (VOSER) terminology [65] (the basis for the 3M Healthcare Data Dictionary [66]). While the list of such qualifying terminologies is not overly extensive, it comprises many that are very important and widely used. Moreover, the author foresees many emerging terminologies being of this ilk and therefore being amenable to the methodologies discussed in this chapter. In fact, the design of SNOMED anticipates the need for extensions and subsets in order to craft terminological artifacts that are tuned to the needs of individual hospitals as well as groups of organizations of all sizes. SNOMED International’s “reference set specification” [67] serves the purpose of extracting components of SNOMED tailored to particular organizational preferences and use-cases. Thus, SNOMED itself is in an ideal position to be the progenitor of a whole family of new terminologies.

Because these methodologies group concepts based on their structure, an auditor may be preferentially directed to review concepts whose structure stands out as being exceptional. This is not necessarily a problem as structural similarity tends to parallel semantic similarity, and semantic errors are liable to be discovered in this manner. However, the methodologies will not readily reveal errors of a semantic nature for concepts whose structure is not particularly exceptional.

The taxonomy derivation and auditing methodology were successfully applied to one small hierarchy of SNOMED, the Specimen hierarchy. However, other hierarchies may potentially yield different results.

For example, hierarchies with low numbers of concepts having multiple parents, such as SNOMED’s Event, Staging and Scales hierarchy or its Dependent Categories hierarchy, will probably have no strict inheritance regions where this auditing methodology focus searching for errors. Some hierarchies have high or low number of relationships that will influence the number of levels of the taxonomies. A more extensive investigation of larger different SNOMED hierarchies is needed to further substantiate and refine this auditing methodology.

2.4.3 Explanation of the Hypotheses

The hypotheses suggest that ever-limited auditing resources be concentrated on small partial-areas of strict inheritance and mixed regions in order to try to maximize the number of errors found for a given amount of effort. The scope of the auditing experiments was limited to the Specimen hierarchy, which represents a relatively small portion of SNOMED. While the tabulated percentages support the hypotheses, the current numbers are too small to achieve statistical significance for two out of the three hypotheses. There is thus a need to apply the auditing methodologies to additional hierarchies to further examine the hypotheses and especially to further support their statistical analysis. Similar results for other hierarchies are expected.

Each hierarchy of SNOMED is different in its size, height, width, number of defined relationships, and pattern of relationship introduction. These characteristics will naturally be reflected in the taxonomies that abstract the hierarchies. It is not clear how those differences will affect the distribution of errors among regions and partial-areas. While the reasoning for the hypotheses suggests a general phenomenon, further experiments are required for verification. In particular, it is difficult to predict the range of small and large partial-areas for the context of the hypotheses. An empirical approach was followed suggesting 7 as the size threshold.

Since SNOMED uses a DL formalism, it can be fruitful to go outside that realm in an effort to uncover errors. As demonstrated in previous sections, SNOMED's DL-classifiers failed to find certain errors (such as the fact that *Eye fluid sample* is a child of *Body fluid sample*) that were found with the structural methodologies.

CHAPTER 3

ABSTRACTION OF COMPLEX CONCEPTS WITH A REFINED PARTIAL-AREA TAXONOMY OF SNOMED

3.1 Introduction

Due to SNOMED's fast growing size and inherent complexity, advanced tools for the display of aspects of SNOMED's conceptual content—facilitating orientation and comprehension—are needed.

In Chapter 2, two high-level abstraction networks have been devised to provide a multi-level abstraction view on top of a SNOMED hierarchy. In addition to their support for orientation to and comprehension of a SNOMED hierarchy, the two networks have served as the bases of the formulation of structural methodologies for auditing SNOMED hierarchies. Importantly, many concept errors were found to have manifested themselves as structural anomalies at the taxonomy level, and thus the taxonomies proved to be effective building blocks for automated auditing regimens. The area taxonomy and partial-area taxonomy for Specimen hierarchy of SNOMED July 2007 release are shown in Figures 3.1 and 3.2.

In this chapter, the taxonomy paradigm is further extended to overcome some deficiencies in the framework in dealing with particularly complex portions of a SNOMED hierarchy. A recurring theme of the previous terminological analyses has been that *complex* concepts—characterized by various structural features—are often obstacles to orientation and comprehension efforts and usually are natural places to look for modeling errors. Of course, there are numerous ways, in different contexts, to qualify the notion of "complex." The idea that concepts are complex when they simultaneously belong to multiple groups along some given categorizing dimension is used here. In the context of SNOMED auditing, as discussed in Chapter 2, concepts appearing in regions of the partial-area taxonomy characterized by the convergence of multiple ancestral inheritance paths were deemed to be complex and given auditing priority.

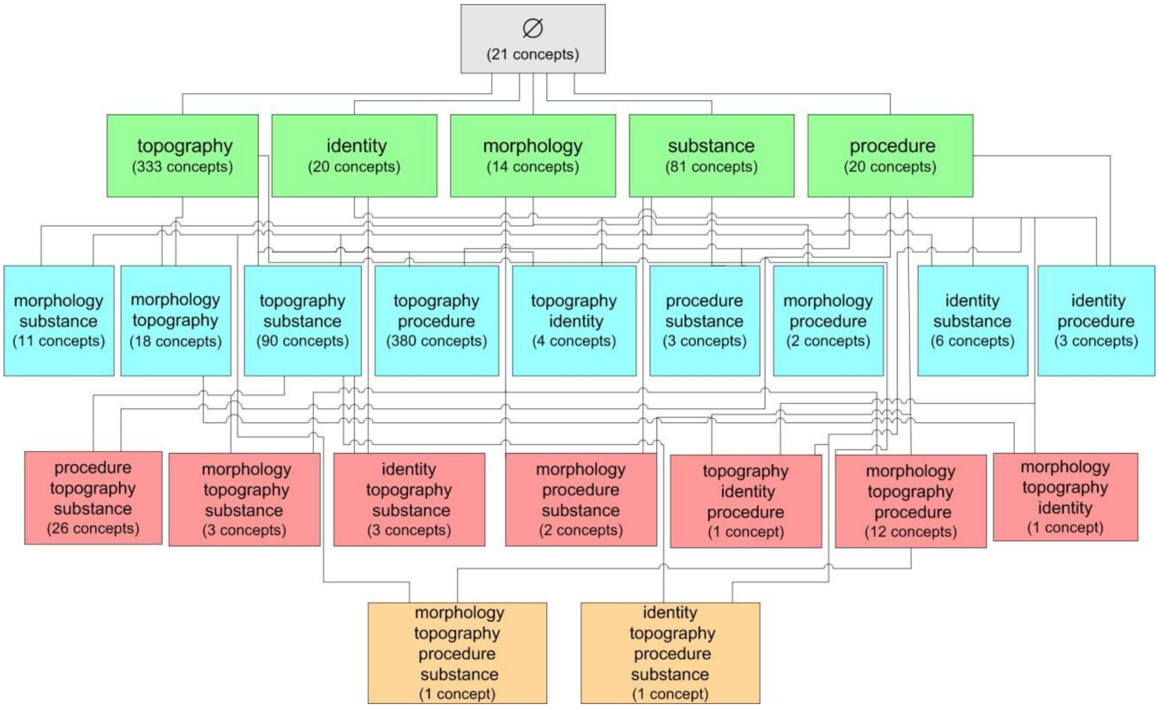


Figure 3.1 Area taxonomy for SNOMED's Specimen hierarchy (July 2007 release).

This chapter focuses on another variety of complex concepts, where again structural feature (relatively easily computed) is being used to determine "complex." In this case, the structural feature is set overlap, and the concepts are those that reside in overlapping portions of two or more partial-areas. As it happens, the entire collection of these overlapping concepts may constitute a highly tangled subhierarchy. It is intended to impose some order on such a subhierarchy to facilitate orientation and comprehension for various users. In particular, an automated methodology is presented to partition the entire set of overlapping concepts to form a *disjoint partial-area taxonomy*, an abstraction network that captures the prevailing hierarchical configuration of the overlaps. Through this taxonomy the user is presented with a view showing the gestalt of the overlaps, allowing for easier comprehension of their content.

One class of user, in particular, that can benefit from the refined, high-level display offered by the new abstraction network is the domain-expert auditor. In this chapter, the details of the abstraction network and its derivation are presented. An enhanced auditing regimen based on this network and the overlapping concepts is expounded in Chapter 4.

3.2 Methods

The partial-area taxonomy has proven to be a useful vehicle for comprehending the overall structure of a SNOMED hierarchy, locating potential errors within it, and identifying modeling aspects that can be improved [48, 49]. However, the taxonomy does lack a characteristic called *semantic uniformity* that has been found useful in the realms of both comprehension and auditing. This deficiency is due to the potential overlap between partial-areas that was alluded to above. For example, the area {*identity*} has two roots, *Device specimen* and *Specimen from patient* (see Figure 3.2). *Device specimen* and its 18 descendants (including *Blood bag specimen*) form one partial-area. *Specimen from patient* and its child *Blood bag specimen, from patient* form another. *Blood bag specimen, from patient* also happens to be a child of *Blood bag specimen*. Thus, *Blood bag specimen, from patient* is in two partial-areas: *Device specimen* and *Specimen from patient*. This situation is illustrated in Figure 3.3. These two partial-areas,

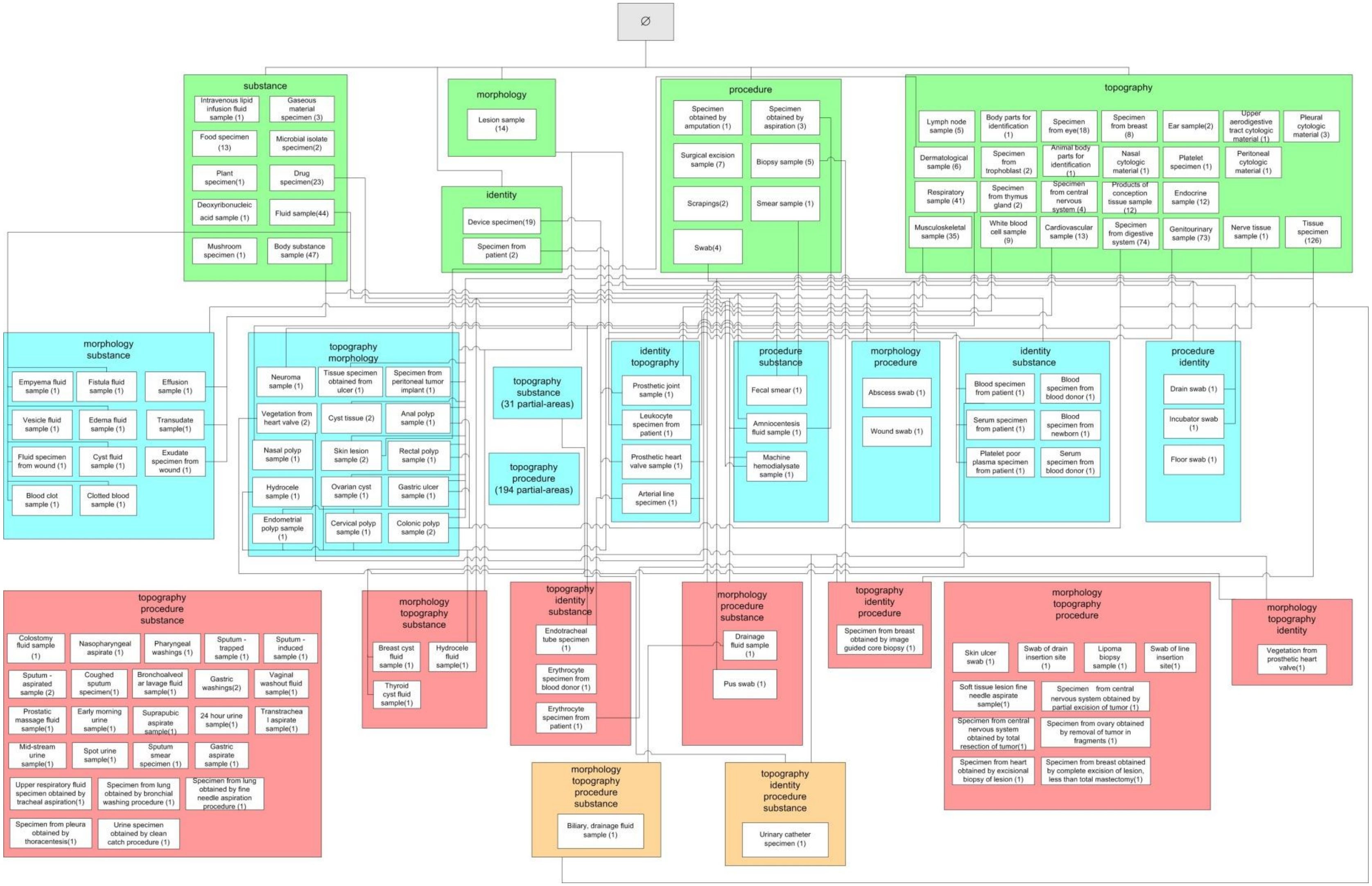

∅
substance
Intravenous lipid infusion fluid sample (1)
Gaseous material specimen (3)
Food specimen (13)
Microbial isolate specimen(2)
Plant specimen(1)
Drug specimen(23)
Deoxyribonucleic acid sample (1)
Fluid sample(44)
Mushroom specimen (1)
Body substance sample (47)
morphology
Lesion sample (14)
identity
Device specimen(19)
Specimen from patient (2)
procedure
Specimen obtained by amputation (1)
Specimen obtained by aspiration (3)
Surgical excision sample (7)
Biopsy sample (5)
Scrapings(2)
Smear sample (1)
Swab(4)
topography
Lymph node sample (5)
Body parts for identification (1)
Specimen from eye(18)
Specimen from breast (8)
Ear sample(2)
Upper aerodigestive tract cytologic material (1)
Pleural cytologic material (3)
Dermatological sample (6)
Specimen from trophoblast (2)
Animal body parts for identification (1)
Nasal cytologic material (1)
Platelet specimen (1)
Peritoneal cytologic material (1)
Respiratory sample (41)
Specimen from thymus gland (2)
Specimen from central nervous system (4)
Products of conception tissue sample (12)
Endocrine sample (12)
Musculoskeletal sample (35)
White blood cell sample (9)
Cardiovascular sample (13)
Specimen from digestive system (74)
Genitourinary sample (73)
Nerve tissue sample (1)
Tissue specimen (126)
morphology substance
Empyema fluid sample (1)
Fistula fluid sample (1)
Effusion sample (1)
Vesicle fluid sample (1)
Edema fluid sample (1)
Transudate sample(1)
Fluid specimen from wound (1)
Cyst fluid sample (1)
Exudate specimen from wound (1)
Blood clot sample (1)
Clotted blood sample (1)
topography morphology
Neuroma sample (1)
Tissue specimen obtained from ulcer (1)
Specimen from peritoneal tumor implant (1)
Vegetation from heart valve (2)
Cyst tissue (2)
Anal polyp sample (1)
Nasal polyp sample (1)
Skin lesion sample (2)
Rectal polyp sample (1)
Hydrocele sample (1)
Ovarian cyst sample (1)
Gastric ulcer sample (1)
Endometrial polyp sample (1)
Cervical polyp sample (1)
Colonic polyp sample (2)
topography substance (31 partial-areas)
topography procedure (194 partial-areas)
identity topography
Prosthetic joint sample (1)
Leukocyte specimen from patient (1)
Prosthetic heart valve sample (1)
Arterial line specimen (1)
procedure substance
Fecal smear (1)
Amniocentesis fluid sample (1)
Machine hemodialysate sample (1)
morphology procedure
Abscess swab (1)
Wound swab (1)
identity substance
Blood specimen from patient (1)
Blood specimen from blood donor (1)
Serum specimen from patient (1)
Blood specimen from newborn (1)
Platelet poor plasma specimen from patient (1)
Serum specimen from blood donor (1)
procedure identity
Drain swab (1)
Incubator swab (1)
Floor swab (1)
topography procedure substance
Colostomy fluid sample (1)
Nasopharyngeal aspirate (1)
Pharyngeal washings (1)
Sputum - trapped sample (1)
Sputum - induced sample (1)
Sputum - aspirated sample (2)
Coughed sputum specimen(1)
Bronchoalveolar lavage fluid sample(1)
Gastric washings(2)
Vaginal washout fluid sample(1)
Prostatic massage fluid sample(1)
Early morning urine sample(1)
Suprapubic aspirate sample(1)
24 hour urine sample(1)
Transtracheal aspirate sample(1)
Mid-stream urine sample(1)
Spot urine sample(1)
Sputum smear specimen (1)
Gastric aspirate sample (1)
Upper respiratory fluid specimen obtained by tracheal aspiration(1)
Specimen from lung obtained by bronchial washing procedure (1)
Specimen from lung obtained by fine needle aspiration procedure (1)
Specimen from pleura obtained by thoracentesis(1)
Urine specimen obtained by clean catch procedure (1)
morphology topography substance
Breast cyst fluid sample (1)
Hydrocele fluid sample(1)
Thyroid cyst fluid sample(1)
topography identity substance
Endotracheal tube specimen (1)
Erythrocyte specimen from blood donor (1)
Erythrocyte specimen from patient (1)
morphology procedure substance
Drainage fluid sample (1)
Pus swab (1)
topography identity procedure
Specimen from breast obtained by image guided core biopsy (1)
morphology topography procedure
Skin ulcer swab (1)
Swab of drain insertion site (1)
Lipoma biopsy sample (1)
Swab of line insertion site(1)
Soft tissue lesion fine needle aspirate sample(1)
Specimen from central nervous system obtained by partial excision of tumor (1)
Specimen from central nervous system obtained by total resection of tumor(1)
Specimen from ovary obtained by removal of tumor in fragments (1)
Specimen from heart obtained by excisional biopsy of lesion (1)
Specimen from breast obtained by complete excision of lesion, less than total mastectomy(1)
morphology topography identity
Vegetation from prosthetic heart valve(1)
morphology topography procedure substance
Biliary, drainage fluid sample (1)
topography identity procedure substance
Urinary catheter specimen (1)

Figure 3.2 Partial-area taxonomy for SNOMED's Specimen hierarchy (July 2007 release).

Device specimen and *Specimen from patient*, are considered "overlap" with each other. The concept *Blood bag specimen, from patient* is called an "overlapping concept." The entire set of overlapping concepts is denoted *V*.

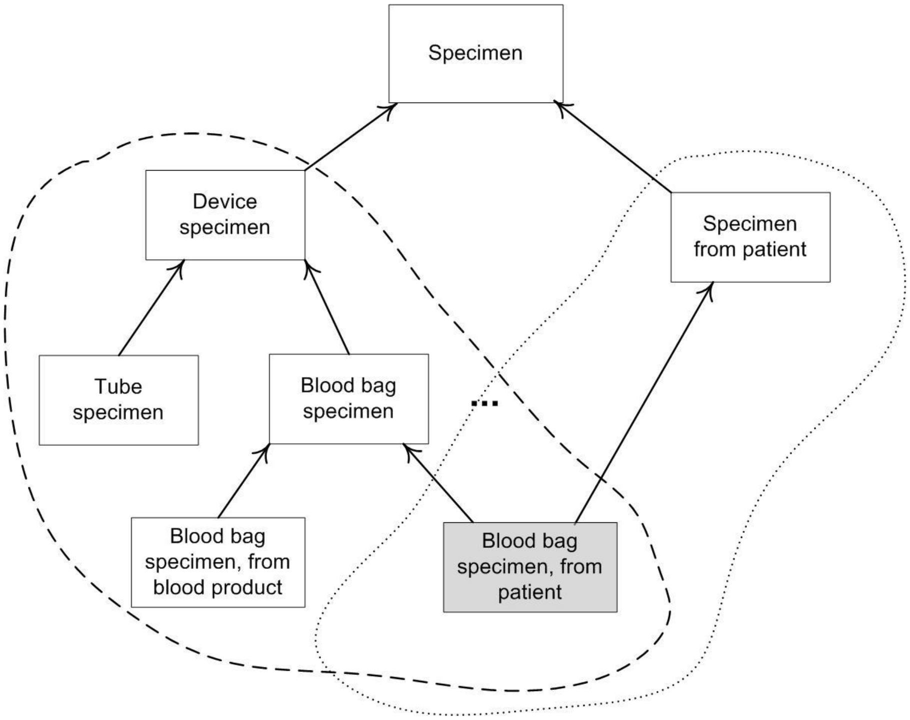


Figure 3.3 The overlapping concept *Blood bag specimen, from patient* resides in two partial-areas, *Device specimen* and *Specimen from patient*, demarcated by the dashed bubbles.

