Spring 5-31-2006

Structural auditing methodologies for controlled terminologies

Hua Min
New Jersey Institute of Technology

ABSTRACT

STRUCTURAL AUDITING METHODOLOGIES FOR CONTROLLED TERMINOLOGIES

by
Hua Min

Several auditing methodologies for large controlled terminologies are developed. These are applied to the Unified Medical Language System (UMLS) and the National Cancer Institute Thesaurus (NCIT). Structural auditing methodologies are based on the structural aspects such as IS-A hierarchy relationships groups of concepts assigned to semantic types and groups of relationships defined for concepts. Structurally uniform groups of concepts tend to be semantically uniform. Structural auditing methodologies focus on concepts with unlikely or rare configuration. These concepts have a high likelihood for errors.

One of the methodologies is based on comparing hierarchical relationships between the META and SN, two major knowledge sources of the UMLS. In general, a correspondence between them is expected since the SN hierarchical relationships should abstract the META hierarchical relationships. It may indicate an error when a mismatch occurs.

The UMLS SN has 135 categories called semantic types. However, in spite of its medium size, the SN has limited use for comprehension purposes because it cannot be easily represented in a pictorial form, it has many (about 7,000) relationships. Therefore, a higher-level abstraction for the SN called a metaschema, is constructed. Its nodes are meta-semantic types, each representing a connected group of semantic types of the SN. One of the auditing methodologies is based on a kind of metaschema called a cohesive metaschema. The focus is placed on concepts of intersections of meta-semantic types. As is shown, such concepts have high likelihood for errors.

Another auditing methodology is based on dividing the NCIT into areas according to the roles of its concepts. Moreover, each multi-rooted area is further divided into p-areas that are singly rooted. Each p-area contains a group of structurally and seman-

tically uniform concepts. These groups, as well as two derived abstraction networks called taxonomies, help in focusing on concepts with potential errors. With genomic research being at the forefront of bioscience, this auditing methodology is applied to the Gene hierarchy as well as the Biological Process hierarchy of the NCIT, since processes are very important for gene information. The results support the hypothesis that the occurrence of errors is related to the size of p-areas. Errors are more frequent for small p-areas.

STRUCTURAL AUDITING METHODOLOGIES FOR CONTROLLED TERMINOLOGIES

by
Hua Min

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

Computer Science Department

May 2006

APPROVAL PAGE

STRUCTURAL AUDITING METHODOLOGIES FOR CONTROLLED TERMINOLOGIES

Hua Min

Dr. Yehoshua Perl, Dissertation Advisor — Date
Professor, New Jersey Institute of Technology

Dr. Michael Halper, Dissertation Co-Advisor — Date
Professor, Kean University

Dr. Barry Cohen, Committee Member — Date
Assistant Professor, New Jersey Institute of Technology

Dr. Gai Elhanan, Committee Member — Date
3M Health Information Systems

Dr. James Geller, Committee Member — Date
Professor, New Jersey Institute of Technology

Dr. Helen Gu, Committee Member — Date
Associate Professor, University of Medicine and Dentistry of New Jersey

BIOGRAPHICAL SKETCH

Author: Hua Min

Degree: Doctor of Philosophy

Date: May 2006

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2006
- Master of Science in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 1999
- Bachelor of Science in Medicine,
  Nanjing TieDao Medical College, Nanjing, P.R. China, 1993

Major: Computer Science

Presentations and Publications:

H. Min, M. Oren, B. Cohen, Y. Perl, and M. Halper, Structural Auditing Techniques for Gene hierarchy in NCI Thesaurus. To be submitted for journal publication.

H. Min, Y. Perl, Y. Chen, M. Halper, J. Geller, and Y. Wang, Auditing as Part of the Terminology Design Life Cycle. Submitted for journal publication.

Y. Wang, M. Halper, H. Min, Y. Perl, Y. Chen, J. Geller, and K. A. Spackman, Structural Techniques for Auditing SNOMED. To be submitted for journal publication.

H. Gu, Y. Perl, G. Elhanan, H. Min, L. Zhang, and Y. Peng, Auditing Concept Categorizations in the UMLS. *Artificial Intelligence of Medicine*, 31(1), pp. 29–44, 2004.