This raises two important issues. First, the entire collection of partial-areas does *not* form a partition of the hierarchy. This is in contrast to the collection of areas which does. Second, when two partial-areas overlap, some concepts in a partial-area, like the concept *Blood bag specimen*, elaborate only the semantics of one root (i.e., *Device specimen*) while the overlapping concepts in that same partial-area, in this case, the concept *Blood bag specimen, from patient,* elaborate the semantics of two roots (i.e., *Device specimen* and *Specimen from patient*). The situation gets worse when three overlapping partial-areas, say, R_1, R_2, and R_3, are involved. In this situation, some concepts in R_1 are elaborating the semantics of the root R_1, while others may be elaborating the semantics of the two roots R_1 and R_2, and others are elaborating the semantics of all three roots R_1, R_2, and R_3. In this sense, the partial-area R_1 is not semantically uniform with respect to its root.

This deficiency of the partial-area taxonomy actually presents the opportunities of further extending and enhancing the taxonomy, which is presented in this chapter. One is the fact that the overlapping concepts lend themselves nicely to auditing scrutiny. Such concepts elaborate the semantics of two or more significant root concepts in the hierarchy and thus warrant the designation "complex concept," which underpins an auditing methodology that will be introduced in Chapter 4.

The second opportunity pertains to the refinement of the partial-area taxonomy. Its theoretical underpinning will be extended and it will be refined to further facilitate comprehending the terminology as well as the job of an auditor. In particular, the overlapping concepts will be partitioned systematically such that each resulting group of concepts is singly-rooted. The single root of each such group will provide a uniform semantics for the whole group. This is important because the overlapping concepts can collectively constitute quite a tangled hierarchy. The partition paves the way for the formation of an enhanced partial-area taxonomy that provides a view of the prevailing hierarchical configuration of the overlapping concepts. This will aid the subject-domain-expert editor and user in seeing the gestalt of the partial-area overlaps and more easily comprehending their content. Furthermore, such enhanced comprehension will enable an auditor to recognize any troublesome aspects. In this context, a new auditing methodology will be presented in Chapter 4.

In the remainder of this section, the issue of the complexity of overlapping concepts is discussed further. After that, a singly-rooted partitioning scheme for the overlapping concepts of an area is devised. This begins with the definition of *overlapping roots*. From the partition, a new refined abstraction network for the concepts of a SNOMED hierarchy will be defined. This refined abstraction network will better support comprehension of a SNOMED hierarchy by maintenance personnel, including editors and auditors, by providing a disjoint partition of the hierarchy's concepts—the overlapping concepts, among them—into semantically uniform groups. It will also form the basis for an enhanced auditing regimen for the overlapping concepts of such a hierarchy, which will be discussed in Chapter 4.

3.2.1 Overlapping Concepts are Complex Concepts

The following example is presented to further motivate the focus on overlapping concepts and see their inherent complexity. In the area {*substance*} (Figure 3.2), the three direct children, *Body substance sample*, *Fluid sample*, and *Drug specimen*, of the top-level concept Specimen induce three partial-areas, respectively. Figure 3.4 shows the three root concepts, along with two of their descendants (shaded). The partial-areas are demarcated with dashed bubbles, where the different border styles denote the different partial-areas. *Body fluid sample*, being a child of both *Body substance sample* and *Fluid sample*, resides in the intersection of the two partial-areas. It inherits the relationship substance directed to *Body fluid* in the Substance hierarchy from both its parents.

The other shaded concept in Figure 3.4, *Acellular blood (serum or plasma) specimen*, sits in the intersection of the partial-areas *Fluid sample* and *Drug specimen*. Thus, it elaborates the semantics of both parents, and inherits the relationship *substance* and the accompanying targets. Different from the previous example, *Acellular blood (serum or plasma) specimen* has two occurrences of the *substance* relationship, one pointing at *Liquid substance* and the other pointing at *Blood component*, a descendant of *Drug or medicament*, the target of the relationship *substance* of *Drug specimen*.

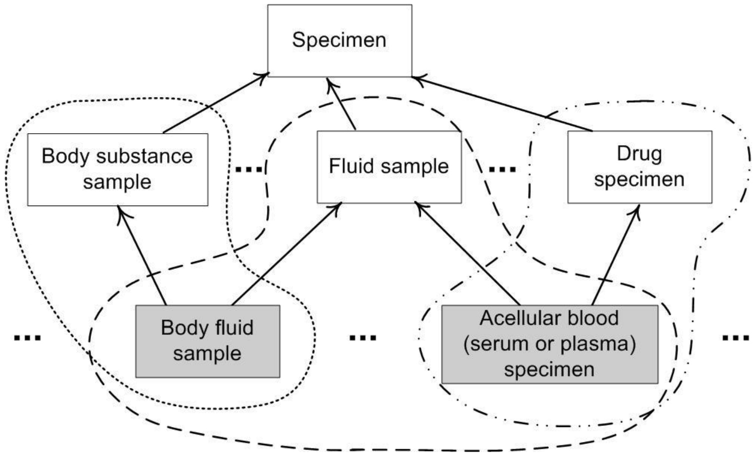


Figure 3.4 The overlapping concepts *Body fluid sample* and *Acellular blood (serum or plasma) specimen* (shaded) in the area {*substance*}.

Overall, the area {*substance*} (Figure 3.2) contains ten partial-areas and has quite a few overlapping concepts. This can be gathered from the fact that the sum of the numbers of concepts in its partial-areas (136) is much higher than the actual number of concepts in the area (81). The increased complexity of overlapping concepts is a consequence of the fact that they represent combination specializations deriving from multiple root concepts. For example, *Body fluid sample* and all its descendants residing in {*substance*} are overlapping concepts belonging to the partial-areas *Body substance sample* and *Fluid sample*. All these concepts that are both body substance and fluid examples, e.g., *Amniotic fluid specimen* and *Lymph sample*, are inherently more complex than concepts that are solely fluid samples, e.g., *Water specimen*, or only body substance samples, e.g., *Calculus specimen*. They each elaborate the semantics of a dual specialization.

The amount of overlapping, and attendant complexity, may increase as traversing downward along the IS-A hierarchy. In {*substance*}, it is found that 15 concepts belonging to exactly two partial-areas, and 20 concepts belonging to three partial-areas. From this, its actual number of concepts is obtained: 136 − (2 − 1) • 15 − (3 − 1) • 20 = 81.

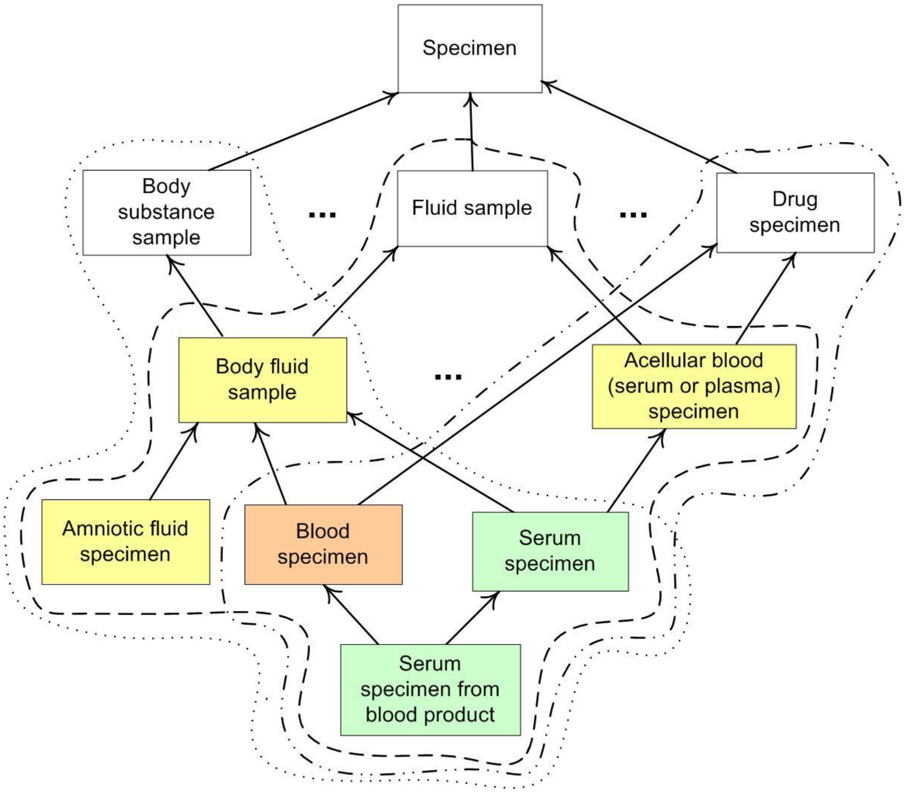


Figure 3.5 Differing degrees of complexity for overlapping concepts in the area {*substance*}. The green overlapping concepts are more complex than the orange overlapping concept which is more complex than the yellow overlapping concepts.

Differing degrees of complexity are seen for the overlapping concepts in Figure 3.5, which contains a small fragment of the Specimen hierarchy consisting of nine concepts from the area {*substance*} (along with the hierarchy's root). The three bubbles with different border styles enclose three partial-areas. Their roots are children of Specimen. All concepts below the roots (in colors) are overlapping concepts. The first of these are the yellow concepts *Body fluid sample* and *Acellular blood (serum or plasma) specimen*. Traversing downward along the IS-A hierarchy, examples of even more complex overlapping concepts were found. For example, one of the children of the overlapping concept *Body fluid sample*, *Blood specimen* (in orange), has another parent *Drug specimen* that is the root of its partial-area. In this case, *Blood specimen* is the specialization of three roots and thus resides in the intersection of three separate partial-areas. But from the complexity point of view, it is a child of one overlapping concept and one root of a partial-area. Hence, it is more complex than the two yellow overlapping concepts that are children of roots of partial-areas.

Other—more complex—cases can be seen with the green concepts, *Serum specimen* and *Serum specimen from blood product*, each having two parents that are overlapping concepts themselves. Note that a move down the hierarchy does not necessarily imply an increase in complexity. This is illustrated by *Amniotic fluid sample*, whose only parent is *Body fluid sample*. Being singly parented, it does not lie at a significant knowledge convergence point and is thus considered no more complex than *Body fluid sample* from a structural standpoint.

3.2.2 Foundations of the Partition: Overlapping Roots

As discussed, the portion of an area consisting of the overlapping concepts may constitute a highly tangled hierarchy. The goal is to impose some order on it by partitioning it in such a way as to obtain a collection of concept groups exhibiting semantic uniformity by satisfying single-rootedness and no overlaps. Thus, the first task is to identify those overlapping concepts that will serve as the roots of the concept groups. They will be called *overlapping roots*. Just like the root of a partial-area, an overlapping root will capture the overarching semantics of its group of overlapping concepts. The grouping process proceeds in a deeply nested (recursive) fashion.

Two kinds of overlapping roots are defined: those at the true "tops" of the overlapping portions of the partial-areas and those residing beneath them—perhaps quite deep in the overlap. Let us first define the fundamental kind of overlapping root called a *base overlapping root*, where, again, *V* is the entire set of overlapping concepts.

Definition (Base Overlapping Root): A concept $L \in V$ is a *base overlapping root* if $\forall C \in parents(L),\ C\ not \in V.$ ■

Examples of overlapping concepts are shown in Figure 3.5. Among them, for instance, *Body fluid sample* is a base overlapping root because both of its parents, *Body substance sample* and *Fluid sample*, are non-overlapping concepts. They are, in fact, partial-area roots. Another example is *Acellular blood (serum or plasma) specimen* with the non-overlapping parents *Fluid sample* and *Drug specimen*.

In the progressive build-up of knowledge that is a concept hierarchy, the significance of a base overlapping root is that it lies at the confluence of multiple independent lines of knowledge—originating from the roots of the area. In this sense, such a concept can be seen as denoting a change of conceptual context within the hierarchy as one moves downward. The roots of a partial-area are significant in terms of unique sets of relationships. The base overlapping roots do not differ from their partial-area roots in regard to their relationships (they have the same relationships, in fact), but each one does represent a new combination in the downward direction of individual knowledge artifacts, each of which was first expressed by some partial-area root.

With the definition of base overlapping root now in place, the general notion of overlapping root can be defined in a recursive manner as follows.

Definition (Overlapping Root): A concept $L \in V$ is an *overlapping root* if either (1) it is a base overlapping root; or there exist concepts C_1 and C_2 ($C_1 \neq C_2$) such that *desc*(L, C_1), *desc*(L, C_2), and either (2) C_1 is an overlapping root and C_2 is a partial-area root or (3) both C_1 and C_2 are overlapping roots. For both Cases (2) and (3), the hierarchical paths from L to C_1 and from L to C_2 do not contain other (intermediate) overlapping roots. ■

Note that the qualifying pair of ancestors (C_1, C_2) is not necessarily unique. That is, more than one pair of ancestors might satisfy the requirements. The definition of overlapping root is well illustrated in Figure 3.5. The yellow concepts, *Body fluid sample* and *Acellular blood (serum or plasma) specimen*, are base overlapping roots (Case (1)). The orange concept *Blood specimen* follows Case (2) since one parent, *Body fluid sample*, is an overlapping root and the other, *Drug specimen*, is a partial-area root. Finally, the green concepts *Serum specimen* and *Serum specimen from blood product* are overlapping roots according to Case (3) since each is a child of two overlapping roots.

Case (1) denotes the fact that base overlapping roots, defined above, form the foundation upon which other overlapping roots are defined. Cases (2) and (3) of the definition (the recurrences) designate certain points in the hierarchy below the level of the base overlapping concepts as being significant convergences of knowledge and thus warranting new grouping structures. A concept satisfying Case (2) or Case (3) in particular is called a *derived overlapping root*.

In Figure 3.5, *Blood specimen* is a derived overlapping root according to Case (2). Its two qualifying ancestors are its parents *Body fluid sample*, a base overlapping root, and *Drug specimen*, a partial-area root. *Serum specimen* and *Serum specimen from blood product* are also derived overlapping roots. The two parents of *Serum specimen* are base overlapping roots. On the other hand, the two parents of *Serum specimen from blood product* are both derived overlapping roots.

The excerpt of the Specimen hierarchy's area {*substance*} in Figure 3.6(a)—some of which already seen in Figure 3.5—shows six of its overlapping roots, highlighted with multi-coloring. (All lines in the figure are IS-As.) This coloring scheme allows for easy identification of an overlapping root's respective partial-area root ancestors. The three partial-area roots are the single-colored concepts on the top level of the figure. For example, *Body fluid sample*, colored orange and blue on the second level, is an overlapping root that is a descendant of *Body substance sample* (orange) and *Fluid sample* (blue). In fact, it happens to be a child of both and is thus a base overlapping root. In Level 2, another base overlapping root *Acellular blood (serum or plasma) specimen* is found, colored blue and yellow, as well as the non-overlapping concept *Stool specimen*, a descendant of only one partial-area root *Body substance sample*. *Fecal fluid sample*, colored orange and blue on Level 3, is also a base overlapping root due to the fact that its two parents are non-overlapping concepts. The derived overlapping roots begin to appear on that level, too. They are the two concepts *Blood specimen* and *Serum specimen*, both colored orange, blue, and yellow. *Blood specimen* is a child of one base overlapping root, *Body fluid sample*, in Level 2 and one partial-area root, *Drug specimen* (see Case (2) of the definition). *Serum specimen* is a child of the two base overlapping roots in Level 2

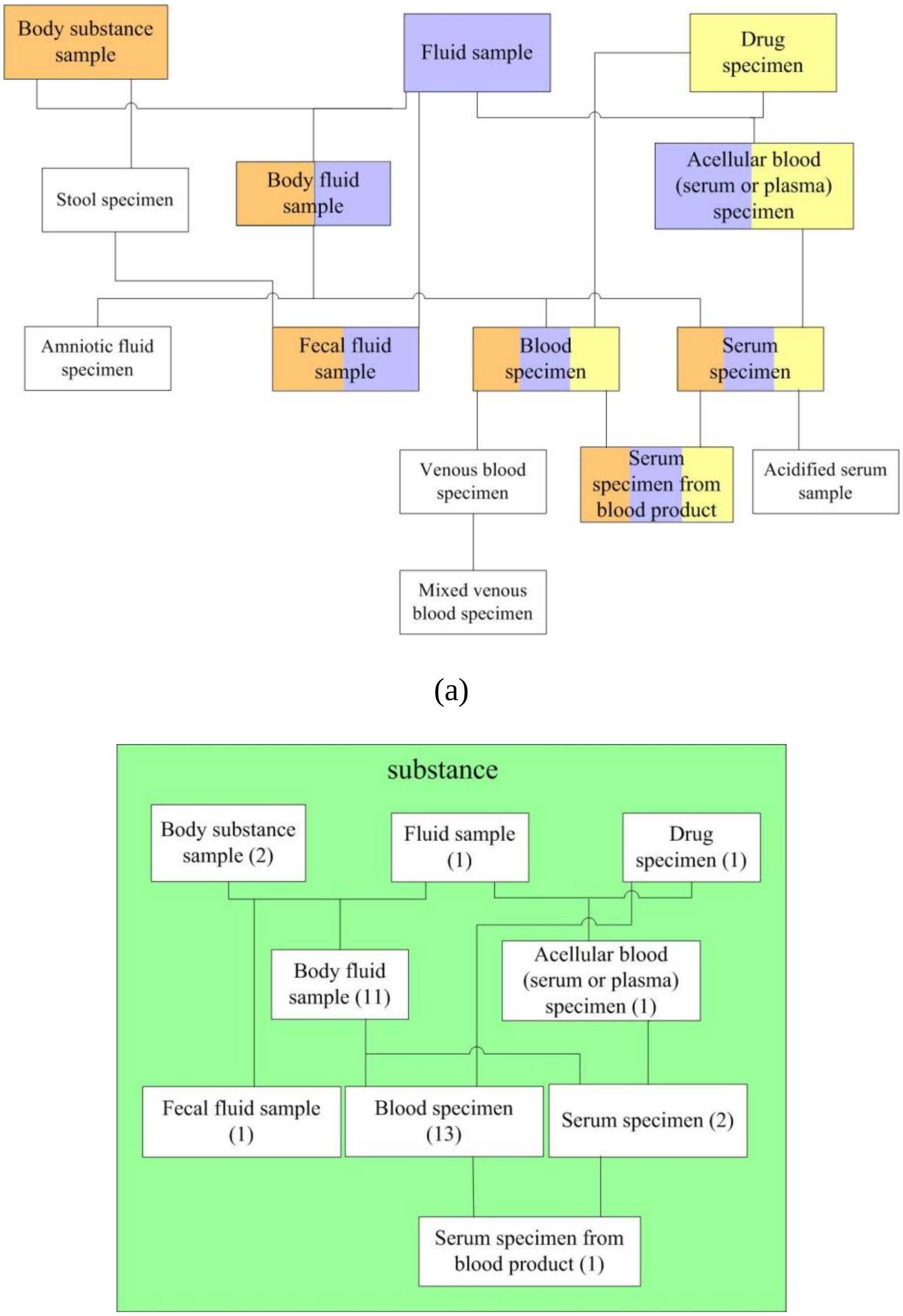


Figure 3.6 (a) Some overlapping roots (shown as multi-colored boxes) from the area {*substance*} in the Specimen hierarchy; (b) corresponding excerpt of the d-partial-area taxonomy representation of {*substance*}, where the embedded boxes are d-partial-areas.

(Case (3)). Note that both have descendants that are not overlapping roots (e.g., *Mixed venous blood specimen*). The last derived overlapping root *Serum specimen from blood product* is found in Level 4.

It should be noted that overlapping concepts having a single parent cannot be overlapping roots. Again, the purpose of this designation is to highlight knowledge convergence points for which multiple parents are necessary. As an example, the concept *Acidified serum sample* has as its only parent the derived overlapping root *Serum specimen* and is thus not an overlapping root (see Figure 3.6(a)). Similarly, the derived overlapping root *Blood specimen* has 12 descendants, such as *Whole blood sample*, *Arterial blood specimen*, and *Cord blood specimen*, none of which are overlapping roots. (Note that these descendants are not shown in the excerpt in Figure 3.6(a). They will be shown in the full figure in Figure 3.8.) As these examples demonstrate, there are overlapping concepts that are not overlapping roots, even though their parents are derived overlapping roots.

3.2.3 Disjoint Partial-Areas

With the definition of overlapping root in place, one can now proceed to establish a partition of an entire area whose partial-areas overlap. Moreover, each of the concept groups collectively forming the partition will be singly-rooted. Such concept groups are referred to as *disjoint partial-areas* (*d-partial-areas,* for short). The initial set of d-partial-areas is derived by removing those portions of the original partial-areas that constitute overlaps, leaving only non-overlapping concepts. For example, the d-partial-area *Body substance sample* contains one additional concept *Stool specimen* beyond its root. It is obtained from the original partial-area of the same name having 47 total concepts by removing the overlapping roots *Fecal fluid sample* and *Body fluid sample* along with the latter's descendants (see Figure 3.6(a)). Clearly, such d-partial-areas are all disjoint with respect to each other and also with respect to the entire set of overlapping concepts. And they are each singly-rooted.

The remainder of the d-partial-areas are created in the context of the set of overlapping concepts based on the overlapping roots. In fact, each overlapping root will be the root of its own newly derived d-partial-area. Intuitively, such a d-partial-area is the portion of the area residing "between" an overlapping root, say, C_R and the descendants of C_R that are also overlapping roots. For example, consider the overlapping root *Body fluid sample*. The concepts that are removed in order to form its d-partial-area are the overlapping root child *Blood specimen* along with all its respective descendants and the other overlapping root child *Serum specimen* with its two children (see Figure 3.6(a)). The concepts that are left in the d-partial-area rooted at *Body fluid sample* are, besides itself, its seven children (e.g., *Amniotic fluid specimen*) and its grandchildren which are children of the child *Cerebrospinal fluid sample*. (Note that only one such child is shown in Figure 3.6(a), as it is an excerpt. All ten descendants appear in the full figure in Figure 3.8.)

More formally, let C_R be an overlapping root. Then it is designated as the root of its own d-partial-area with the name "C_R." Furthermore, let C be an overlapping concept—but not an overlapping root—which is a descendant of C_R such that there are no other overlapping roots on the paths between C and C_R. Then C is a member of the d-partial-area C_R. For example, consider the overlapping root *Blood specimen* and its descendant *Mixed venous blood specimen* in Figure 3.6(a). Since the intermediate concept *Venous blood specimen* on the only path from *Mixed venous blood specimen* to *Blood specimen* is not an overlapping root, *Mixed venous blood specimen* belongs to the d-partial-area *Blood specimen*. It is possible to prove that C_R is unique for any given C, and hence C's membership in a d-partial-area is well-defined. Moreover, it is possible to prove that for each overlapping concept C there is always such a C_R.

3.2.4 Disjoint Partial-area Taxonomy

From the d-partial-areas, an abstraction network is formed, which enhances the partial-area taxonomy framework introduced in Chapter 2 and highlights the structural subtleties of the overlapping portions of the partial-areas. This new network is called *the disjoint partial-area taxonomy* (*d-partial-area taxonomy*, for short). Those d-partial-areas derived directly from the existing partial-areas—and consisting only of non-overlapping concepts—hold the same place as their predecessors in the d-partial-area taxonomy. Moreover, partial-areas originally having no overlapping concepts retain their places as nodes and are also designated d-partial-areas in the new network. The *child-of* relationships emanating from these d-partial-areas and extending into other areas are derived as done previously for the partial-areas.

The d-partial-areas comprising overlapping concepts are also elevated to the status of nodes in the d-partial-area taxonomy. Each is displayed as a box with its name (i.e., its unique overlapping root) inside and its number of concepts in parentheses. *Child-of* links are defined for these new nodes in a similar manner to those for areas and partial-areas, but here the overlapping roots play a role. Let *A* and *B* be two d-partial-areas, such that the concept *A* (the overlapping root of the former) has a parent in the latter. Then there exists a *child-of* from the d-partial-area *A* to the d-partial-area *B*. A portion of the d-partial-area taxonomy for the area {*substance*} derived from the excerpt of its hierarchy shown in Figure 3.6(a) can be seen in Figure 3.6(b). For example, there is a *child-of* from the d-partial-area *Fecal fluid sample* to the d-partial-area *Body substance sample* since, in Figure 3.6(a), there is an IS-A from the concept *Fecal fluid sample* to the concept *Stool specimen* which resides in *Body substance sample*. As can be seen in Figure 3.6(b), the d-partial-area nodes, like the partial-area nodes, are embedded in their respective area, which in this case is {*substance*}, colored green following Figure 3.1.

3.2.5 Enhanced Abstraction of the Complex Overlapping Concepts in Disjoint Partial-areas

The described taxonomies provide abstraction-level views of the content of a SNOMED hierarchy. For example, the area taxonomy (Figure 3.1) shows that there are 81 concepts having exactly the one relationship substance. The partial-area taxonomy (Figure 3.2) also conveys the overarching semantics of these concepts. There are 44 fluid samples, 23 drug specimens, 47 body substance samples, and 13 food specimens. Those four large groups constitute most of the concepts representing specimens with only one relationship to the Substance hierarchy of SNOMED. There are some other small groups, including *Gaseous material specimen* (3), *Microbial isolate specimen* (2), and *Plant specimen* (1). Reviewing this information, the user gets a summary of the content of this area. In contrast, the area {*morphology*} has just one partial-area *Lesion sample* of 14 concepts. (This consolidated view was obtained following the auditing of the 2004 release of the Specimen hierarchy supported by the taxonomies [48]. The area {*morphology*} had six partial-areas in the earlier version, but the auditing found that all fall under *Lesion sample*.)

When users want to view concepts with both *substance* and *morphology* relationships, they can utilize the area {*morphology, substance*} in the second level having 11 concepts. This area is a child of both {*substance*} and {*morphology*} (Figure 3.1). As it happens, the area has 11 partial-areas of one concept each, e.g., *Effusion sample* and *Cyst fluid sample* (Figure 3.2). As shown in Chapter 2, this view provided by the partial-area taxonomy was very helpful in exposing errors in the Specimen hierarchy.

The partial-area taxonomy view is particularly useful when the different partial-areas of an area are disjoint, but it is somewhat deficient when the partial-areas overlap. As was discussed above, those overlapping parts of a partial-area contain concepts that are semantically more complex than concepts of non-overlapping parts of the same partial-area. Furthermore, the unit of a partial-area with an overlap is not semantically uniform. Hence, the difficulty of comprehending such concepts is magnified. For example, out of the 23 drug-specimen concepts in the partial-area of that name in the area {*substance*}, 21 are also fluid samples, while 20 are also body substance samples. Furthermore, 12 concepts are both fluid samples and body substance samples. Hence, the knowledge conveyed by the partial-areas of the area {*substance*} (Figure 3.2) is hiding a more complex situation. They provide a relatively superficial perspective where a more refined view is needed. Furthermore, as shown in Figure 3.1, the area {*substance*} contains only 81 concepts, where overlapping concepts appear in multiple counts of the sizes of the partial-areas in Figure 3.2.

The desired refined view of an area with overlapping partial-areas is provided by the d-partial-area taxonomy introduced above. In Figure 3.6(b), the overlap of the three partial-areas just discussed is concentrated under two d-partial-areas: *Body fluid sample* of 11 concepts, capturing an overlap of *Body substance sample* and *Fluid sample*; and *Acellular blood (serum or plasma) specimen* of one concept, capturing an overlap of *Drug Specimen* and *Fluid sample*. In the d-partial-area taxonomy, the children of these two d-partial-areas, *Blood specimen* of 13 concepts and *Serum specimen* of two concepts, denote the overlaps of the three partial-areas. In turn, a deeper level of overlap is indicated by the grandchild d-partial-area *Serum specimen from blood product* of one concept. The names (overlapping roots) of the d-partial-areas communicate more precise knowledge of the content of the overlapping concepts. The full d-partial-area taxonomy for that portion of the area {*substance*} from which Figure 3.6(b) was extracted will appear in Figure 3.9. More such knowledge was excluded from Figures 3.6(a) and 3.6(b) for the sake of brevity and clarity.

Importantly, each d-partial-area of the overlapping concepts consists of a semantically uniform group, where its name, e.g., *Blood specimen*, characterizes the concepts of the group very well. Hence, the d-partial-area taxonomy is a vehicle for more readily comprehending the nature of the overlapping concepts. In another example corresponding to Figure 3.3, the d-partial-area taxonomy will have a minimal overlap of just one concept, *Blood bag specimen, from patient*, between the two partial-areas of Figure 3.3, *Device specimen* and *Specimen from patient*. This overlap appears as one d-partial-area, *Blood bag specimen, from patient*, containing only that concept. Note that in the d-partial-area taxonomy, this d-partial-area is the child of the two semantically uniform d-partial-areas *Device specimen* (18) and *Specimen from patient* (1), which are now uniform due to the removal of the overlapping concept (Figure 3.7). Thus, the d-partial-area taxonomy reveals both the uniform semantics of the overlapping subgroup and the precise size of its extent (by the number appearing alongside the name) as well as the uniform semantics of the d-partial-areas obtained by the removal of the overlapping concepts from the partial-areas of the partial-area taxonomy. This enhanced view afforded by the d-partial-area taxonomy supports a better auditing regimen for the complex overlapping concepts, which will be demonstrated in Chapter 4.

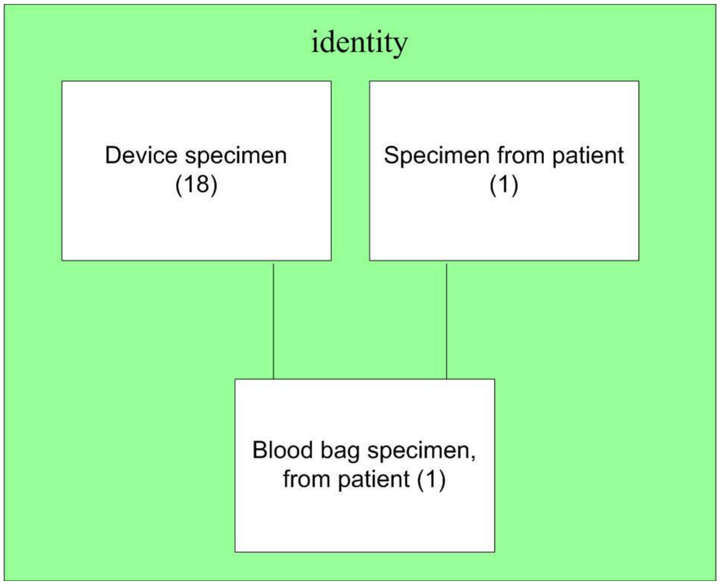

identity
Device specimen (18)
Specimen from patient (1)
Blood bag specimen, from patient (1)

Figure 3.7 The d-partial-areas *Device specimen*, *Specimen from patient*, and *Blood bag specimen, from patient* of the area {*identity*}.

There are two issues regarding the display of the d-partial-area taxonomy. One is the arrangement of d-partial-areas within an area. In the partial-area taxonomy (e.g., Figure 3.2), no *child-of* hierarchical relationships exist between partial-areas of the same area because each is based on and contains a root of the area. When one partial-area is displayed below another (see, e.g., the area {*substance*} in Figure 3.2), no hierarchical arrangement is implied. It is just a layout expediency.

In the d-partial-area taxonomy, there are *child-of*'s between d-partial-areas in a given area. In fact, any d-partial-area rooted at an overlapping root (be it base or derived) has multiple *child-of*'s to other d-partial-areas of the same area. To reflect the hierarchical nature of these *child-of*'s, the author try to position the d-partial-areas such that they are below their respective parents, and the *child-of*'s are in an upward direction.

As a result, there is a contrast between the detailed display of an area of many overlapping concepts, such as {*substance*} in Figure 3.6(b), and an area without overlapping concepts, such as {*morphology*}. The d-partial-area taxonomy contains both kinds of areas. Thus, there is a disparity in the display of these two kinds of areas in regard to their nature and level of detail. It will be discussed in the following sections that the three taxonomies are best used in concert in a kind of multi-scale display.

3.3 Results

The July 2007 release of the Specimen hierarchy of SNOMED consists of 1,056 active concepts, of which 162 are overlapping. The July 2007 release has been used in this chapter because in Chapter 4, the application of a systematic auditing regimen to both the July 2007 and 2009 releases will be reported. The partial-area taxonomy and the d-partial-area taxonomy for July 2009, whose contents were affected by the audit of the July 2007 release, will appear in Chapter 4. Most of the overlapping concepts reside in Level l areas, i.e., those having one relationship. In fact, roughly one third (155 out of 468) of the Level 1 concepts are overlapping, and these are found primarily in {*topography*} and {*substance*}. Overlapping concepts also appear in the partial-areas of areas with two relationships, but in far fewer numbers. In fact, there are only seven of them. Six are in {*topography*, *procedure*}, and the other is in {*topography*, *morphology*}. The statistics of the overlapping concepts in Levels 1 and 2 are given in

Table 3.1. For each area, its total number of concepts *C* (Column 2), number of overlapping concepts *V* (Column 3), the percentage of overlapping concepts (Column 4), the number of d-partial-areas with overlapping roots *D* (Column 5), and the average number of overlapping concepts per d-partial-area: *V* /*D* (Column 6), are listed. For example, {*substance*} has 81 concepts and 35 of them are overlapping (43%). It also has nine overlapping roots which head d-partial-areas, with about four concepts per each such d-partial-area, on average.

Table 3.1 Statistics of Overlapping Concepts at Levels 1 and 2

Area	***C***	***V***	***V* / *C* (%)**	***D***	**Avg = *V* / *D***
substance	81	35	43	9	3.9
topography	333	116	35	52	2.2
procedure	20	3	15	3	1.0
identity	20	1	5	1	1.0
topography, procedure	380	6	2	6	1.0
topography, morphology	18	1	6	1	1.0
Total:	**852**	**162**	**19**	**72**	**2.3**

C = # concepts; ***V*** = # overlapping concepts; ***D*** = # overlapping roots

Most overlapping concepts in the area {*topography*} are found in intersections with the partial-area *Tissue specimen* which contains 126 concepts. These results have been tabulated separately in Table 3.2. For example, the partial-area *Specimen from eye* has 18 concepts. Its intersection with *Tissue specimen* has 12 of them (67%).

The full complement of nine overlapping roots from the area {*substance*} can be seen as the multi-colored boxes in the excerpt in Figure 3.8. This figure follows the color conventions of Figure 3.6(a). The top four concepts are the area's roots. Among the overlapping roots, five are base overlapping roots and four are derived overlapping roots.

The remaining white concepts are overlapping concepts that elaborate the semantics of the overlapping roots of their respective d-partial-areas.

Table 3.2 Intersections Involving Partial-area *Tissue specimen*

Second Partial-area	***C***	***V***	***V / C* (%)**
Specimen from eye	18	12	67
Ear sample	2	1	50
Specimen from breast	8	4	50
Cardiovascular sample	13	3	23
Products of conception tissue sample	12	3	8
Genitourinary sample	73	22	27
Dermatological sample	6	2	33
Specimen from digestive system	74	30	39
Musculoskeletal sample	35	22	63
Respiratory sample	41	7	16
Endocrine sample	12	3	25
Specimen from central nervous system	4	1	25
Specimen from thymus gland	2	1	50
Specimen from trophoblast	2	1	50
Total:	**302**	**112**	**35**

C = # concepts; ***V*** = # overlapping concepts

The portion of the d-partial-area taxonomy for the area {*substance*} corresponding to the concept diagram in Figure 3.8 is shown in Figure 3.9. It presents a precise abstraction of the configuration of the overlapping concepts within {*substance*}. Note that the numbers of concepts listed for the top-level d-partial-areas are actually the numbers of non-overlapping concepts appearing in the original partial-areas from which these d-partial-areas are derived. For example, *Drug specimen* (2) has the two non-overlapping concepts from the partial-area of the same name, containing a total of 23 concepts, in Figure 3.2. They are the area root *Drug specimen* plus a non-overlapping child not shown in Figure 3.8. The entire content of the partial-area *Drug specimen* is distributed among the d-partial-*area Drug specimen* and all its descendants. This can be seen by summing up the numbers of concepts in those d-partial-areas: 2 + 1 + 13 + 2 + 4 + 1 = 23. The same holds true for the other top-level d-partial-areas and their respective descendants in Figure 3.8.

The complete node for {*substance*} in the d-partial-area taxonomy is shown in Figure 3.10, which differs from Figure 3.9 only in the inclusion of the six additional d-partial-areas derived from the corresponding six partial-areas (Figure 3.2) that do not contain any overlapping concepts. The isolation of these d-partial-areas from the others conveys the absence of overlaps. Overall, this network can be used, for example, as a vehicle for comprehending the details of the kinds of overlapping concepts and their numbers in the underlying SNOMED hierarchy.

Figure 3.11 provides a larger excerpt of the portion of the d-partial-area taxonomy appearing within the area {*topography*}, highlighting the extensive overlapping among its partial-areas. As shown in Table 3.1, this area has 116 overlapping concepts distributed among 52 d-partial-areas. Most of the overlapping concepts *have Tissue specimen* as one of their partial-areas, as listed in Table 3.2. In the top level of Figure 3.11, 15 d-partial-areas are obtained by removing all overlapping concepts from the original partial-areas. On the next level down, 13 d-partial-areas are found having base overlapping roots. Two d-partial-areas with derived overlapping roots appear on the bottom level. Many other d-partial-areas with few concepts have been omitted. Again, it should be noted that the intersection of two partial-areas may contain several overlapping roots. For example, the intersection of *Tissue specimen* and *Cardiovascular sample* has

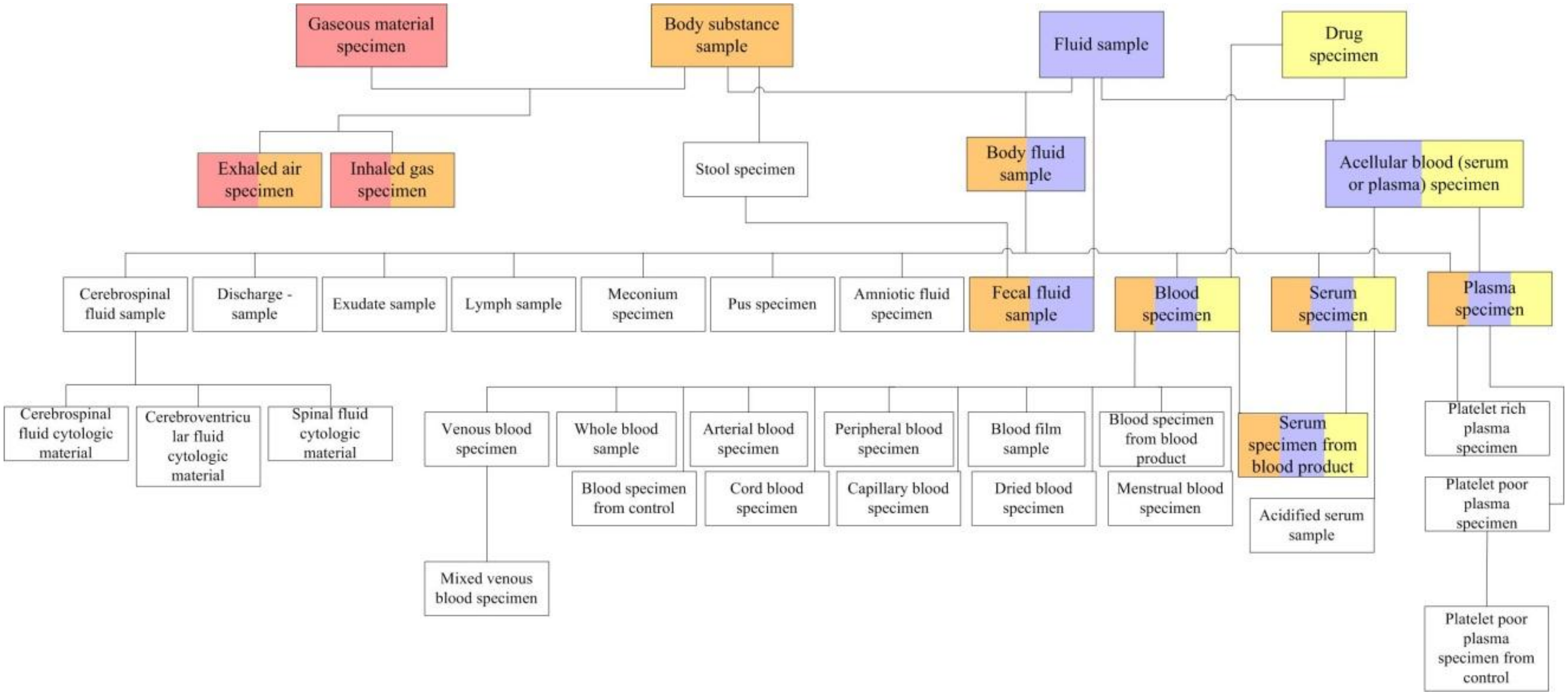


Figure 3.8 The nine overlapping roots from the area {*substance*} are shown as multi-colored boxes among other concepts.

three overlapping roots, as shown in the figure: *Tissue specimen from heart*, *Heart valve tissue*, and *Native heart valve sample*.