JJ. Cimino, H. Min, and Y. Perl, Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *Journal of Biomedical Informatics*, 36(6), pp. 450–461, Dec. 2003.

H Gu, H. Min, Y. Peng, and L. Zhang, Using the metaschema to audit UMLS classification errors. In I. S. Kohane, editor, *Proc. of the 2002 AMIA Annual Symposium*, pp. 310–314, San Antonio, TX, Nov. 2002.

To my husband, Fang and my daughter, Julia

ACKNOWLEDGMENT

I would like to extend my sincere gratitude to my advisors, Dr. Yehoshua Perl and Dr. Michael Halper. Their invaluable guidance and encouragement have contributed significantly to the work presented in this dissertation. I would also like to extend a warm thanks to Dr. James Geller for his guidance and help throughout this research. Special thanks to Dr. James J. Cimino for providing invaluable advice.

I would also like to thank Dr. Barry Cohen, Dr. Gai Elhanan, and Dr. Helen Gu for serving as members of my committee.

I thank Yan Chen and Yue Wang for their great help and support throughout the four years of my research and for the joyful days in the Medical Informatics Lab.

I will always be indebted to my parents. Without their moral and intellectual guidance throughout my life, all this would have been impossible. Also, I wish to thank my brother for his continuous support and encouragement.

I dedicate this dissertation to my husband, Fang Zhou, for his love, understanding, help, and support.

TABLE OF CONTENTS

Chapter		Page
1	INTRODUCTION	1
	1.1 Terminologies and Their Complexity	1
	1.2 Importance of Auditing Terminologies	2
	1.3 Literature Review	5
	1.4 UMLS and NCIT	5
	1.5 Structural Auditing	6
	1.6 Overview	6
2	CONSISTENCY ACROSS THE HIERARCHIES OF THE UMLS SEMANTIC NETWORK AND METATHESAURUS	10
	2.1 Background: UMLS	10
	2.2 Methods	11
	2.3 Results	15
	2.3.1 *Clinical Drug* Relationship Sets	17
	2.3.2 *Medical Device* Relationship Sets	18
	2.3.3 *Body Part, Organ or Organ Component* Relationship Sets	18
	2.3.4 *Body Location or Region and Body Space or Junction* Relationship Sets	20
	2.3.5 *Disease or Syndrome and Pathologic Function* Relationship Set	20
	2.3.6 Small Unexplained Relationship Sets	20
	2.3.7 Missing-Ancestor-Descendant-Link	22
	2.4 Discussion	23
	2.5 Conclusions	26
3	AUDITING CONCEPT CATEGORIZATIONS IN THE UMLS USING A META SCHEMA OF SN	27
	3.1 Background: Metaschema of the SN	27
	3.2 Auditing Methodology	28

TABLE OF CONTENTS
(Continued)

Chapter **Page**

3.2.1 Intersection of Semantic Types 30

3.2.2 Meta-semantic Type Association 31

3.2.3 Intersection of Meta-semantic Types 32

3.2.4 Pure Intersection of Meta-semantic Types 34

3.3 Results 37

3.3.1 Analysis of Small Pure Intersections 37

3.3.2 Analysis of Large Pure Intersections 44

3.4 Discussion 48

3.5 Conclusion 52

4 AUDITING AS PART OF THE TERMINOLOGY DESIGN LIFE CYCLE . . . 54

4.1 Background: NCI Thesaurus 54

4.2 Methods 55

4.2.1 Dividing a Terminology into Areas 55

4.2.2 Dividing an Area into P-Areas 58

4.2.3 Auditing Methodology 60

4.3 Results 63

4.3.1 AT and PAT for a NCIT Hierarchy 63

4.3.2 Errors Found in P-areas 67

4.3.3 Testing the Hypotheses 73

4.4 Discussion 75

4.5 Conclusions 79

5 STRUCTURAL AUDITING TECHNIQUES FOR GENE HIERARCHY IN NCI THESAURUS 80

5.1 Background 80

5.1.1 Structural Characteristics of the NCIT Gene Hierarchy 80

5.1.2 Importance of Auditing Gene Hierarchy in the NCIT 82

TABLE OF CONTENTS
(Continued)