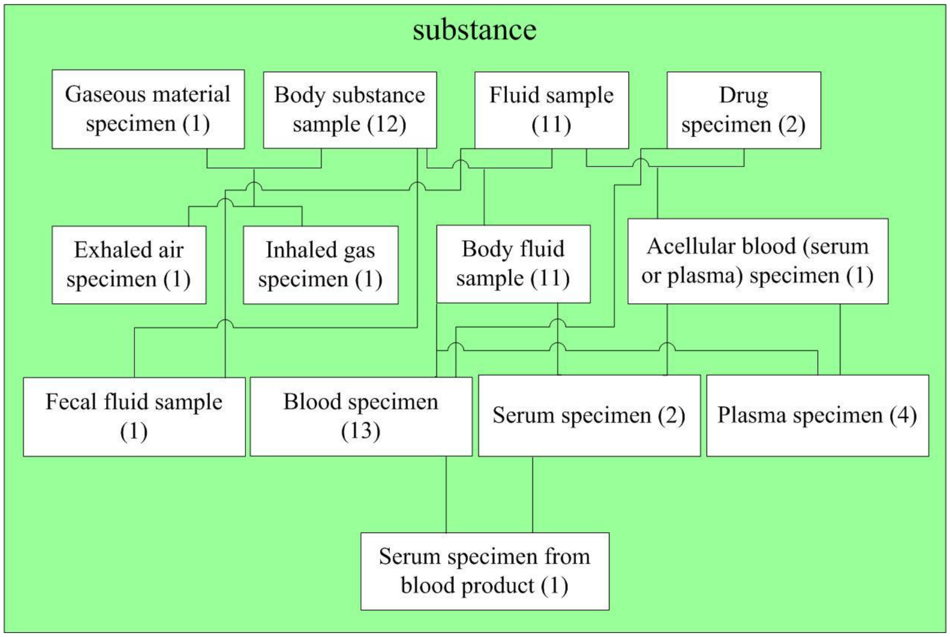


Figure 3.9 The d-partial-area taxonomy excerpt consisting of 13 d-partial-areas corresponding to the concept network appearing in Figure 3.8.

To illustrate the general applicability of this abstraction approach, the author has applied it to all seven of the SNOMED hierarchies that have outgoing lateral relationships. (The other 12 hierarchies have no such relationships, rendering this methodology inapplicable to them.) The results are listed in Table 3.3. For each of the seven hierarchies, the table gives its total number of concepts, the number of overlapping concepts and their percentage, and the number of overlapping roots. For example, the Pharmaceutical Product hierarchy has a total of 17,410 concepts, of which 1,047 are overlapping (6.1%). The number of overlapping roots is 949. Note that in Pharmaceutical Product almost all the overlapping concepts are overlapping roots (1,047 compared to

949). As it happens, the hierarchies Event and Body Structure have no overlapping concepts whatsoever.

Table 3.3 Concept Distributions in Seven SNOMED Hierarchies

Hierarchy	***C***	***V***	***V/C* (%)**	***D***	C_{mult}	C_{mult}/C **(%)**	V_{mult}
Event	3,661	0	0	0	86	2.4	0
Situation	3,237	86	2.7	67	387	12.0	67
Pharmaceutical Product	17,140	1,047	6.1	949	7,721	45.1	963
Procedure	52,687	7,878	15.0	3,374	27,031	51.3	5,846
Specimen	1,330	191	14.4	80	788	59.3	130
Body Structure	31,155	0	0	0	13,418	43.1	0
Clinical Finding	98,414	13,943	14.2	3,127	44,544	45.3	9,841
Total:	**207,624**	**23,145**	**11.2**	**7,597**	**93,975**	**45.3**	**16,847**

C = # concepts; ***V*** = # overlapping concepts; ***D*** = # overlapping roots; C_{mult} = # concepts having multiple parents; V_{mult} = # overlapping concepts having multiple parents

As a point of comparison, Table 3.3 lists the number of concepts having multiple parents (and their percentage), along with the number of overlapping concepts having that characteristic. These numbers will be discussed further below. The Pharmaceutical Product hierarchy has 7,721 concepts (45%) with multiple parents, of which only 963 (5.6%) are overlapping. As can be seen, there are only 14 (= 963 − 949) non-root overlapping concepts having multiple parents. Note that 84 (= 1047 − 963) overlapping concepts have only one parent.

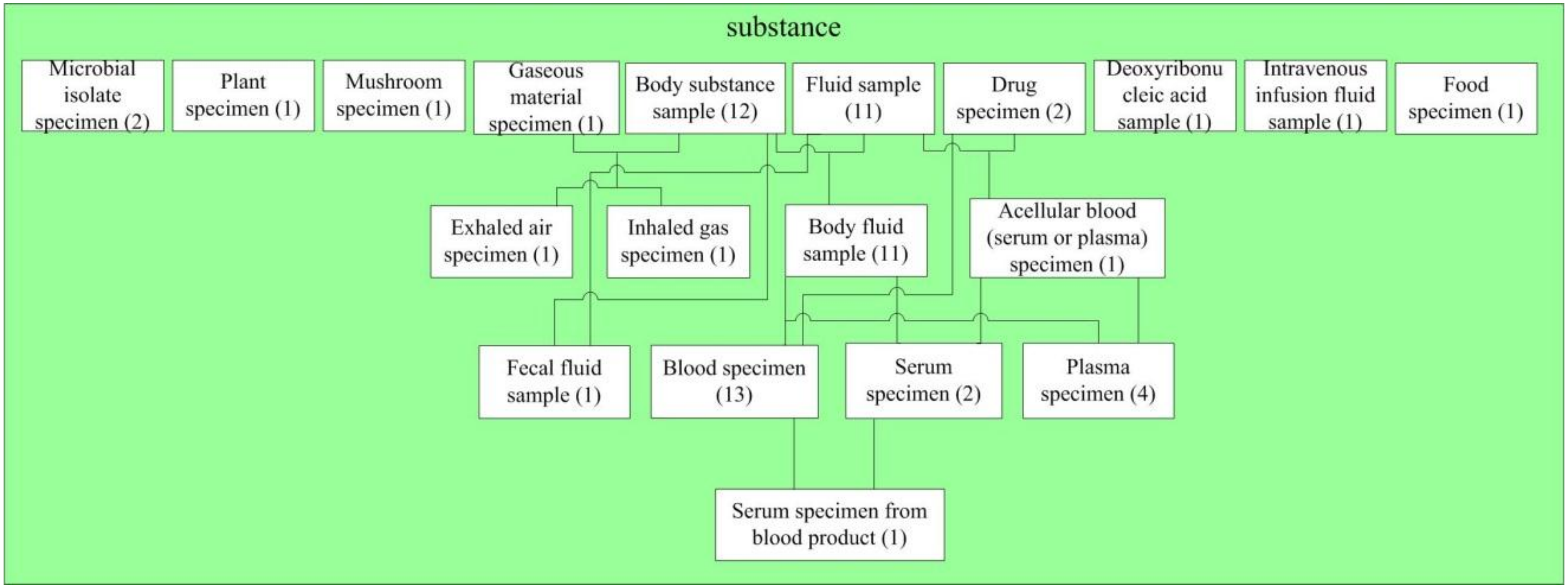


Figure 3.10 The d-partial-area taxonomy node for the area {*substance*} containing 19 embedded d-partial-areas. The numbers in parentheses indicate the numbers of concepts in the respective d-partial-areas.

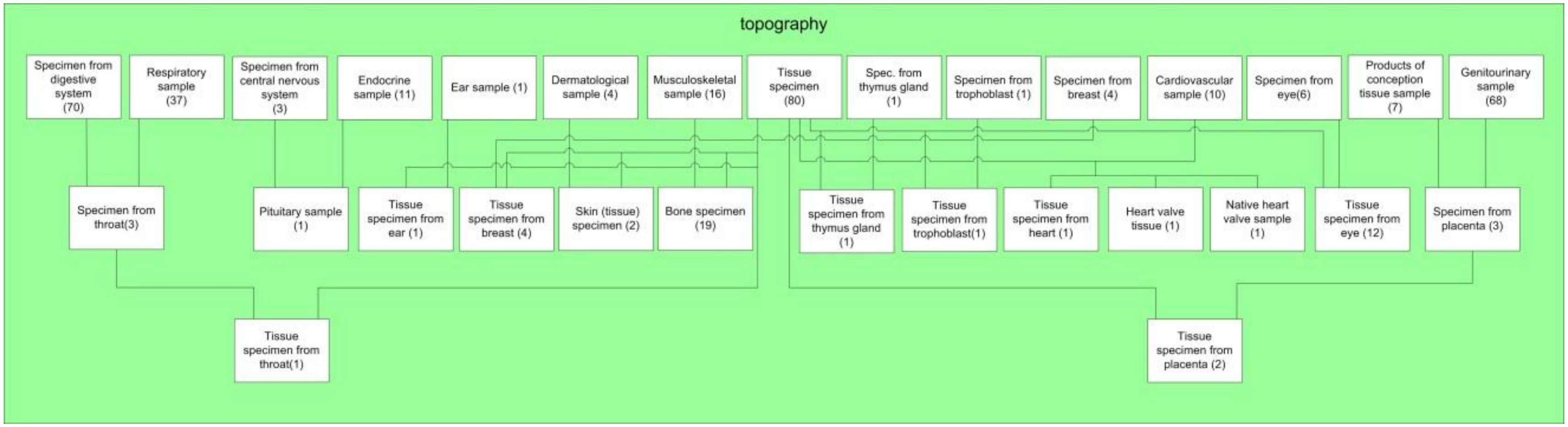

topography
Specimen from digestive system (70)
Respiratory sample (37)
Specimen from central nervous system (3)
Endocrine sample (11)
Ear sample (1)
Dermatological sample (4)
Musculoskeletal sample (16)
Tissue specimen (80)
Spec. from thymus gland (1)
Specimen from trophoblast (1)
Specimen from breast (4)
Cardiovascular sample (10)
Specimen from eye(6)
Products of conception tissue sample (7)
Genitourinary sample (68)
Specimen from throat(3)
Pituitary sample (1)
Tissue specimen from ear (1)
Tissue specimen from breast (4)
Skin (tissue) specimen (2)
Bone specimen (19)
Tissue specimen from thymus gland (1)
Tissue specimen from trophoblast(1)
Tissue specimen from heart (1)
Heart valve tissue (1)
Native heart valve sample (1)
Tissue specimen from eye (12)
Specimen from placenta (3)
Tissue specimen from throat(1)
Tissue specimen from placenta (2)

Figure 3.11 An excerpt of the d-partial-area taxonomy for the area {*topography*} consisting of 30 d-partial-areas.

3.4 Discussion

3.4.1 Taxonomy Support for Presentation of Terminology Content

The value of a terminological knowledge base depends on the accuracy and reliability of its constituent knowledge. This is true from the perspective of both ad hoc users and developers of software systems, such as EHR software and decision-support systems, that are dependent on that knowledge. Moreover, the ability to visualize and assess the knowledge's underlying structural organization is a critical factor contributing to terminology usability, deployment, and maintenance. The area and partial-area taxonomy abstraction networks have been shown to support maintenance efforts for SNOMED [48, 49] and the NCIt [46]. However, in this chapter, some deficiencies in these abstraction networks have been discussed regarding complex portions of the terminology involving what is called overlapping concepts. The d-partial-area taxonomy that was introduced extends the area taxonomy paradigm to more properly present the overlapping concepts by highlighting semantically uniform groups and their sizes. For example, Figure 3.9 highlights the groups *Blood specimen* (13), *Serum specimen* (2), and *Plasma specimen* (4), which were originally hidden but tacitly accounted for multiple times in *Body substance sample* (47*), Fluid sample* (44), and *Drug specimen* (23) in Figure 3.2.

In Figure 3.11, showing the area {*topography*}, only two d-partial-areas with derived overlapping roots are found. More than twice that number is found, with many concepts in their d-partial-areas, in the excerpt of {*substance*} in Figure 3.9. What is seen in {*topography*} is extensive overlapping with many base overlapping roots but not as complex a pattern as is found in {*substance*}. An interesting finding revealed by Figure

3.11 is that *Products of conception tissue sample*, the second d-partial-area from the right in Level 1, represents a modeling error. Its root should not actually have been a root but rather an overlapping concept of *Tissue specimen* and *Genitourinary sample*.

In this chapter, the complexity of overlapping concepts was studied, finding what is called “overlapping roots” that represent the convergence of multiple hierarchical paths originating at the roots of an area (see Section 3.2.1). A variety, called “base overlapping root”, is less complex than the “derived overlapping root.” Within the latter, different kinds have been identified according to Cases (2) and (3) of the definition (see Section 3.2.2). The organizational subtleties of the various kinds of overlapping concepts are abstracted in the d-partial-area taxonomy which was introduced in Section 3.2.4. The network breaks down the highly tangled group of overlapping concepts of an area into subsets in a manner that summarizes their hierarchical configuration and supports orientation into their nature. This phenomenon is demonstrated, for example, in Figure 3.10, where nine d-partial-areas (rooted at derived overlapping roots) on Levels 2 and 3 expose the very complex modeling of the 35 overlapping concepts in a clear and unambiguous way, while all this knowledge is hidden “under the hood” in the partial-area taxonomy of Figure 3.2. The refined view helps in assessing the correctness of the modeling of this highly complex portion of the SNOMED hierarchy.

3.4.2 Further Applicability of the Methodology

While the abstraction methodology presented in this chapter was formulated in the context of SNOMED, its applicability extends to other DL-based terminologies such as the NCIt. Moreover, terminologies such as Kaiser-Permanente’s CMT [62] and the VA’s ERT [61], that have been derived in part from SNOMED, may prove to be fertile grounds for additional applications. By 2015, SNOMED is slated to become a standard for problem-list encoding in EHRs under the HITECH initiative [2]. It is thus reasonable to assume that further derivatives from SNOMED will emerge. SNOMED's design, in fact, anticipates the need for extensions and subsets in order to craft terminological artifacts that are tuned to the needs of individual hospitals and other organizations. Its "reference set specification" [67] serves the purpose of extracting components of SNOMED tailored to particular organizational preferences and use-cases.

3.4.3 Limitations and Future Work

The area and partial-area taxonomies are available only for DL-based terminologies. Abstraction of terminologies is very delicate, and no one model of abstraction networks is expected to fit all terminologies. However, more research is needed to explore abstraction networks for other families of terminologies and terminological systems. The benefits obtained from abstraction networks in regard to auditing should motivate more research in this direction.

A limitation of the taxonomy approach is that it depends on the existing relationships defined for a hierarchy of SNOMED. Hence, the methodology of this research is not applicable to a SNOMED hierarchy without any outgoing relationships at all. An initial effort to handle such a hierarchy based on converse relationships appeared in [68]. Moreover, the d-partial-area taxonomy is only pertinent when there are overlapping partial-areas within the partial-area taxonomy. Otherwise, the two taxonomies are identical.

In general, an abstraction network should represent a significant reduction in size (i.e., number of nodes) vis-à-vis its underlying concept network. For the Specimen hierarchy, the area taxonomy provides a 0.023 reduction factor (24 areas versus 1,056 concepts). The partial-area taxonomy has a reduction factor of 0.34 (361 partial-areas versus 1,056 concepts). The d-partial-area taxonomy only has a reduction factor of 0.41 (433 d-partial-areas versus 1,056 concepts). Note that the higher reduction factor for the d-partial-area taxonomy is the justifiable price paid for the enhanced view obtained by the inclusion of the d-partial-areas that abstract the more complex overlapping concepts. There is no impact on the representation of those partial-areas experiencing no overlap in the partial-area taxonomy. Experiments with more SNOMED and NCIt hierarchies of various sizes are needed to shed more light on reduction factors obtained for various kinds of taxonomies. Also note that the relatively high reduction factor for the partial-area taxonomy is a result of a large number of partial-areas containing just one concept each (so-called "singletons"). As was shown in Chapter 2, such partial-areas tend to signal errors. It is interesting to see if the number of such partial-areas will decrease as a result of auditing them. An initial promising result is brought up in [69]. Further research into this issue is required.

The reduction factors aside, the three taxonomies complement each other in terms of granularity of display, with a zooming effect achieved as one moves successively through them starting from the area taxonomy. When used together in this manner, they provide a multi-scale display. The area taxonomy offers a global view of the hierarchy's layout and the partial-area taxonomy provides a more semantically focused view of the areas, whereas the real benefits of the d-partial-area taxonomy are seen at the local level—on the scale of an individual area—where it helps to reveal the complexity of the configuration of the overlapping concepts.

One might question whether there are simpler ways to identify “complex” concepts rather than having to go through the abstraction analysis presented in this chapter. For example, one might choose to consider the easily identified concepts having multiple parents as being complex. Note that the overlapping concepts are not simply a subset of the multi-parent concepts. Only a root overlapping concept must have, by definition, more than one parent. As seen in Table 3.1, only 72 out of 162 overlapping concepts, in the Specimen hierarchy of July 2007, are overlapping roots. The other overlapping concepts have mostly a single parent. See also Figure 3.8 where only the nine overlapping roots are multi-parented. Similar statistics are seen in Table 3.3. In the seven hierarchies for which the analysis is applicable, a total of 93,975 multi-parented concepts (45.3%) were found. In that same context, there are a total of 23,145 overlapping concepts, with 16,847 being multi-parented.

3.5 Summary

SNOMED is one of the leading terminologies being used in a variety of applications worldwide. However, it contains hundreds of thousands of concepts and has an inherent complexity that could hinder its further adoption as well as its ongoing maintenance. A new abstraction network, called the disjoint partial-area taxonomy, has been introduced to provide a better high-level view of portions of a SNOMED hierarchy containing concepts of a particularly complex nature. It refines the previous abstraction network, the partial-area taxonomy, for SNOMED introduced in Chapter 2. The new network focuses on the location and number of such complex concepts and highlights their modeling and local neighborhoods. Overall, users are provided with a summary account of the “lay of the land" that can facilitate orientation to and assessment of SNOMED's content. The methodology was demonstrated by applying it to SNOMED's Specimen hierarchy. In Chapter 4, a systematic auditing regimen based on the disjoint partial-area taxonomy will be presented, demonstrating its utility to terminology maintenance personnel.

CHAPTER 4

AUDITING OVERLAPPING CONCEPTS OF SNOMED USING A REFINED HIERARCHICAL ABSTRACTION NETWORK

4.1 Introduction

One of the driving themes in this dissertation has been that "complex" concepts, as defined by various criteria, are worth concentrating on in auditing efforts. By their very nature, such concepts are more difficult to model and should therefore be scrutinized more closely by auditors. In Chapter 3, a category of *complex* concepts, referred to as *overlapping concepts*, is identified based on the partial-area taxonomy. The author presented a methodology for hierarchically clustering such concepts and automatically constructing a novel abstraction network for their presentation. A portion of the new network, the *disjoint partial-area taxonomy*, is a directed acyclic graph of nodes representing groups of overlapping concepts where increased conceptual complexity is encountered as one navigates downward in the terminological hierarchy.

This chapter continues to follow the theme of focusing auditing on complex concepts. A methodology for auditing the overlapping concepts based on the disjoint partial-area taxonomy presented in Chapter 3 is introduced. The methodology constitutes a systematic review of the overlapping concepts as determined by their hierarchical ordering within the disjoint partial-area taxonomy. The methodology is applied to the July 2009 release of SNOMED's Specimen hierarchy. The results are compared to those obtained from an audit carried out on the July 2007 release and based on a preliminary methodology that also focused on overlapping concepts [70].

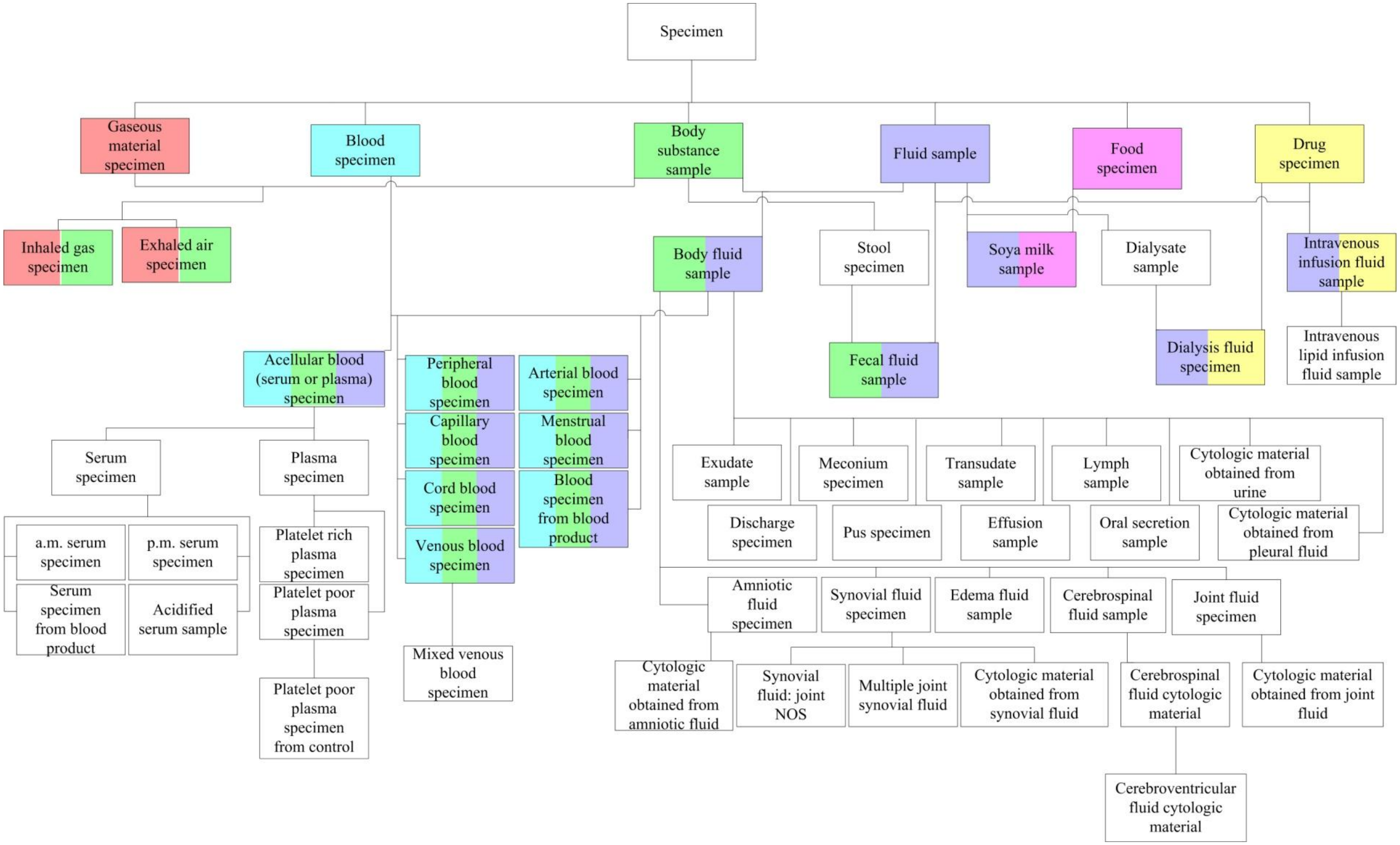

Specimen
Gaseous material specimen
Blood specimen
Body substance sample
Fluid sample
Food specimen
Drug specimen
Inhaled gas specimen
Exhaled air specimen
Body fluid sample
Stool specimen
Soya milk sample
Dialysate sample
Intravenous infusion fluid sample
Fecal fluid sample
Dialysis fluid specimen
Intravenous lipid infusion fluid sample
Acellular blood (serum or plasma) specimen
Peripheral blood specimen
Arterial blood specimen
Capillary blood specimen
Menstrual blood specimen
Cord blood specimen
Blood specimen from blood product
Venous blood specimen
Mixed venous blood specimen
Serum specimen
Plasma specimen
a.m. serum specimen
p.m. serum specimen
Serum specimen from blood product
Acidified serum sample
Platelet rich plasma specimen
Platelet poor plasma specimen
Platelet poor plasma specimen from control
Exudate sample
Meconium specimen
Transudate sample
Lymph sample
Cytologic material obtained from urine
Discharge specimen
Pus specimen
Effusion sample
Oral secretion sample
Cytologic material obtained from pleural fluid
Amniotic fluid specimen
Synovial fluid specimen
Edema fluid sample
Cerebrospinal fluid sample
Joint fluid specimen
Cytologic material obtained from amniotic fluid
Synovial fluid: joint NOS
Multiple joint synovial fluid
Cytologic material obtained from synovial fluid
Cerebrospinal fluid cytologic material
Cytologic material obtained from joint fluid
Cerebroventricular fluid cytologic material

Figure 4.1 The 15 overlapping roots from the area {*substance*} of the Specimen hierarchy (July 2009) are shown as multi-colored boxes among other concepts. The coloring indicates their ancestry.

4.2 Methods

Different auditing methodologies are applied in the first phase and the second phase of this study. The former is with respect to the July 2007 release of SNOMED, when all the overlapping concepts were reviewed without utilizing any grouping structures or ordering; the latter, with respect to the July 2009 release, when topological ordering was employed in auditing.

4.2.1 Phase 1: Unordered Auditing

As discussed in Chapter 3, the overlapping concepts are complex concepts due to their multiple classification with respect to the partial-area taxonomy and are thus targeted for auditing. For Phase 1, two domain experts, Dr. Gai Elhanan, Chief Medical Information Officer of Halfpenny Technologies, and Junchuan Xu, MD, were called upon, each of whom has training in medicine as well as training and experience in medical terminologies. The overlapping concepts of the July 2007 Specimen hierarchy are reviewed individually by each of the two auditors. The concepts are presented to the auditors with the following data for each: concept ID, preferred term, area, and d-partial-area. The auditor is given a standardized form containing two fields for completion. The first field is used to indicate the error type (if any). The choice is to be made from a menu of seven types of errors: incorrect parent, missing parent, incorrect child, missing child, incorrect relationship type, missing relationship, and incorrect relationship target. The second field is used by the auditors to suggest a correction for the error discovered.

The auditors' review in this phase involves the examination of all overlapping concepts without regard to any specific order [70]. After that, the two auditors together review concepts for which their individual reports differ, and analyze the discrepancies until a consensus is reached. A consensus report is then given to Dr. Kent A. Spackman—who is currently the Chief Terminologist of IHTSDO [3]—for further review. Only his accepted results are reported for Phase 1.

4.2.2 Phase 2: Topologically Ordered Auditing

As discussed in Chapter 3, some overlapping concepts are seen to be more complex than others when moving down through the hierarchy. With this idea in mind, the following auditing regimen is proposed that utilizes the paradigm of "group-based" auditing [48]. In the group-based approach applied to overlapping concepts, the concepts are reviewed in groups exhibiting semantic uniformity, that is, all the overlapping concepts of a d-partial-area are reviewed together with an eye toward the overlapping root which expresses the overarching semantics of the group. Furthermore, the concepts in the immediate neighborhoods of the overlapping concepts (consisting of parents, children, siblings, and targets of relationships) are audited. This "neighborhood auditing" may help to uncover propagated errors, which might otherwise be missed if the review were limited to the overlapping concepts alone.

Since SNOMED is DL based, relationships are inherited by a child concept from its parent(s) along the IS-A hierarchy. Thus, an error such as an incorrect relationship will be inherited, too. Furthermore, even an error such as an omitted relationship may be "inherited" in the sense that if it is missing from the parent, it will probably be missing from the child (unless it is explicitly defined at the child).

As a consequence, it is preferred in an audit of a group of hierarchically related concepts that the review follow a top-down order. Following such an order may help in detecting more errors as well as in accelerating the review process. In particular, when a child is scrutinized, the auditor is already aware of any errors with the parents and is alert to their potential propagation. The topological sort [71] of a directed acyclic graph (DAG)—the structure exhibited by a SNOMED hierarchy—offers a traversal of concepts in a manner where each is processed only after all its parents have been processed. Because the d-partial-areas and their *child-of* relationships also constitute a DAG [50], the disjoint partial-area taxonomy enables the utilization of the topological sort order at two different levels: the d-partial-area level and the concept level, with the latter nested in the former.

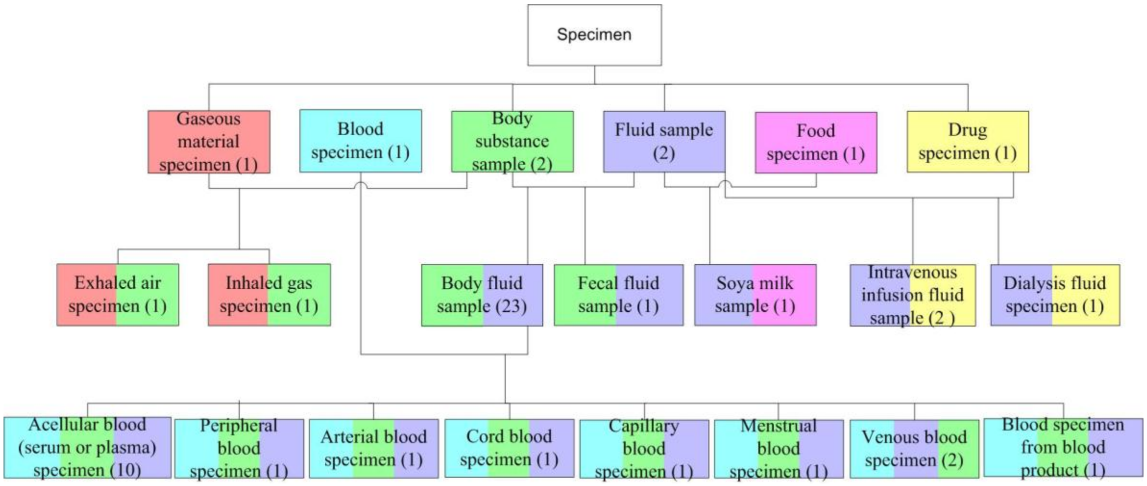

Specimen
Gaseous material specimen (1)
Blood specimen (1)
Body substance sample (2)
Fluid sample (2)
Food specimen (1)
Drug specimen (1)
Exhaled air specimen (1)
Inhaled gas specimen (1)
Body fluid sample (23)
Fecal fluid sample (1)
Soya milk sample (1)
Intravenous infusion fluid sample (2)
Dialysis fluid specimen (1)
Acellular blood (serum or plasma) specimen (10)
Peripheral blood specimen (1)
Arterial blood specimen (1)
Cord blood specimen (1)
Capillary blood specimen (1)
Menstrual blood specimen (1)
Venous blood specimen (2)
Blood specimen from blood product (1)

Figure 4.2 The portion of the disjoint partial-area taxonomy for the area {*substance*} corresponding to the concept network in Figure 4.1 (July 2009).

The following describes the auditing methodology for overlapping concepts based on the disjoint partial-area taxonomy. It should be noted that overlapping roots come in two varieties: *base* and *derived*. The details can be found in Chapter 3. The important distinction between the two in this context is that the base overlapping roots occur toward the top of the concept hierarchy and are above all the derived overlapping roots. Also note that some d-partial-areas do not have any overlapping concepts at all. They are the ones at the very top of the disjoint partial-area taxonomy that were residually left over after the lower-level d-partial-areas—containing overlapping concepts—were removed from their original partial-areas. For example, the top d-partial-area *Drug specimen* (1), comprising a single, non-overlapping concept, was left over as a result of extracting the d-partial-areas *Intravenous infusion fluid sample* (2) and *Dialysis fluid specimen* (1) (see Figure 4.2) from the original partial-area also named "*Drug specimen*" that contained a total of four concepts. Those upper-level d-partial-areas are not considered in this auditing methodology.

1. **Taxonomy level**: The d-partial-areas are processed in topological sort order starting with those having base overlapping roots. The processing proceeds through their children, grandchildren, etc., down to the very bottom of the disjoint partial-area taxonomy. As discussed in Chapter 3, the lower d-partial-areas are rooted at more complex overlapping concepts.
2. **Concept level**: On arrival at a particular d-partial-area in (1), all its constituent concepts are reviewed in a topological sort order starting with its unique root and progressing downwards. The concepts are presented to the auditor in an indented hierarchical (textual) format for inspection. The indented display neatly supports the top-down processing where each concept is reviewed only after all its respective parents are reviewed.

It is noted that the topological sort order leaves degrees of freedom with regards to the order with which the nodes of the graph are visited—and reviewed. For example, in a level-by-level traversal, all nodes on a given level are processed before any node on the next level. Another choice is a "preorder traversal," where the processing proceeds from a parent node to its children and even its grandchildren, assuming all their parents were already processed at that point. For the effectiveness of the auditing regimen, the preorder traversal is recommended. In this way, the scrutiny of a child follows that of the parent as quickly as possible, allowing an auditor to more readily retain knowledge of errors discovered at the parent and potentially propagating to the child.

To illustrate the Taxonomy level, the review will begin with the bicolored d-partial-areas in Figure 4.2, including *Exhaled air specimen*, *Inhaled air specimen*, etc. Once the review reaches *Body fluid sample*, the only bicolored d-partial-area with children, it proceeds to the bottom level containing eight tricolored d-partial-areas, i.e., *Acellular blood (serum or plasma) specimen*, *Peripheral blood specimen*, and so on. When all child d-partial-areas of *Body fluid sample* have been audited, the processing continues with the rest of the bicolored d-partial-areas, e.g., *Dialysis fluid specimen*. Again, the d-partial-areas of one color in Figure 4.2 do not have overlapping concepts and are therefore not part of the auditing regimen.

Within the d-partial-area *Body fluid sample*, the Concept level processing would begin with the root *Body fluid sample* and then proceed to its 22 children, including *Exudate sample* and *Discharge specimen* (Figure 4.1). When a concept with children is encountered, the children are processed immediately after the parent to support the auditor in detecting error propagation from parent to child. For example, *Amniotic fluid specimen* is followed by its child *Cytologic fluid specimen obtained from amniotic fluid*. An example of a propagation of an error that is easily detectable when reviewing a d-partial-area can be seen with the concept *Synovial fluid specimen* in the d-partial-area *Body fluid sample* (Figure 4.1). A missing topography relationship is detected with the target *Articular space* in the Body Structure hierarchy. The same missing relationship is detected for its three children: *Multiple joint synovial fluid*, *Cytologic material obtained from synovial fluid*, and *Synovial fluid joint NOS*. Arriving later at the d-partial-area *Acellular blood (serum or plasma) specimen*, the root would be examined first. Note that the root's overlapping parent *Body fluid sample* would already have been examined according to the Taxonomy level ordering. The review of its child *Serum specimen* and its four children would follow. Only after that would the review of the sibling *Plasma Specimen* and its three descendants occur (see Figure 4.1).

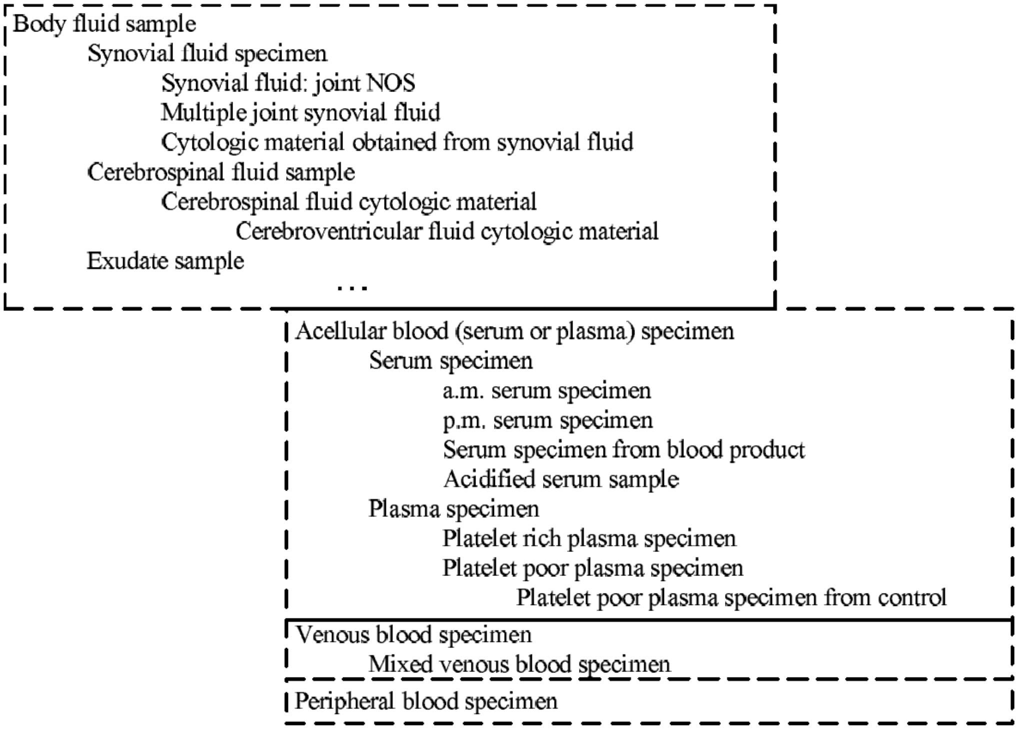

Body fluid sample
Synovial fluid specimen
Synovial fluid: joint NOS
Multiple joint synovial fluid
Cytologic material obtained from synovial fluid
Cerebrospinal fluid sample
Cerebrospinal fluid cytologic material
Cerebroventricular fluid cytologic material
Exudate sample
...
Acellular blood (serum or plasma) specimen
Serum specimen
a.m. serum specimen
p.m. serum specimen
Serum specimen from blood product
Acidified serum sample
Plasma specimen
Platelet rich plasma specimen
Platelet poor plasma specimen
Platelet poor plasma specimen from control
Venous blood specimen
Mixed venous blood specimen
Peripheral blood specimen

Figure 4.3 An indented display of four d-partial-areas and their constituent concepts illustrating the topological-sort-order processing.

For further illustrative purposes, Figure 4.3 shows an excerpt of four d-partial-areas, *Body fluid sample*, *Acellular blood (serum or plasma) specimen*, *Venous blood specimen*, and *Peripheral blood specimen*, of the area {*substance*}, where both the d-partial-areas, drawn as boxes, and the concepts, listed inside the boxes, are displayed in an indented format to illustrate the topological-sort-order processing. The auditing proceeds left-to-right and downward, following the indentation. Only a sample of the concepts are shown for the d-partial-area *Body fluid sample*.

For this phase, the auditing is performed by three domain experts, Dr. Gai Elhanan, Dr. Junchuan Xu, and Dr. Yan Chen, an associate professor from Borough of Manhattan Community College, each of whom has training in medicine as well as training and experience in medical terminologies. All the overlapping concepts of SNOMED's Specimen hierarchy (July 2009), within all its areas, are audited. The data presented to them for each concept are exactly the same in this phase as they are in Phase 1. Additionally, the same error-reporting form is used. In Section 4.3, a sample of the various types of errors is listed.

In the Phase 2 review, the author seeks to achieve a better agreement regarding the combined reported results. Thus, the auditors' findings are anonymized and summarized. The three experts are then requested to review the summary report and mark whether they agree or disagree with the errors listed. One expert might overlook an error discovered by another, and may eventually agree with it once the potential error is reported. All errors asserted by at least one auditor are reviewed by Dr. James T. Case of the SNOMED US National Release Center (NRC) at the NLM for possible inclusion in the US extension of SNOMED. Only errors confirmed by him are considered in the results. Any changes approved by him for inclusion in the US extension of SNOMED are eventually transferred to the IHTSDO for review and potential inclusion in SNOMED's international release.