Chapter **Page**

5.2 Auditing Methodology 83

5.2.1 Review of the Top-level Area: $\emptyset$ 85

5.2.2 Review First-level Areas having No Children 85

5.2.3 Review Large Areas with Number of P-areas Close to Number of Concepts 85

5.2.4 Hypotheses 86

5.3 Results 86

5.3.1 AT and PAT for the NCIT Gene Hierarchy 86

5.3.2 Role Errors Discovered 93

5.3.3 Error Distributions in P-areas and Areas 96

5.4 Discussion 99

5.4.1 Advantages of the AT and PAT 99

5.4.2 Improving the Modeling of the Gene Hierarchy 100

5.4.3 Transfer of Concepts between Areas 104

5.5 Conclusion 105

6 SUMMARY 108

REFERENCES 110

LIST OF TABLES

Table		Page
2.1	Reasons for Unexplained Relationship Sets from a Sample of Small Sets	21
3.1	Distribution of Number of Concepts for Pure Intersections	38
3.2	Examples of Various Types of Incorrect Categorizations	40
3.3	Analysis of Errors in Small Pure Intersections	44
3.4	Largest Pure Intersections and Their Cardinalities	46
3.5	Semantically Suspicious Medium-size Pure Intersections	48
4.1	Analysis of Errors by P-areas Sizes	74
4.2	Distribution of Areas by Their Cardinality and Number of $\bar{3}$-p-areas	74
4.3	Analysis of Errors in $\bar{3}$-p-areas of Different Kinds of Areas	75
4.4	Analysis of Errors in $\bar{3}$-p-areas of Different Kinds of Areas	77
5.1	Degree Distribution for Category Concepts of the Gene Hierarchy	81
5.2	Characteristics of the Gene Hierarchy Levels	81
5.3	Distribution of Areas, P-areas, and Concepts by Level	90
5.4	Areas with Their Numbers of Concepts and P-areas	91
5.5	P-area Size Distribution	93
5.6	Missing Roles for Concepts in $\emptyset$	94
5.7	Erroneous Concept Distributions by Size of P-areas	97
5.8	Error Distribution of Areas and P-areas	98
5.9	Concepts with Multiple Parents	103
5.10	List of New Intersection Category Concepts	106
5.11	Movement of Concepts with Missing Roles	107

LIST OF FIGURES

Figure		Page
2.1	Examples of expected IS-A relationships between concepts.	12
2.2	Examples of unexpected IS-A relationships between concepts.	13
2.3	Unexplained relationship sets with greater than 30 concept pairs.	16
2.4	One example of an unexpected IS-A relationship found in the META.	18
3.1	The cohesive metaschema hierarchy of the UMLS Semantic Network.	29
3.2	Example of the intersection of semantic types.	31
3.3	Example of meta-semantic type association.	32
3.4	Example of the intersection of meta-semantic types.	32
3.5	Example of pure intersections.	35
4.1	Examples of an area taxonomy and p-area taxonomy.	57
4.2	AT for the Biological Process hierarchy.	64
4.3	PAT for the Biological Process hierarchy.	65
4.4	A portion of the PAT for the Biological Process hierarchy.	66
4.5	Descendants of ORGANISMAL PROCESS in (a) NCIT hierarchy indented format, and (b) as selected areas and p-areas in the diagram format.	68
4.6	Areas and p-areas of the Organismal Process subhierarchy after corrections.	70
5.1	The flow chart of the auditing methodology.	84
5.2	Area taxonomy for the Gene hierarchy.	88
5.3	Excerpt of the p-area taxonomy for the Gene hierarchy.	89
5.4	Another excerpt of the p-area taxonomy.	90
5.5	Another excerpt of the p-area taxonomy.	92
5.6	Example of new modeling of the p-area taxonomy.	101
5.7	Example of a transformation of the hierarchy of concepts with multiple parents.	105

CHAPTER 1

INTRODUCTION

Large biomedical terminologies have become increasingly important resources for medical researchers. Modern biomedical data sets are annotated with standard terms to describe the data and to support data linking between terminologies. Many controlled terminologies are created for different applications. Some of them integrate others. The accuracy of one system impacts other systems. Thus, quality assurance plays important role in the life cycle of controlled terminologies.

1.1 Terminologies and Their Complexity

Controlled medical terminologies have been recognized as important tools in a variety of medical informatics applications ranging from patient-record systems to decision support systems. A controlled terminology usually has large size and high complexity, as it may have hundreds thousands and up to over a million concepts and each concept may have many relationships. The size and complexity of a terminology can make it difficult to comprehend and use [30]. Its size also poses great challenges in system maintenance.

Controlled terminology may have one domain (e.g., National Cancer Institute Thesaurus (NCIT)). It may also have multiple domains that integrate other terminologies. For example, the Unified Medical Language System (UMLS) [10, 33, 34, 40] integrates about 100 well established medical terminologies into a unified knowledge representation framework. The integration of terminology sources may introduce inconsistencies, leading to further confusion. Every time one of the individual terminologies changes, those changes must be reflected in the integrated terminology.

1.2 Importance of Auditing Terminologies

As the software industry matured, different models for software life cycle processes have been developed. However, in recent years it has become clear that no such model is complete without activities dedicated to assuring the correctness of the software. Typically, software life cycle models list auditing as part of quality assurance, one of the support activities (see [62] for example). It is normally assumed that a team that is independent of the development team performs auditing. Such life cycle models have also been expanded to knowledge-based systems. For an application of auditing in the development of knowledge-based expert systems for business and finance, see [66]. However, it is observed that auditing has been typically absent from the life cycle of many ontologies, terminologies, and controlled vocabularies, and that this omission needs to be rectified. Auditing is essential since terminologies underlie decision-support systems, clinical patient records, health care administrative systems, etc. Errors in a terminology will propagate to errors in these systems, which in turn may result in endangering the life or quality of life of a patient and unnecessary cost.

Categorizations of concepts by experts are not necessarily consistent since domain experts may have different knowledge backgrounds, views, and priorities. It is unavoidable that some errors are introduced. The accuracy of a terminology is critical for its developers and users. Recognizing the importance of auditing as an integral part of the terminology design life cycle is essential for the terminology industry.

A terminology may be integrated into another terminology. The quality assurance effort for a source terminology impacts the quality of an integrated terminology, as the integration may spread the errors of one terminology into others. The integration process may also introduce inconsistencies.

The common perception in the terminology "industry" reflected in anecdotal evidence is that customers want to increase coverage, and this is what they are willing to "pay" for. Note that the terminology "industry" include departments in corporations, government

agencies, hospitals, and academic institutions that design, maintain, and use terminologies. If a customer discovers an error and complains about it, it will be fixed. But undertaking an extensive auditing effort is typically not what the customer wants.

Such a situation is common in emerging industries, but not in mature ones. Just imagine a public company not willing to audit its financial statements. The SEC would quickly penalize it since the trust of shareholders is based on the assumption that the financial statements accurately reflect the "value" of the company and its transactions. Similarly, one cannot imagine nowadays the pharmaceutical, automobile, or airline industry without extensive investment in quality assurance of their products.

Terminologies are now being created by a maturing industry. The recent emergence of a generation of medical terminologies satisfying the desiderata of Cimino [16,18] support this claim. These terminologies have sound theoretical models such as description logics [2, 50] and frames [49]. Examples include the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [61], the Foundational Model of Anatomy (FMA) [56], the National Cancer Institute Thesaurus (NCIT) [51], Lab LOINC [47], and the Medical Entities Dictionary (MED) [18]. They all have accompanying software tools—either commercial or of commercial quality—that provide users with convenient interfaces: e.g., Protégé [55] used for FMA and Apelon's Terminology Development Environment (TDE) [1] used for SNOMED and NCIT. These terminologies tend to be of substantial size and complexity, and they keep growing; e.g., the recent version (July '05) of SNOMED contains 366,179 concepts, while the January '03 version contained only 344,609 concepts. Similarly, NCIT has grown in two years from about 25,000 concepts to 42,404 concepts.