4.2.3 Hypotheses and Control Sample

There are two hypotheses that were investigated in regard to this study. The first distinguishes between overlapping concepts and non-overlapping concepts. The second distinguishes between overlapping roots of d-partial-areas and other overlapping concepts.

Hypothesis 4.1: Concepts residing in d-partial-areas having overlapping roots (i.e., overlapping concepts) are more likely to have errors than concepts residing in d-partial-areas containing no overlapping concepts. ■

Hypothesis 4.2: Overlapping roots of d-partial-areas are more likely to have errors than non-root overlapping concepts. ■

The first hypothesis asserts that these more complex concepts indeed exhibit a higher number of errors. The second hypothesis refers to the more significant overlapping concepts as the overlapping roots, where the convergence of multiple inheritance paths occurs and where higher concentrations of errors is expected.

As a basis for comparison, a control sample, which comprises concepts gleaned from partial-areas having no overlaps whatsoever, is also audited. Both kinds of concepts are audited by the same auditors. Figure 4.4 presents a flow diagram that summarizes this study.

To compare overlapping concepts with those in the control sample, the proportion of erroneous concepts is examined. The d-partial-area is used as the unit of analysis, and across levels (because of the small number of concepts at Level 2). Both hypotheses are tested for Phases 1 and 2 of the auditing on the two releases of SNOMED, two years apart. The double bootstrap [59] and Fisher’s exact test two-tailed [72] are employed to calculate the statistical significance of the difference of the proportions, for Hypothesis 4.1 and 4.2, respectively.

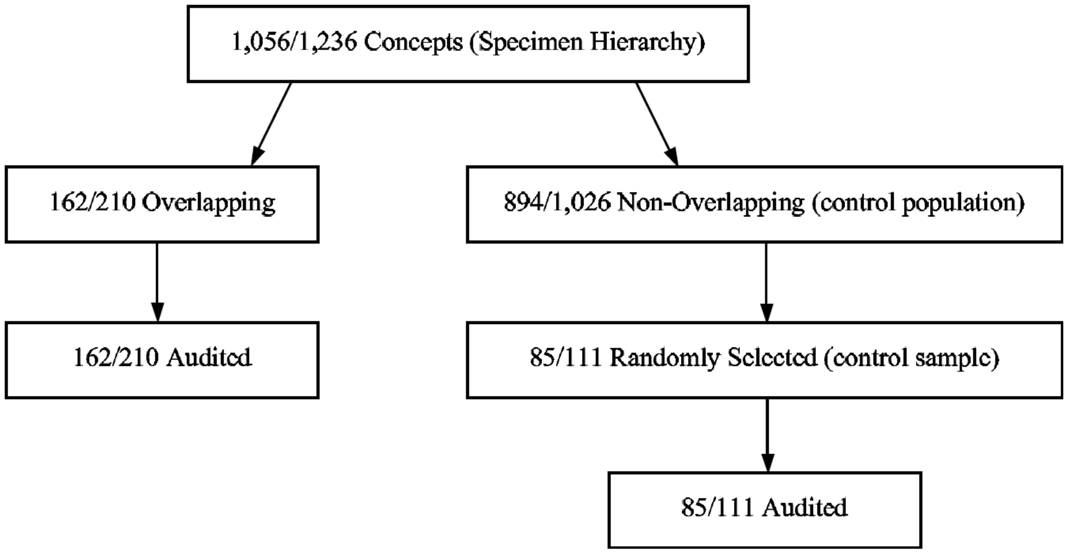

1,056/1,236 Concepts (Specimen Hierarchy)
162/210 Overlapping
894/1,026 Non-Overlapping (control population)
162/210 Audited
85/111 Randomly Selected (control sample)
85/111 Audited

Figure 4.4 Flow diagram summarizing the audits of SNOMED 2007 and 2009. The numbers in each box represent the respective numbers from the 2007 and 2009 versions of SNOMED. For example, “1,056/1,236” in the top box indicates that there are 1,056 concepts in the Specimen hierarchy in SNOMED 2007 and 1,236 in 2009.

4.3 Results

The results are reported for Phase 1 in Section 4.3.1 and for Phase 2 in Section 4.3.2. The results pertaining to the hypotheses (see Section 4.2.3) are distributed in these sections according to the respective phase.

Two phases of results obtained with respect to two releases of SNOMED are reported. Phase 1 for the July 2007 release and Phase 2 for July 2009. In Phase 2, the methodology described in the previous section is utilized and based on the disjoint partial-area taxonomy. During Phase 1, the methodology was not yet developed and therefore and exhaustive audit of all overlapping concepts was carried out without regard to any structural configuration or ordering. A preliminary report with some results of Phase 1 appeared in [70].

4.3.1 Phase 1: Auditing of July 2007 SNOMED

The July 2007 release of the Specimen hierarchy consists of 1,056 concepts, of which 162 are overlapping. For its partial-area taxonomy, see Figure 3.2. Most of the overlapping concepts reside in Level 1 areas, i.e., those having one relationship. In fact, roughly one third (155 out of 468) of the Level 1 concepts are overlapping. And these are found primarily in the area {*topography*} and {*substance*}. A portion of the disjoint partial-area taxonomy of {*substance*} can be seen in Figure 4.5, which should be compared with the 2009 version appearing in Figure 4.2. The d-partial-areas of {*substance*} and {*topography*} can be seen in Figure 3.10 and 3.11, respectively. Overlapping concepts also appear in the partial-areas of areas with two relationships but in far fewer numbers. In fact, there are only seven of them. Six are in {*topography*, *procedure*}, and the other is in {*topography*, *morphology*}.

Table 4.1 presents the results of auditing the 35 overlapping concepts (see Figure 3.8) distributed across nine d-partial-areas in the area {*substance*} (Figure 4.5). For each d-partial-area, the following are listed: number of overlapping concepts V, number of erroneous overlapping concepts V_{err}, the number of errors E_{root} exhibited by the overlapping root, and the total number of errors E for all overlapping concepts. For example, the largest d-partial-area *Blood specimen* has 13 concepts, of which five were found to be in error. The root *Blood specimen* had two errors, and overall the d-partial-area's concepts had seven. For this d-partial-area, 50% (six out of 12) of the non-root overlapping concepts are erroneous, while the root itself exhibits two errors. The result, for one example of a d-partial-area, gives support to Hypothesis 4.2.

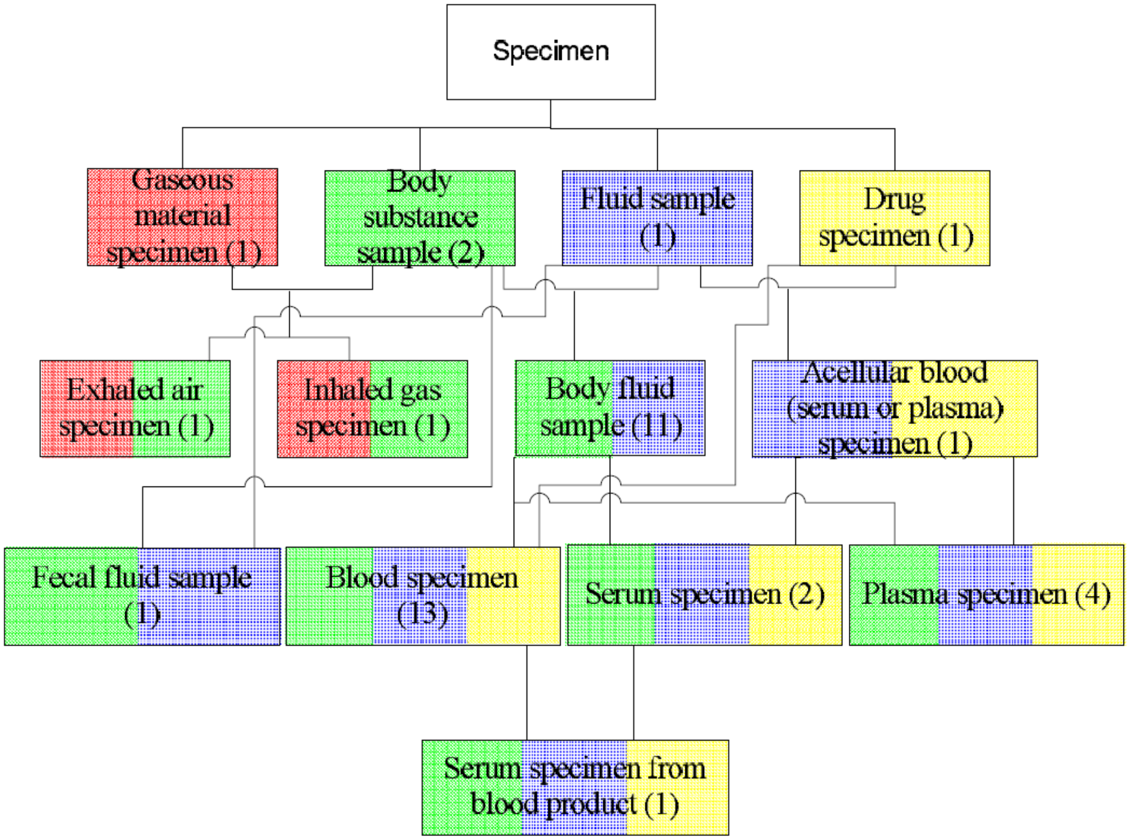


Figure 4.5 A portion of the disjoint partial-area taxonomy for the area {*substance*} (July 2007). The multicolored boxes are the d-partial-areas containing overlapping concepts.

The auditing results for all overlapping concepts are listed by area Table 4.2. For each area, its total number of concepts *C*, number of overlapping concepts *V*, number of overlapping roots *D*, number of erroneous overlapping concepts V_{err}, total number of errors *E* for the overlapping concepts, number of erroneous overlapping roots D_{err}, number of errors E_{root} exhibited by the set of overlapping roots, and a number of relevant ratios are shown. For example, {*substance*} has 81 concepts, of which 35 are overlapping. Eleven (31%) of the latter were found to have a total of 31 errors or an average of 2.8 per erroneous concept, as detailed in Table 4.2. The ratio of the total number of errors at the overlapping concepts to the number of overlapping concepts is 0.89. Of the nine overlapping roots, five (56%) were found to be in error – with a combined 24 errors among them (or 4.8 errors per erroneous root). But only 23% (= (11-5)/(35-9)) of the non-root overlapping concepts had errors. Note that for some areas (e.g., {*procedure*}), the ratio in the last column is not applicable (undefined) since singletons (i.e., d-partial-areas containing just one concept) have no non-root overlapping concepts. Other ratios may not be applicable due to a lack of errors. Nevertheless, the total ratios at the bottom of the table are defined across all the areas with overlapping concepts.

Table 4.1 Auditing Results for Overlapping Concepts of {*substance*} Arranged by Disjoint Partial-area

Disjoint partial-area	V	V_{err}	E_{root}	E
Exhaled air specimen	1	0	0	0
Inhaled gas specimen	1	0	0	0
Fecal fluid sample	1	0	0	0
Acellular blood (serum or plasma) specimen	1	1	1	1
Serum specimen from blood product	1	1	3	3
Serum specimen	2	0	0	0
Plasma specimen	4	1	1	1
Body fluid sample	11	3	17	19
Blood specimen	13	5	2	7
Total:	**35**	**11**	**24**	**31**

V = # overlapping concepts; V_{err} = # erroneous overlapping concepts;
E_{root} = # errors at the overlapping root; E = total # errors at overlapping concepts

Table 4.2 Auditing Results for Overlapping Concepts by Area

Area	C	V	D	V_{err}	E	E/V_{err}	E/V	D_{err}	E_{root}	E_{root}/D_{err}	D_{err} /D (%)	$(V_{err}$-$D_{err})$ /$(V$-$D)$ (%)
substance	81	35	9	11	31	2.8	0.89	5	24	4.80	56	23
topography	333	116	52	71	110	1.6	0.95	39	62	1.59	75	50
procedure	20	3	3	3	9	3.0	3.00	2	9	4.50	66	N/A
identity	20	1	1	0	0	N/A	0	0	0	N/A	0	N/A
topog., proc.	380	6	6	4	9	2.3	1.50	4	9	2.30	66	N/A
topog., morph.	18	1	1	0	0	N/A	0	0	0	N/A	0	N/A
Total:	**852**	**162**	**72**	**89**	**159**	**1.8**	**0.93**	**50**	**104**	**2.1**	**69**	**43**

C = # concepts; V = # overlapping concepts; D = # overlapping roots;
V_{err} = # erroneous overlapping concepts; E = total # errors at overlapping concepts;
D_{err} = # erroneous overlapping roots; E_{root} = # errors at the overlapping roots; N/A = Not applicable

Most overlapping concepts in {*topography*} are found in intersections of partial-areas involving *Tissue specimen* containing 126 concepts. These results have been tabulated separately in Table 4.3. For example, the partial-area *Specimen from eye* has 18 concepts. Its intersection with *Tissue specimen* has 12 of them. Eight of those are in error.

The control sample was gleaned from partial-areas from partial-areas that had no intersections whatsoever with other partial-areas and from d-partial-areas having no overlapping concepts (i.e., those left over after the removal of the d-partial-areas with overlapping concepts from a partial-area; see, e.g., the six d-partial-areas at Level 1 of Figure 4.2). Furthermore, only partial-areas that contained more than one concept are used. The reason for the last requirement is that, as alluded to, partial-areas of one concept are already known to be error-prone [46, 49]. Thus, they do not make for a proper control sample.

Table 4.3 Results of Auditing Intersections Involving Partial-area *Tissue specimen*

Second Partial-Area	***C***	***V***	***V_{err}***	***V_{err} / V (%)***
Specimen from eye	18	12	8	67
Ear sample	2	1	0	0
Specimen from breast	8	4	2	50
Cardiovascular sample	13	3	1	33
Products of conception tissue sample	12	1	1	100
Genitourinary sample	73	20	17	85
Dermatological sample	6	2	0	0
Specimen from digestive system	74	29	18	62
Musculoskeletal sample	35	22	15	68
Respiratory sample	41	6	5	83
Endocrine sample	12	3	0	0
Specimen from central nervous system	4	1	0	0
Spec. from thymus gland	2	1	0	0
Specimen from trophoblast	2	1	0	0

A control sample of 78 concepts is used from Level 1, half of its overlapping concepts (155). From Level 2, seven concepts are gathered for the control sample, an equal number to the overlapping concepts. Hence, there are 155+7=162 overlapping concepts, and the control sample has 78+7 = 85 concepts. Since the purpose was to audit overlapping concepts, a smaller control sample is used that was large enough to support statistical significance for the result presented below.

Table 4.4 gives the results of the auditing carried out on these two groups of concepts. *C* denotes the number of concepts, *E* (Column 3) denotes the total number of errors, and C_{err} is the number of erroneous concepts (Column 5)—with a given concept potentially having more than one error. The average erroneous-concept rate among the overlapping concepts was 55%, and among the control sample it was 29% (Column 6). The difference was significant (using the double bootstrap [72]) at the 0.05 level, supporting Hypothesis 4.1. Let the author point out that there was nearly one error (0.98) on average per overlapping concept as compared to 0.36 on average within the control sample (Column 4). Moreover, erroneous concepts in the overlapping group had 1.8 errors on average (last column) versus 1.2 errors on average for the control sample, showing further difference between the two.

Table 4.4 Auditing Results for Overlapping Concepts vs. Control Sample (Phase 1)

	C	***E***	***E / C***	***C_{err}***	***C_{err}/C(%)***	***E/C_{err}***
Overlapping	162	158	0.98	89	55	1.8
Control Sample	85	31	0.36	25	29	1.2

In examining the auditing results, overlapping roots are found to be more error-prone than other overlapping concepts. For example, in {*procedure*} and {*topography*, *procedure*}, all errors are found in overlapping roots. As shown in Table 4.2, in the area {*substance*}, five out of nine roots (55%) versus six (= 11-5) out of 26 (=35-9) non-root overlapping concepts (23%) were found to be erroneous. To assess Hypothesis 4.2, the data from Table 4.2 are used for the entire collection of overlapping concepts. The percentage of erroneous concepts for overlapping roots is 69% (=50/72). The percentage of erroneous concepts in the set of non-root overlapping concepts is 43% (=(89-50)/(162-72)). The difference in the percentages of erroneous concepts between the overlapping roots (69%) and the non-root overlapping concepts (43%) is statistically significant (Fisher's exact test two-tailed [72], p-value = 0.0014), supporting Hypothesis 4.2.

4.3.2 Phase 2: Auditing of July 2009 SNOMED

The results of Phase 1 were submitted to CAP for consideration and incorporation into the Specimen hierarchy. As a result, there were many changes in the overlapping concepts of this hierarchy as reflected in SNOMED's July 2009 release. The area taxonomy and the partial-area taxonomy for the July 2009 release appear in Figures 4.6 and 4.7, respectively. A comparison of the area taxonomies of 2007 (Figure 3.1) and 2009 (Figure 4.6) exposes many differences in the Specimen hierarchy. For example, the total number of concepts with one relationship—which is equal to the sum of the sizes of the (green) areas on Level 1—went down from 468 to 420. At the same time, the area {*substance*} grew from 81 to 107 concepts. The number of areas with three relationships went down from seven to five with the loss of the two areas {*morphology*, *procedure*, *substance*} and {*topography*, *identity*, *procedure*}. On the other hand, the area {*procedure*, *topography*, *substance*} grew from 26 concepts in 2007 to 288 concepts in 2009.

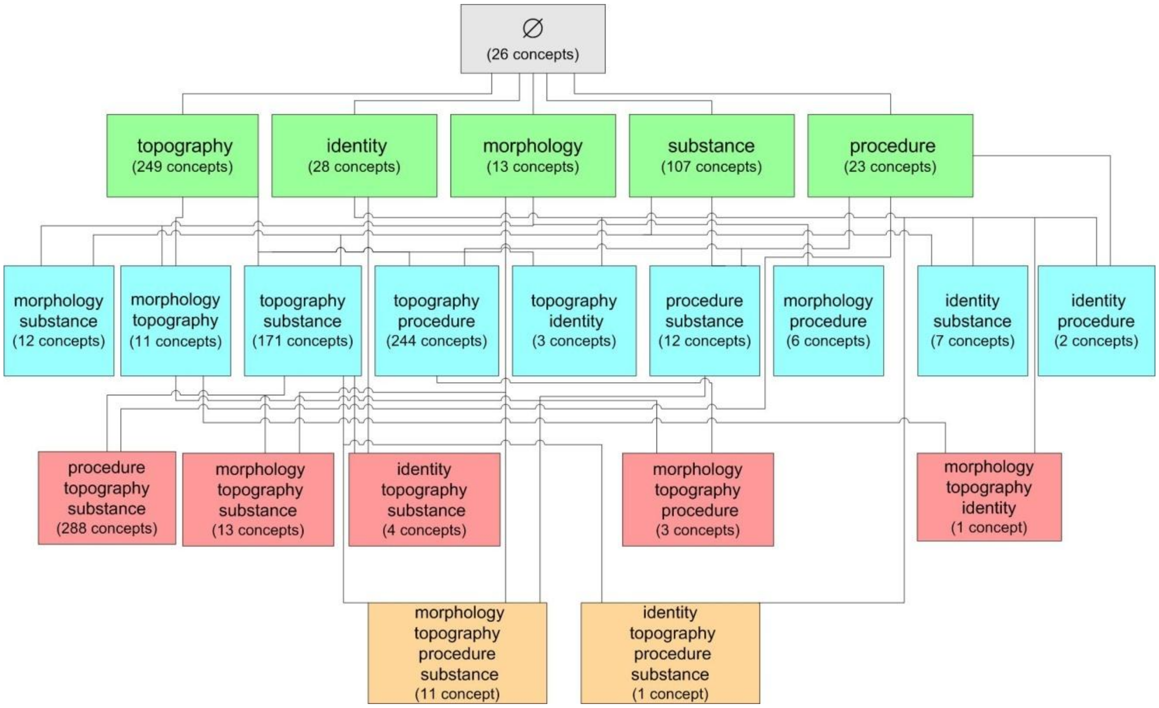


Figure 4.6 Area taxonomy for SNOMED's Specimen hierarchy (July 2009 release).