The NJIT Medical Informatics research group, jointly with Dr. James J. Cimino and Dr. George Hripcsak of Columbia University, distributed a questionnaire about Unified Medical Language System (UMLS) [10] users' applications and priorities to the UMLS users mailing list. There are 70 responses. Three questions dealt with auditing. Two asked to what extent the user is bothered by a list of twelve kinds of errors, with the

choice of answers: "not at all," "a little," "moderately," and "a lot," coded by the values 0, 1, 2, and 3, respectively. It became clear from the results of the survey that there is a strong demand for high-quality auditing. For example, the average user is approximately "bothered moderately" (1.97) by errors such as incorrect hierarchical relationships (1.97), incorrect associative relationships (2.11), incorrect semantic-type assignments (2.15), missing hierarchical relationships (1.86), and missing semantic-type assignments (1.76). For the other kinds of errors listed in the questionnaire, the average user was bothered to an extent that is between "a little" and "moderately" (1.46).

Furthermore, the responding UMLS users clearly saw auditing as a high priority since, on average, they would allocate 35% of a putative NLM budget to auditing, the highest of all given options by a large margin. The three trailing categories, "designing a derived terminology," "improving interfaces," and "extending coverage," were assigned only 24%, 20%, and 16% of the budget, respectively.

In summary, the results of our study showed that users of the UMLS care about eliminating errors and would like to see a substantial portion of the available budget allocated to auditing activities for quality assurance. These results confirm the claims that UMLS users are demanding serious auditing efforts so they can rely on the represented knowledge with a reasonable level of confidence. It is a research issue whether users of other medical terminologies share similar opinions to those expressed by the UMLS users in this study.

Assuring the consistency and correctness of a terminology is an ongoing challenge facing its designers. It is a difficult (even overwhelming) task to audit all concepts and their related knowledge in a terminology. Usually, there are not enough resources available for such a task. Thus, auditing methodologies need to be developed to assist human reviewers in accomplishing this job. To be effective, these methodologies should focus the attention of auditors on a relatively small number of concepts with a high likelihood for errors. This way, the limited resources available for auditing are best utilized.

1.3 Literature Review

A number of researchers have developed different methodologies to help with auditing terminologies. There are several auditing approaches for the UMLS [10,40]. For example, semantic methods are being used to detect concept classification errors [15]. Techniques have been developed for discovering errors in concept hierarchies, e.g., circular hierarchical relationships [4]. The problems of concept redundancy and ambiguity were addressed in [17], while redundant categorization is considered in [52]. Revising the UMLS Semantic Network (SN) [42,43,45,46] through the reclassification of semantic types was discussed in [57]. Object-oriented models have been constructed to support navigation, maintenance, and auditing [5, 29]. A method for finding undetected synonymy in the UMLS has been developed in [32].

Auditing the Systematized Nomenclature for Medicine (SNOMED) [61] based on ontological and linguistic techniques is discussed in [12, 13]. There are serious defects in NCIT when its conformity is assessed with principles of good practice in terminological and ontological design [11, 37]. A technique for auditing the Medical Entities Dictionary (MED) [18] based on an object-oriented database representation appears in [26]. Error detection [21] for the Diagnoses for Intensive Care Evaluation (DICE) system [22] is based on migration to Description Logics. Detecting errors caused by the design problems of the Gene Ontology is addressed in [36,38,58,59,60]. A technique for auditing the Foundational Model of Anatomy (FMA), in its previous name UWDA (University of Washington Digital Anatomist), based on identifying shortcuts, circles, and diamond structures in the FMA presents in [27].

1.4 UMLS and NCIT

The Unified Medical Language System (UMLS) [10,33,34,40] of the National Library of Medicine (NLM), started in 1986, integrates a large number of well established medical terminologies into a unified knowledge representation framework . It also helps to improve

the ability of computer programs to capture biomedical meaning and to use this to retrieve and integrate relevant machine-based information. The UMLS provides users with extensive and up-to-date information which helps to improve decision making and ultimately the quality of patient care as well as research in the healthcare field. For more background on UMLS see Section 2.1.

The National Cancer Institute Thesaurus (NCIT) was designed in response to a need for a consistent, shared vocabulary for the various projects and initiatives at the NCI, as well as in the broader cancer research community. The NCIT covers clinical and basic research as well as administrative terminology. For more background on NCIT see Section 4.1.

1.5 Structural Auditing

Structural auditing is based on techniques that utilize structural aspects such as IS-A hierarchical relationships and the set of relationships defined for a concept. Due to their nature, using efficient programming, concepts can be grouped according to structural aspects to guide manual review of concepts with high likelihood errors. The underlying theme of structural auditing is that there is typically a correspondence between similar structure and similar semantics of concepts. Thus, structurally uniform groups tend to be semantically uniform, although this is not always the case. An unlikely or rare configuration may indicate some concepts with likely errors and structural auditing direct limited manual editing work to these limited number of concepts. Such an approach tend to maximize the impact of typically limited auditing resources.

1.6 Overview

The purpose of this dissertation is to design structural auditing methodologies for the UMLS and NCIT. These auditing methodologies focus the limited resources of human reviewers on the problematic areas that are most likely to contain erroneous concepts.

Based on the structural characters of the UMLS and NCIT, several methodologies are designed as follows:

The UMLS contains two knowledge resources, the Metathesaurus (META) and the Semantic Network (SN). Both resources include hierarchical information: the SN organizes semantic types in a strict IS-A hierarchy, while META has a collection of a variety of hierarchical relationships between pairs of concepts. The two resources are connected by the assignment of one or more semantic types from the SN to each concept in META. Due to the large size and complexity of the META, the creation and maintenance of the system is difficult. Automated tools are developed to assist human reviewers with the management tasks [63]. The automated methods can help focus the limited resources of human review to the cases most likely to need attention. One automated methodology is designed in Chapter 2 [19]. It can automatically identify inconsistencies in the META by comparing the parent-child relationships between concepts in the META and the ancestor-descendant relationships between the corresponding semantic types in SN. The auditing is focused on high error likelihood area identified by the inconsistencies.

Though SN forms an abstraction of META, SN is still large and complex. It may be difficult to view and comprehend. Metaschema is one approach to help SN comprhension. It partitions the SN into structurally uniform sets of semantic types based on the distribution of the relationships within the SN. One methodology for partitioning the SN of UMLS has been introduced in [14]. It groups closely related semantic types into semantic-type collections represented as meta-semantic types. The network of meta-semantic types connected by hierarchical and semantic relationships is called metaschema [54]. For background on metaschema see Section 3.1. The metaschema provides a higher-level abstract view of the SN. Each concept can be assigned to several semantic types. It can also be assigned to several meta-semantic types. The concept is more likely to be erroneously assigned to semantic types of different meta-semantic types than to semantic types of the same meta-semantic type because of larger semantic distance. The auditing effort should

concentrate on concepts that are assigned to different meta-semantic types. The idea is that such concepts have a high likelihood of errors and inconsistencies. A auditing methodology based on the cohesive metaschema of the UMLS is designed in Chapter 3 [28]. The auditing methodology consists of three parts: (1) identify all concepts of intersections of two or more meta-semantic types. (2) refine each one of the intersections into multiple pure intersections. (3) domain experts review each pure intersection containing a small number of concepts of similar semantics. Furthermore, the combination of intersecting semantic types of each pure intersection containing medium and large numbers of concepts are reviewed to verify that it is semantically sound. The concepts of the pure intersections, which are not semantically sound, are reviewed by a domain expert. This auditing methodology is designed to minimize the effort and maximize the likelihood of finding errors.

For the UMLS, the SN and metaschema provide high-level abstractions. The auditing methodologies can identify high error likelihood areas from the inconsistencies of high-level abstractions. The high-level abstractions provide a framework to design the structural auditing methodologies. Unfortunately, there is no such kind of high-level abstractions for most controlled terminologies (e.g., NCIT). It may be one necessary step to construct high-level abstractions for such terminologies. NCIT is selected to conduct the experiment due to its relatively small size. The auditing methodology for terminologies satisfying systematic inheritance (Chapter 4) comprises two major phases: (1) the automated preparatory phase; and (2) the manual guided-discovery phase. Phase (1) consists of four steps. First, the terminology's concepts are divided into groups. The concepts of each group have the exact same roles. This division provides structurally uniform collections of concepts. From this division, the second step constructs a compact abstraction network, called an *area taxonomy*. Thirdly, the division is refined into groups of concepts called p-areas that are both structurally uniform and singly rooted. Finally, an enhanced abstraction network, called the *p-area taxonomy* is derived. It is very difficult to comprehend terminologies because they are typically huge in size (number of concepts) and have high complexity