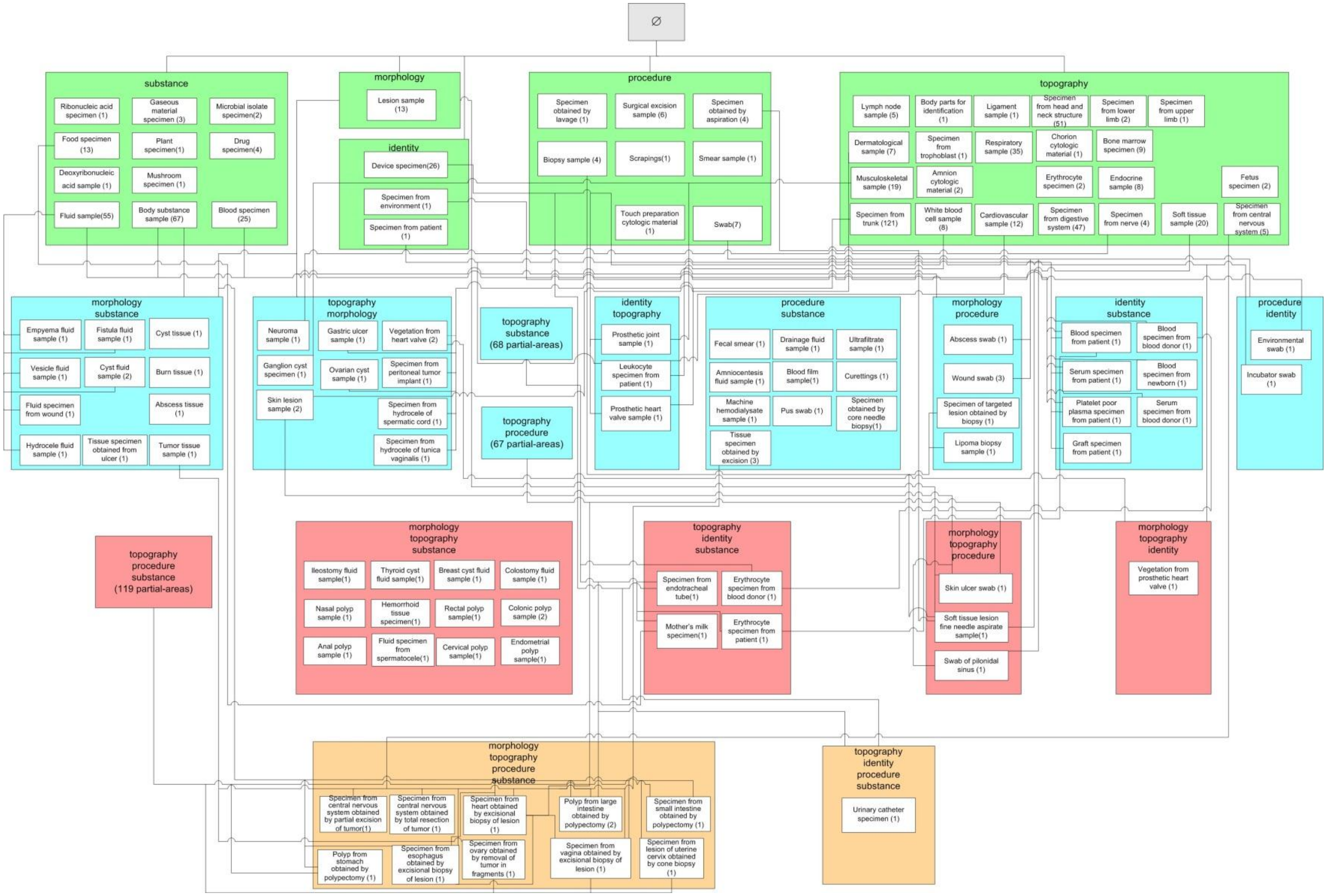

∅
substance
Ribonucleic acid specimen (1)
Gaseous material specimen (3)
Microbial isolate specimen(2)
Food specimen (13)
Plant specimen(1)
Drug specimen(4)
Deoxyribonucleic acid sample (1)
Mushroom specimen (1)
Fluid sample(55)
Body substance sample (67)
Blood specimen (25)
morphology
Lesion sample (13)
identity
Device specimen(26)
Specimen from environment (1)
Specimen from patient (1)
procedure
Specimen obtained by lavage (1)
Surgical excision sample (6)
Specimen obtained by aspiration (4)
Biopsy sample (4)
Scrapings(1)
Smear sample (1)
Touch preparation cytologic material (1)
Swab(7)
topography
Lymph node sample (5)
Body parts for identification (1)
Ligament sample (1)
Specimen from head and neck structure (51)
Specimen from lower limb (2)
Specimen from upper limb (1)
Dermatological sample (7)
Specimen from trophoblast (1)
Respiratory sample (35)
Chorion cytologic material (1)
Bone marrow specimen (9)
Musculoskeletal sample (19)
Amnion cytologic material (2)
Erythrocyte specimen (2)
Endocrine sample (8)
Fetus specimen (2)
Specimen from trunk (121)
White blood cell sample (8)
Cardiovascular sample (12)
Specimen from digestive system (47)
Specimen from nerve (4)
Soft tissue sample (20)
Specimen from central nervous system (5)
morphology substance
Empyema fluid sample (1)
Fistula fluid sample (1)
Cyst tissue (1)
Vesicle fluid sample (1)
Cyst fluid sample (2)
Burn tissue (1)
Fluid specimen from wound (1)
Abscess tissue (1)
Hydrocele fluid sample (1)
Tissue specimen obtained from ulcer (1)
Tumor tissue sample (1)
topography morphology
Neuroma sample (1)
Gastric ulcer sample (1)
Vegetation from heart valve (2)
Ganglion cyst specimen (1)
Ovarian cyst sample (1)
Specimen from peritoneal tumor implant (1)
Skin lesion sample (2)
Specimen from hydrocele of spermatic cord (1)
Specimen from hydrocele of tunica vaginalis (1)
topography substance (68 partial-areas)
topography procedure (67 partial-areas)
identity topography
Prosthetic joint sample (1)
Leukocyte specimen from patient (1)
Prosthetic heart valve sample (1)
procedure substance
Fecal smear (1)
Drainage fluid sample (1)
Ultrafiltrate sample (1)
Amniocentesis fluid sample (1)
Blood film sample(1)
Curettings (1)
Machine hemodialysate sample (1)
Pus swab (1)
Specimen obtained by core needle biopsy(1)
Tissue specimen obtained by excision (3)
morphology procedure
Abscess swab (1)
Wound swab (3)
Specimen of targeted lesion obtained by biopsy (1)
Lipoma biopsy sample (1)
identity substance
Blood specimen from patient (1)
Blood specimen from blood donor (1)
Serum specimen from patient (1)
Blood specimen from newborn (1)
Platelet poor plasma specimen from patient (1)
Serum specimen from blood donor (1)
Graft specimen from patient (1)
procedure identity
Environmental swab (1)
Incubator swab (1)
topography procedure substance (119 partial-areas)
morphology topography substance
Ileostomy fluid sample(1)
Thyroid cyst fluid sample(1)
Breast cyst fluid sample (1)
Colostomy fluid sample (1)
Nasal polyp sample (1)
Hemorrhoid tissue specimen(1)
Rectal polyp sample(1)
Colonic polyp sample (2)
Anal polyp sample (1)
Fluid specimen from spermatocele(1)
Cervical polyp sample(1)
Endometrial polyp sample(1)
topography identity substance
Specimen from endotracheal tube(1)
Erythrocyte specimen from blood donor (1)
Mother's milk specimen(1)
Erythrocyte specimen from patient (1)
morphology topography procedure
Skin ulcer swab (1)
Soft tissue lesion fine needle aspirate sample(1)
Swab of pilonidal sinus (1)
morphology topography identity
Vegetation from prosthetic heart valve (1)
morphology topography procedure substance
Specimen from central nervous system obtained by partial excision of tumor(1)
Specimen from central nervous system obtained by total resection of tumor (1)
Specimen from heart obtained by excisional biopsy of lesion (1)
Polyp from large intestine obtained by polypectomy (2)
Specimen from small intestine obtained by polypectomy (1)
Polyp from stomach obtained by polypectomy (1)
Specimen from esophagus obtained by excisional biopsy of lesion (1)
Specimen from ovary obtained by removal of tumor in fragments (1)
Specimen from vagina obtained by excisional biopsy of lesion (1)
Specimen from lesion of uterine cervix obtained by cone biopsy (1)
topography identity procedure substance
Urinary catheter specimen (1)

Figure 4.7 Partial-area taxonomy for the Specimen hierarchy (July 2009 release).

Similarly, comparing the partial-area taxonomies for 2007 and 2009 reveals many differences. For example, the area {*substance*} changed from having ten to 11 partial-areas. But that small numerical change is misleading, as one can guess, considering the 32% increase in the size of the area. Only six partial-areas did not change. A new partial-area is *Blood specimen* with 25 concepts. Note that there was a d-partial-area with that name consisting of 13 concepts in 2007 (Figure 4.5). At the same time, *Drug specimen* shrank from 23 to four concepts, mainly due to the removal of blood specimen concepts. *Body substance sample* expanded from 47 to 67 concepts, while *Fluid sample* grew from 44 to 55 concepts. Such large changes on the partial-area level seem to indicate an increase in the overlap size when compared to the overall increase of 26 concepts observed on the area level. As another example, the area {*morphology*, *topography*, *substance*} went from having three partial-areas to 12. The area {*morphology*, *topography*, *procedure*, *substance*} grew from one to ten.

The number of overlapping concepts increased by 48 from 162 to 210 (30%). Clearly, the landscape of the overlapping portions of partial-areas changed meaningfully from the time of the July 2007 release. For example, as was predicted above, in the area {*substance*}, there were 35 overlapping concepts in nine d-partial-areas in 2007 (Figure 3.9), but 48 overlapping concepts in 15 d-partial-areas in 2009 (Figure 4.2).

These changes motivated the application of the new methodology based on the disjoint partial-area taxonomy in this phase to the July 2009 release's overlapping concepts. The author's expectation was also that this new methodology employing a detailed order of review would expose errors missed during Phase 1.

Table 4.5 Sample of Error Types of Overlapping Concepts for July 2009 Release

Concept	Partial-areas	Error Type(s)	Correction(s)
Serum specimen from blood product	Blood specimen / Fluid sample/Body substance sample	Missing parent	Add parent: Blood specimen from blood product
Dentin specimen	Specimen from digestive system/Specimen from head and neck structure	Incorrect Parent: Oral cavity sample	Correct parent: Specimen from tooth
a.m. serum specimen	Blood specimen/Fluid sample(specimen)/Body substance sample	Missing relationship	Add relationship: TIMEASPECT with the value of – am-ante meridiem
Specimen from tooth	Specimen from digestive system/Specimen from head and neck structure	Incorrect relationship target: Oral cavity structure	Refine with: Tooth structure
Specimen obtained by fine needle aspiration procedure	Specimen obtained by aspiration/Biopsy sample	Missing child	Add children: *Breast fine needle aspirate sample; *Soft tissue lesion fine needle aspirate sample; *Specimen from heart obtained by fine needle aspiration procedure; *Specimen from thymus gland obtained by fine needle aspiration biopsy
Tissue specimen from placenta	Tissue specimen from genital system/Products of conception tissue sample	Other error type: missing ancestor “Soft tissue sample”	Create a proper concept to parent it in the “Soft tissue sample” tree.

A sample of different types of errors agreed upon by all three auditors and confirmed after a review (by Dr. James T. Case) is listed in Table 4.5. For example, it was agreed that *Serum specimen from blood product* is missing a parent *Blood specimen* *from blood product* that should be added. Table 4.6 summarizes the number of occurrences for each type of error found in the overlapping concepts of the July 2009 release. Missing parents, for example, were found for 23 concepts.

In the Phase 2 review, a better agreement regarding the combined reported results is tried to be achieved. One expert might have overlooked an error discovered by another, and may have agreed with it, once the potential error was reported. The level of agreement improved after the second-stage review. All overlapping concepts are reported as potential errors to the SNOMED United States NRC having at least one auditor reporting an error for them. The report was reviewed by Dr. Case (who works at the NRC). Only errors confirmed by him are considered in the results presented in the following.

Table 4.6 Distribution of Types of Errors in the Second Phase of Auditing Overlapping Concepts

Error Type	**# Concepts**
Missing parent	23
Incorrect parent	22
Missing child	6
Incorrect child	2
Missing relationship	55
Incorrect relationship target	2
Other error type	6

The auditing results for Phase 2 are listed by area in Table 4.7, in the same format used in Table 4.2 for Phase 1. In this case, for example, {*topography*} has 249 concepts, with 110 of them being overlapping. Fifty-two out of the 110 (47%) were found to have a total of 57 errors or an average of 1.10 per erroneous concept. The ratio of the total number of errors to the number of overlapping concepts is 0.52. Twenty of the 37 overlapping roots (54%) were found to be in error – with a combined 22 errors among them (or 1.10 errors per root). Finally, 44% (=(52-20)/(110-37)) of the non-root overlapping concepts had errors.

For the entire set of overlapping concepts summarized in the bottom row of Table 4.7, 127 out of 210 (60%) were found to be erroneous. This result is applicable in assessing Hypothesis 4.1 (as shown in Table 4.8).

Table 4.7 Phase 2 Auditing Results for Overlapping Concepts by Area

Area	C	V	D	V_{err}	E	E/V_{err}	E/V	D_{err}	E_{root}	E_{root}/D_{err}	D_{err}/D (%)	$(V_{err}-D_{err})/(V-D)$ (%)
substance	107	48	15	28	36	1.29	0.75	8	11	1.38	53	61
topography	249	110	37	52	57	1.10	0.52	20	22	1.10	54	44
procedure	23	2	1	1	1	1.00	0.50	1	1	1.00	100	0
topog., proc.	244	29	16	28	38	1.36	1.31	15	19	1.27	94	100
topog., subst.	171	5	4	3	4	1.33	0.80	3	4	1.33	75	0
subst., topog., proc.	288	16	14	15	25	1.67	1.56	14	23	1.64	100	50
Total:	**1,082**	**210**	**87**	**127**	**161**	**1.27**	**0.77**	**61**	**80**	**1.30**	**70**	**54**

C = #concepts; V=#overlapping concepts; D=#overlapping roots;
V_{err} = #erroneous overlapping concepts; E=total #errors;
D_{err} = # erroneous overlapping roots; E_{root} = #errors at the roots

The control sample for Phase 2 was taken strictly from partial-areas and d-partial-areas that had no intersections whatsoever. As with Phase 1, only partial-areas that contained more than one concept are used. The sample consisted of 111 concepts from the same areas as the overlapping concepts. And as in Phase 1, the number of sample concepts taken from areas with small numbers (i.e., 2 – 16) of overlapping concepts was about the same as the number of overlapping concepts taken from those areas. The sample concepts numbered about half the overlapping concepts for areas with larger numbers of overlapping concepts. As with Phase 1, the purpose was to audit overlapping concepts, and a smaller control sample is used that was nevertheless big enough to support statistical significance of the result.

Like Table 4.4, Table 4.8 juxtaposes the results of auditing the overlapping concepts and those in the control sample. The average erroneous-concept rate among the overlapping concepts was 60%, versus 13% for the control sample (Column 6). The difference was significant at the 0.05 level, supporting Hypothesis 4.1. Note that there were 0.77 errors on average per overlapping concept as compared to 0.13 on average within the control sample (Column 4). Erroneous concepts in the overlapping group had 1.27 errors on average (last column) versus 1.00 errors on average for the control sample, showing further difference between the two samples.

Table 4.8 Auditing Results for Overlapping Concepts vs. Control Sample (Phase 2)

	C	***E***	***E/C***	***C_{err}***	***C_{err}/C (%)***	***E/C_{err}***
Overlapping	210	161	0.77	127	60	1.27
Control Sample	111	14	0.13	14	13	1.00

For the assessment of Hypothesis 4.2, the results obtained for all overlapping concepts are used, reflected in the bottom row of Table 4.7. Among the 87 overlapping roots, 61 (70%) were erroneous, while for the 123 (=210-87) non-root overlapping concepts, 66 (=127-61 or 54%) were found to be in error. The difference in the percentages of erroneous concepts between the overlapping roots (70%) and the non-root overlapping concepts (54%) is statistically significant (Fisher’s exact test two-tailed, p-value = 0.0217).

4.4 Discussion

4.4.1 Auditing Theme: Complex Concepts

This study is motivated by a general theme that more “complex” concepts tend to have more errors than simpler concepts. The theme of being more complex may manifest itself in a variety of ways. One manifestation of this theme for partial-areas was the group of concepts residing in “strict inheritance” partial-areas described in Chapter 2 [48, 49]. In the context of the current study, this theme appears twice: the first time in identifying overlapping concepts as more complex than non-overlapping concepts due to their elaborating the multiple semantics of the multiple partial-areas they belong to; the second in the distinction between overlapping roots and non-root overlapping concepts. The reason for the higher complexity of overlapping roots stems from their being at the junction points where multiple hierarchical paths from ancestors converge. Each such path contributes a portion of a diverse collection of inherited knowledge at the overlapping root. Hypothesis 4.1 addresses the first appearance. Hypothesis 4.2 pertains to the second.

As was shown in Chapter 2 with regards to strict inheritance partial-areas, the results of the study confirm the auditing theme that complex concepts have relatively more errors. In view of the fact that modeling complex concepts is more challenging than modeling simpler concepts, it is not really surprising to find more errors in the former. The research challenge is to discover various characterizations of “complex” concepts. In particular, it is fruitful to identify structural characterizations that can be computed automatically, as in the current study and in Chapter 2. The higher error rate shown here and in Chapter 2 will help achieve higher productivity from quality-assurance personnel in their review of such concepts. It is suggested that the design of partial-area taxonomies and the auditing of the complex concepts discussed here and in Chapter 2 should become integral parts of the design cycle for terminologies such as SNOMED and the NCIt [46]. Such techniques will also help interface terminologies such as Kaiser-Permanente’s CMT [62] or the VA’s ERT [61], which were derived initially from SNOMED and were enhanced with local vocabulary as well as integrated parts of other terminologies. It is a research challenge to identify more manifestation of complex concepts using taxonomies or other structural techniques for SNOMED and similar terminologies.

One may wonder why there are more errors in overlapping roots than there are in other overlapping concepts (as stated in Hypothesis 4.2), in spite of the expectation that this methodology will expose error propagation from parents to children, which implies that errors at an overlapping root would be "inherited" by the other concepts in its d-partial-area. One should realize that indeed missing or incorrect relationship errors are "inherited," but that is not true of other errors, e.g., an incorrect parent. Furthermore, many d-partial-areas have just a single concept (which serves as the respective root), with no children below to inherit the errors. Hence, this methodology is designed to expose the cross-generational error propagation to the extent that it exists.

4.4.2 Repeated Application of an Auditing Methodology

In this dissertation, various methodologies for auditing a SNOMED hierarchy are presented. A question to consider is whether there is a reason to reapply the same auditing technique to the hierarchy obtained following corrections derived from the earlier auditing phase that used the same technique. Should it be assumed that not all errors were found and corrected? In the context of this research, the question was: should the overlapping concepts be audited again following the first phase reported in [70]? Furthermore, how many times should the same technique be applied? Another way to phrase this last question is: how the convergence of the auditing process is identified?

There were several reasons to re-audit the overlapping concepts. First, in Phase 1, only the set of all overlapping concepts were audited without utilizing any structure among them. In this chapter, the new “group auditing” methodology of overlapping concepts was introduced, where d-partial-areas were utilized as the grouping unit following the new framework described in Chapter 3 [50]. Furthermore, the new methodology employs a top-down ordering within each d-partial-area and among various d-partial-areas.

Another reason for repeating the auditing on the overlapping concepts is the large increase in their numbers and the number of d-partial-areas. For example, see Figure 4.2 for the d-partial-areas in the area {*substance*} in comparison to the corresponding Figure 3.9. Only four d-partial-areas without overlapping concepts are seen in Figure 3.9 at the first level and nine d-partial-areas comprising overlapping concepts. In Figure 4.2, showing the overlapping concepts of {*substance*} in 2009, there are six top d-partial-areas without overlapping concepts and 15 d-partial-areas with overlapping concepts. Moreover, when one reviews the details of the two figures, many internal changes can be seen. For example, the d-partial-area *Body fluid sample* had 11 concepts in 2007 and 23 in 2009. *Blood specimen* had 13 overlapping concepts in Level 3 originally, and in 2009 it is a top d-partial-area of one concept only. It has eight child d-partial-areas containing 18 overlapping concepts on Level 3, which are shared jointly by the parent d-partial-area *Body fluid sample* (see Figure 4.2). The latter was a parent of *Blood specimen* in Figure 3.9.