(proportional to the number of relationships) [30]. Auditing, which requires comprehension, is even more difficult since it is like finding needles in a haystack. The two taxonomies derived in Phase (1) provide compact, comprehensible views of the terminology. Such representations tend to highlight relevant features of the terminology while at the same time hiding unimportant details. In Phase (2)—the actual auditing phase—elements of the p-area taxonomy are used to guide the auditor to suspicious parts of the terminology. It shows that areas of small size tend to denote irregularities in the terminology and therefore may reveal errors. The p-area taxonomy readily exposes such situations to the auditor. A hypothesis reflects the concentration of errors in some groups of concepts exposed by the taxonomies. An application of our methodology to the Biological Process hierarchy of the NCIT is presented. The results are analyzed and confirm the hypothesis.

Genomic research is at the forefront of bioscience. The recent achievements in genomic research are attracting increasing public and scientific interest. The gene terminology plays a critical role for genomic research. The rapid growth of genomic information over the past few years makes it ever more important to provide a methodology of quality assurance for the genomic components of medical terminologies. The above structural auditing methodology is also applied to audit role errors of the Gene hierarchy of the NCIT (Chapter 5). Due to the special structure of the Gene hierarchy, i.e., all genes are leaves, there is a need to design a slightly modified auditing methodology for it. Results are presented where many role errors are exposed. Due to the large number of role errors, they first need to be corrected before trying to audit for other errors in the Gene hierarchy.

All together, several different structural auditing methodologies are described in this dissertation. They are all similar in utilizing structural aspects of terminologies to identify groups of concepts with relatively high likelihood for errors. Such methodologies focus the limited resources available for auditing to optimize the impact of improving the quality of the knowledge of terminologies.

CHAPTER 2

CONSISTENCY ACROSS THE HIERARCHIES OF THE UMLS SEMANTIC NETWORK AND METATHESAURUS

2.1 Background: UMLS

The two major UMLS knowledge sources are the META and the SN. The META serves as the central repository of concepts used in the biomedical field. It contains detailed information on concepts that appear in different biomedical terminologies. The META also preserves the meaning, hierarchical connections, and other relationships between concepts represented in its source terminologies.

The basic unit of information in the META is the concept, which is identified by a Concept Unique Identifier (CUI). Each concept in the META is assigned one or more semantic types from the SN. For example, the concept ORGAN (C0178784)[1] has been assigned the semantic type *Body Part, Organ, or Organ Component* (T023, A.1.2.3.1). A second concept, ANATOMIC STRUCTURES (C0700276), has the semantic type *Anatomical Structure* (T017, A1.2).

The META includes a variety of relationships between concepts, provided in a file called MRREL. Relationships include PARENT-CHILD, BROADER-NARROWER, LIKE, and OTHER. They may be further characterized with specific semantic relationships, such as IS-A and PART-OF. For example, MRREL includes a PARENT-CHILD IS-A relationship between the concepts ORGAN and ANATOMIC STRUCTURES, indicating that the former is a more specific concept than the latter.

The purpose of the SN is to provide a consistent categorization of all concepts represented in the META. The SN contains 135 semantic types, as well as hierarchical and non-

[1]Concepts names and relations will be depicted in a "small cap" style. Semantic types and relations from the SN will be depicted in italics.

hierarchical relationships between semantic types [45, 64]. The SN serves as a high-level abstract view [42] of the META.

The semantic types in the SN are arranged in a strict hierarchy (that is, each concept has at most one parent) of *ancestor-descendant* relationships, implicit in the tree address provided for each semantic type. For example, in the current SN, *Anatomical Structure* (with the tree address A1.2) is the immediate *ancestor-of Fully Formed Anatomical Structure* (with the tree address A1.2.3) which, in turn, is the immediate *ancestor-of Body Part, Organ, or Organ Component* (with the tree address A1.2.3.1). The *ancestor-descendant* relationship between the semantic types can be obtained from their tree addresses.

An automated method is presented in this chapter. It identifies inconsistencies in the META by comparing the parent-child relationships between concepts in the META and the ancestor-descendant relationships between the corresponding semantic types in SN.