When realizing the extent of the changes, it was possible that new errors were introduced and that the new disjoint partial-area taxonomy would lead to exposure of errors not reported in the review of the 2007 release. The results shown in Table 4.7 justify the decision for the second auditing phase. While a meaningful amount of errors are expected to be found in Phase 2, it is surprising by their magnitude. Both the percentages of erroneous concepts among overlapping concepts (60% vs. 55%) and among overlapping roots (70% vs. 69%) were little changed in spite of this being a second round of auditing. Part of the explanation may be the improved methodology employed in this study. Another reason may be the large increase in the number of overlapping concepts (from 162 to 210). A further factor might be that in practice the proper modeling of these complex concepts demands more than one iteration.

On the other hand, the ratio of errors per erroneous concept was reduced (0.93 to 0.77) for all overlapping concepts, as was the ratio for erroneous overlapping roots (2.1 to 1.3). Hence, while the percentage of erroneous concepts persisted, the average number of errors fell. That is, fewer concepts with multiple errors are found. This last observation seems in line with the speculation above that multiple iterations are required for the proper modeling of complex concepts.

One could certainly question the expectation of the need for an additional phase of auditing after all corrections from the overlapping-concept regimen have been implemented. That is particularly true when the corrections have made their way into SNOMED’s international release following the report of Dr. Case at the NRC to IHTSDO. To better understand the phenomenon of finding more errors in a subsequent phase of auditing overlapping concepts mentioned above, one needs to keep in mind the restructuring undergone by d-partial-areas due to the discovered errors. For example, in the description of the methodology in Section 4.2, a concept *Synovial fluid specimen* in the d-partial-area *Body fluid sample* is mentioned, which together with its children is missing the relationship *specimen topography* to *Articular space*. But reviewing the complete audit report for the overlapping concepts in {*substance*}, one may realize that the same concept was found to have an incorrect parent, *Body fluid sample*, which was replaced by *Joint fluid specimen*. This latter concept was independently found to be missing the same *topography* relationship, as was its child *Cytologic material obtained from joint fluid*. Furthermore, another concept *Synovial fluid cells* in the area {*topography*} was also made a child of *Synovial fluid specimen* instead of *Synovial sample*. What is seen is a movement of many concepts into the d-partial-area rooted at *Joint fluid specimen*, which before had only one child. Moreover, this d-partial-area would move from the area {*substance*} to the area {*substance*, *topography*} due to the additional *topography* relationship. When all these corrections are incorporated into a future release of SNOMED, the disjoint partial-area taxonomy will convey the refined modeling of all joint fluid specimen concepts, contributing to better overall comprehension. However, this new modeling may expose errors not yet detected and deserves the analysis provided by the disjoint partial-area taxonomy.

If the new disjoint partial-area taxonomy for the Specimen hierarchy obtained as a result of the Phase 2 audit, and possibly reflecting a future release of SNOMED, were to differ meaningfully from the disjoint partial-area taxonomy of the 2009 release of SNOMED, then it may be advisable to reapply the auditing regimen utilizing this new view.

4.4.3 Error Rates and the Complexity of the Disjoint Partial-area Taxonomy

In Phase 1 of the auditing, the bulk of the erroneous overlapping concepts and the overlapping concept errors occurs for the areas {*substance*} and {*topography*}. It is interesting to compare the various ratios of errors for these two areas. The percentage of erroneous overlapping concepts in {*topography*} (61%) is about double that in {*substance*} (31%). However, when measuring the ratios of errors to overlapping concepts, the values for the two areas, 0.95 and 0.89, respectively, are close. This is a result of a much higher ratio of errors to erroneous concepts for {*substance*} (2.8) than for {*topography*} (1.6). This observation indicates a correlation between the ratio of the number of errors to the number of erroneous concepts and the level of complexity of overlapping concepts, as expressed in the structure of the disjoint partial-area taxonomy. As was discussed and shown in Figures 3.9 and 3.10 in Chapter 3, the nature of the overlap is much more complex for {*substance*} with several levels in its disjoint partial-area taxonomy, while it is simpler and relatively flat for {*topography*}.

4.4.4 An Audit Report from Several Auditors

The auditing in Phase 1 was performed by two auditors (Dr. Elhanan and Dr. Xu), and their error report was obtained by a consensus from their individual findings. Anecdotal evidence from the auditors was that the face-to-face consensus process seemed to follow more of a social give-and-take rather than a deep investigation about the concepts. Similar anecdotal evidence was obtained for a study of auditor performance regarding a consensus-building stage [73].

As a result, it was decided to avoid the discussion-based, consensus-building effort in the Phase 2 auditing. Instead, a combined report derived from the three auditors' Phase 2 reports was circulated. This report was anonymized and contained listings of the number of auditors for each identified error. In this second stage, each auditor was asked to indicate their agreement with each of the errors. Errors that had the support of at least one auditor were passed on for further review. It seems that a second review of others' audit reports carried out by each auditor individually without the pressure of direct social interaction is functioning well in achieving an agreement level. Not only was a better level of agreement reached, but the author also witnessed auditors backing off from certain errors, when noticing that the other auditors did not mark them.

4.4.5 Limitations and Future Work

As can be seen from Tables 4.4 and 4.8, according to all reported measures, there is a significantly higher return for the auditing effort obtained for the overlapping concepts compared to concepts in partial-areas without overlaps. Such higher return seems to justify concentrating auditing efforts on the more complex overlapping concepts. The results confirm Hypothesis 4.1. More experiments with different and larger hierarchies of SNOMED and similar terminologies, e.g., NCIt [46], are needed to further confirm the finding. One idea expressed in Chapter 3 that was not confirmed by this study was that “derived” overlapping roots (of d-partial-areas) would be more error-prone than “base” overlapping roots due to their higher complexity. The current results did not support such a phenomenon. Future studies should look again at whether this extra inherent complexity manifests itself in higher error rates in other SNOMED hierarchies.

The interest of the author in this dissertation was not in studying the auditing process per se, but in the distribution of the unquestionable errors resulting from it. Auditor performance and the impact of various protocols in achieving better agreement among a group of auditors may be investigated further in the future.

4.5 Summary

The author proceeded from the assumption that “complex” concepts warrant particular attention in quality-assurance activities pertaining to SNOMED. Toward that end, an auditing methodology based on a refined abstraction network for a SNOMED hierarchy is presented, called the *disjoint partial-area taxonomy*, formulated in Chapter 3. The complex concepts in this study were taken to be those residing in elements of the disjoint partial-area taxonomy that represented certain overlapping subsets of portions of a SNOMED hierarchy. These so-called overlapping concepts in the Specimen hierarchy (in two different releases of SNOMED) were identified programmatically and then put through rigorous audits. Comparing these auditing results with those from control sets, a statistically significant higher error rate among the overlapping concepts is found. Furthermore, among the overlapping concepts, roots have a statistically significantly higher error rate than do non-roots. Thus, the auditing methodology based on the disjoint partial-area taxonomy and its overlapping concepts can be seen as an important addition to the existing suite of SNOMED and SNOMED-related terminology auditing regimens.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Biomedical terminologies, such as SNOMED, have attained an important position in the medical information domain, underlying applications ranging from electronic medical records and clinical laboratory systems to outcomes assessment and telemedicine. As such, it is critical that the conceptual content of terminologies be kept as accurate and up-to-date as possible. Due to SNOMED's large volume and continuing expansion, quality assurance is a daunting challenge facing the biomedical community.

This dissertation takes an approach based entirely on the structural aspects of the SNOMED hierarchies, aiming at developing automated or semi-automated methods that can identify concepts deserving special attention, and consequently enhance the efficacy and efficiency of the auditing process.

A partitioning methodology is applied to a SNOMED hierarchy which yields small groups of concepts similar in both structure and semantics. Three different abstraction networks, the area taxonomy, partial-area taxonomy and disjoint partial-area taxonomy, are derived programmatically from the partitions. These three taxonomies complement each other in terms of granularity of display, providing a high-level contextual view of the underlying terminology in a multi-scale display.

The taxonomies form the basis for a number of systematic auditing regimens proposed and implemented in this dissertation. Often times, concept errors are manifested as anomalies at the taxonomy level. For example, by examining the area taxonomy and partial-area taxonomy, three kinds of concept groups, strict-inheritance regions, mixed regions, and small partial-areas, are found to be fruitful in bringing errors to light [48,49]. The disjoint partial-area taxonomy is devised as a refinement of partial-area taxonomy, which helps to reveal the complexity of the configuration of the overlapping concepts [50]. Following the assumption that “complex” concepts warrant particular attention in quality assurance activities, the overlapping concepts in SNOMED Specimen hierarchy are identified programmatically and then put through a rigorous audit. Comparing these auditing results with results from a control set, a statistically significant of higher error rate among the overlapping concepts has been found. In addition, two phases of auditing were carried out with respect to two releases of SNOMED in different fashions. Results show a statistically significant higher error rate among the overlapping concepts.

In general, the taxonomy-based auditing methodology presented in this dissertation can be seen as complementary to other auditing approaches. Since different auditing techniques typically expose some kinds of errors while missing others, there is a need for a suite comprising a variety of techniques to provide quality-assurance support for terminologies.

In the future, the current study will be extended in the following directions. The current research and experiments are mostly done using SNOMED’s Specimen hierarchy. Applying the methodologies to hierarchies with large numbers of concepts and rich sets of relationships may shed more light on the manageability and scalability of the techniques described in this dissertation.

One of the limitations of the taxonomy approach is that it depends on the existing relationships defined for a hierarchy of SNOMED. More research is required on how to handle the hierarchies without any outgoing relationships. The use of converse relationships is investigated in [68] as an initial attempt to resolve the issue.

Following the research theme emphasized in this dissertation that complex concepts are more error prone and thus deserve special attention, only one particular kind of “complex” concepts, the overlapping concepts, is investigated in this research. Further research is needed to classify other kinds of complex concepts, by different structural and semantic features of the concepts, which may require further refinement of the taxonomies discussed here.

Furthermore, the taxonomies implemented in this dissertation are only applicable to DL-based terminologies. There may be certain kinds of structures that can only occur when primitive concepts are present. Thus, they may very well have an impact on the complexity that is seen in this work. More research is needed to explore the feasibility of extending the techniques for other families of terminologies and terminological system.

REFERENCES

[1] SNOMED CT. Available at http://www.ihtsdo.org/snomed-ct/. Accessed March 29, 2012.

[2] Department of Health and Human Services, Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Record Technology; Final Rule, 45 CFR Part 170, July 28, 2010.

[3] IHTSDO: International Health Terminology Standards Development Organisation. Available at http://www.ihtsdo.org. Accessed March 29, 2012.

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

[5] Spackman K. SNOMED RT and SNOMED CT. Promise of an international clinical terminology. MD Comput. 2000 Nov-Dec;17(6):29.

[6] Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp. 1997:640-4.

[7] McCray AT. Representing Biomedical Knowledge in the UMLS Semantic Network. In: Broering NC, editor. Proc. High-Performance Medical Libraries: Advances in Information Management for the Virtual Era. Westport, CT; 1993. p. 45-55.

[8] McCray AT, Hole WT. The Scope and Structure of the First Version of the UMLS Semantic Network. In: Proc. 14th Annual SCAMC. Los Alamitos, CA; 1990. p. 126-130.

[9] Gu H, Perl Y, Geller J, Halper M, Liu L, Cimino JJ. Representing the UMLS as an OODB: Modeling Issues and Advantages. JAMIA 2000;7(1):66-80. Selected for reprint in: R. Haux and C. Kulikowski, editors, Yearbook of Medical Informatics: Digital Libraries and Medicine (International Medical Informatics Association), pages 271-285, Schattauer, Stuttgart, Germany, 2001.

[10] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003;36(6):414-432.

[11] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. In: Proc. Medinfo2001. London, UK; 2001. p. 171-175.

[12] 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 2003;35(3):194-212.

[13] Zhang L, Perl Y, Halper M, Geller J, Hripcsak G. A Lexical Metaschema for the UMLS Semantic Network. Artif Intell Med 2005;33(1):41-59.

[14] Perl Y and Geller J, editors, Special Issue on Structural Issues in UMLS Research, J Biomed Inform, 36(6):409-517, December 2003.

[15] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: Representing Different Views of Biomedical Concepts. Bull Med Libr Assoc 1993;81(2):217-222.

[16] Tuttle MS, Sherertz DD, Erlbaum MS, et al. Adding your Terms and Relationships to the UMLS Metathesaurus. In: Proc. 15th Annual Symposium on Computer Applications in Medical Care; 1991. p. 219-223.

[17] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-Based Approaches to the Maintenance of a Large Controlled Medical Terminology. JAMIA 1994;1(1):35-50.

[18] Liu L, Halper M, Geller J, Perl Y. Using OODB Modeling to Partition a Vocabulary into Structurally and Semantically Uniform Concept Groups. IEEE Trans Knowledge & Data Engineering 2002;14(4):850-866.

[19] Liu L, Halper M, Geller J, Perl Y. Controlled Vocabularies in OODBs: Modeling Issues and Implementation. Distributed and Parallel Databases 1999;7(1):37-65.

[20] Gu H, Halper M, Geller J, Perl Y. Benefits of an Object-Oriented Database Representation for Controlled Medical Terminologies. JAMIA 1999;6(4):283-303.

[21] Niles I, Pease A. Origins of the IEEE Standard Upper Ontology. In: Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology. Seattle, WA; 2001.

[22] Niles I, Pease A. Towards a Standard Upper Ontology. In: Proc. FOIS 2001. Ogunquit, ME; 2001.

[23] Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge, MA: The MIT Press; 1998.

[24] Niles I, Pease A. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In: Proc. 2003 Int’l Conference on Information and Knowledge Engineering (IKE’03). Las Vegas, NV; 2003.

[25] Penz JF, Brown SH, Carter JS, et al. Evaluation of SNOMED coverage of Veterans Health Administration terms. Stud Health Technol Inform. 2004;107(Pt 1):540-4.

[26] Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. J Am Med Inform Assoc. 1996 May-Jun;3(3):224-33.

[27] Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. J Am Med Inform Assoc. 1997 May-Jun;4(3):238-51.

[28] Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. J Am Med Inform Assoc. 2009 Jan-Feb;16(1):89-102.

[29] Zhang GQ, Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annu Symp Proc. 2010:922-6.

[30] Campbell KE, Tuttle MS, Spackman KA. A "lexically-suggested logical closure" metric for medical terminology maturity. Proc AMIA Symp. 1998:785-9.

[31] Pacheco E, Stenzhorn H, Nohama P, Paetzold J, Schulz S. Detecting underspecification in SNOMED CT concept definitions through natural language processing. AMIA Annu Symp Proc. 2009: 492-6.

[32] Agrawal A, Elhanan G, Halper M. Dissimilarities in the logical modeling of apparently similar concepts in SNOMED CT. AMIA Annu Symp Proc. 2010:212-6.

[33] Mendonca EA, Cimino JJ, Campbell KE, Spackman KA. Reproducibility of interpreting "and" and "or" in terminology systems. AMIA Annu Symp Proc. 1998:790-4.

[34] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Stud Health Technol Inform. 2004;107(Pt 1):482-6.

[35] Ceusters W, Smith B, Kumar A, Dhaen C. Mistakes in medical ontologies: where do they come from and how can they be detected? Stud Health Technol Inform. 2004;102:145-63.

[36] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating Subsumption in DL-Based Terminologies: A Case Study in SNOMED CT. In: Hahn U, Schulz S, Cornet R, editors. First Int'l Workshop on Formal Biomedical Knowledge Representation (KR-MED 2004). Whistler, Canada; 2004. p. 12-20.

[37] Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. Artif Intell Med. 2007 Mar;39(3):183-95.

[38] Schlobach S, Huang Z, Cornet R, Van Harmelen F. Debugging incoherent terminologies. Journal of Automated Reasoning 39 (2007) 317-349.

[39] Cornet R, Abu-Hanna A. Auditing description-logic-based medical terminological systems by detecting equivalent concept definitions. Int'l Journal of Medical Informatics (2008), doi:10.1016/j.ijmedinf.2007.06.008.

[40] Wade G, Rosenbloom ST. The impact of SNOMED CT revisions on a mapped interface terminology: terminology development and implementation issues. J Biomed Inform. 2009 Jun;42(3):490-3.

[41] Zhu X, Fan JW, Baorto DM, Weng C, J. J. Cimino. A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomed Inform. 2009 Jun;42(3):413-25.

[42] Schulz EB, Barrett JW, Price C. Read Code quality assurance: From simple syntax to semantic stability. J Am Med Inform Assoc. 1998 Jul-Aug;5(4):337-346.

[43] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):41-51.

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

[45] McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods of Information in Medicine 34 (1995) 193-201.

[46] 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 Nov-Dec;13(6):676-90.

[47] NCI Thesaurus. Available at http://ncit.nci.nih.gov/. Accessed March 29, 2012.

[48] 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.

[49] Halper M, Wang Y, Min H, et al. Analysis of error concentrations in SNOMED. AMIA Annu Symp Proc. 2007:314-8.

[50] 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.

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

[52] Jurisica I, Mylopoulos J, Yu E. Ontologies for Knowledge Management: An Information Systems Perspective. Knowledge and Information Systems 2004;6(4):380-401.

[53] Noy NF, Hafner CD. The State of the Art in Ontology Design: A Survey and Comparative Review. AI Magazine 1997;18(3):53-74.

[54] Garey MR, Johnson DS. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York, NY: W. H. Freeman; 1979.

[55] TDE - Terminology Development Environment. Available at http://www.apelon.com/Products/TDE/tabid/100/Default.aspx. Accessed March 29, 2012.

[56] SNOMED CT CliniClue Browser. Available at http://www.cliniclue.com/. Accessed March 29, 2012.

[57] The Protégé. Ontology Editor and Knowledge Acquisition System. Available at http://protege.stanford.edu/. Accessed March 29, 2012.

[58] Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing Concept Categorizations in the UMLS. Artif Intell Med 2004;31(1):29-44.

[59] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Boca Raton, FL: CRC Press; 1993.

[60] Lomax J, McCray AT. Mapping the Gene Ontology into the Unified Medical Language System. Comparative and Functional Genomics 2004;5(5):354-361.

[61] Lincoln MJ, Brown SH, Nguyen V, Cromwell T, et al. U.S. Department of Veterans Affairs Enterprise Reference Terminology Strategic Overview. In: Fieschi M, et al., editors. Proc. Medinfo2004. San Francisco, CA; 2004. p. 391-395.

[62] Dolin RH, Mattison JE, Cohn S, et al. Kaiser Permanente's Convergent Medical Terminology. In: Fieschi M, Coiera E, Li YC, editors. Proc. Medinfo 2004. San Francisco, CA; 2004. p. 346-350.

[63] Rosse C, Mejino JLV. A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics 2003;36(6):478-500.

[64] RxNorm. Available at http://www.nlm.nih.gov/research/umls/rxnorm/. Accessed March 29, 2012.

[65] Rocah RA, Huff SM, Haug PJ, Warner HR. Design a controlled medical vocabulary server: the VOSER project. Computer and Biomedical Research 1994;27(6):472-507.

[66] 3M Worldwide. URL: http://www.3m.com/product/information/Healthcare-Data-Dictionary.html. Accessed March 29, 2012.

[67] SNOMED Clinical Terms Reference Sets: Technical Specification. College of American Pathologists.

[68] Wei D, Halper M, Elhanan G, et al. Auditing SNOMED relationships using a converse Abstraction Network. AMIA Annu Symp Proc. 2009:685-9.

[69] Wei D, Wang Y, Perl Y, Xu J, Halper M, Spackman KA. Complexity measures to track the evolution of a SNOMED hierarchy. AMIA Annu Symp Proc. 2008:778-82.

[70] Wang Y, Wei D, Xu J, et al. Auditing complex concepts in overlapping subsets of SNOMED. AMIA Annu Symp Proc. 2008:273-7.

[71] Even S. Graph algorithms. Potomac (MD): Computer Science Press; 1979.

[72] Good P. Permutation, Parametric, and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling. 3rd ed. New York, NY: Springer; 2005.

[73] Gu H, Hripcsak G, Chen Y, et al. Evaluation of a UMLS auditing process of semantic type assignments. AMIA Annu Symp Proc. 2007:294-8.