2.2 Methods

The presence of hierarchies in both the SN and META, and the tight connection between the semantic types and the concepts, suggests a certain symmetry. Given the meaning of "IS-A" (both in plain English and in formal knowledge representation), if CONCEPT 1 IS-A CONCEPT 2, it seems reasonable to assume that both concepts are either of the same semantic type, or else the semantic type of CONCEPT 1 should have an *ancestor-descendant* relationship to the semantic type of CONCEPT 2, either immediate or indirect.[2] Indeed, this is the case with the example presented above: ORGAN IS-A ANATOMIC STRUCTURE, and *Body Part, Organ, or Organ Component is-descendant-of Anatomical Structure.*

Some definitions are needed to describe relationships between concepts in the META that are connected by the IS-A relationship.

[2]The *ancestor-descendant* relationship is transitive; since *Body Part, Organ, or Organ Component is-descendant-of Fully Formed Anatomical Structure*, and *Fully Formed Anatomical Structure is-descendant-of Anatomical Structure*, it is also implied that *Body Part, Organ, or Organ Component is-descendant-of Anatomical Structure.*

Definition (Expected IS-A Relationship): An expected IS-A relationship between a child and a parent concept holds if the semantic types of the parent concept are identical to, or are *ancestor-of*, any of the semantic types of the child concept.

Definition (Unexpected IS-A Relationship): An unexpected IS-A relationship between a child and a parent concept holds if none of the semantic types of the parent concept is identical to, or an *ancestor-of*, any of the semantic types of the child concept.

Figure 2.1 shows examples of "expected IS-A relationships" between concept pairs based on their semantic types, and Figure 2.2 shows examples of "unexpected IS-A relationships" between concept pairs based on the semantic types.

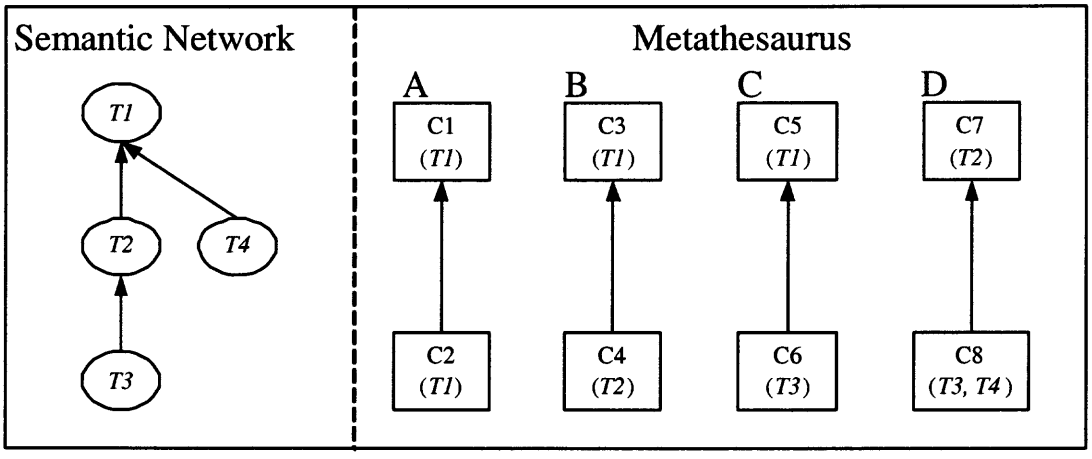


Figure 2.1 Examples of expected IS-A relationships between concepts. A hypothetical subtree of the UMLS Semantic Network is shown on the left, consisting of four semantic types (*T1-T4*). A hypothetical piece of the UMLS META is shown on the right, consisting of eight concepts (C1-C8), arranged in pairs of IS-A relationships (drawn as arrows). Each concept is assigned one or more semantic types, shown in parentheses. In A, both concepts have the same semantic type. In B and C, the relationship between the two semantic types is *ancestor-descendant*. In D, the child concept has two semantic types, but since one of them (*T3*) is a *descendant-of* the parent concept's type (*T2*), it is an expected IS-A relationship.

The method examines the cases where there is an inconsistency between the semantic types assigned to the concepts in META that have an IS-A relationship. Specifically, all instances in MRREL where none of the semantic types of the parent concept is identical to, or an *ancestor-of*, any of the semantic types of the child concept are examined.

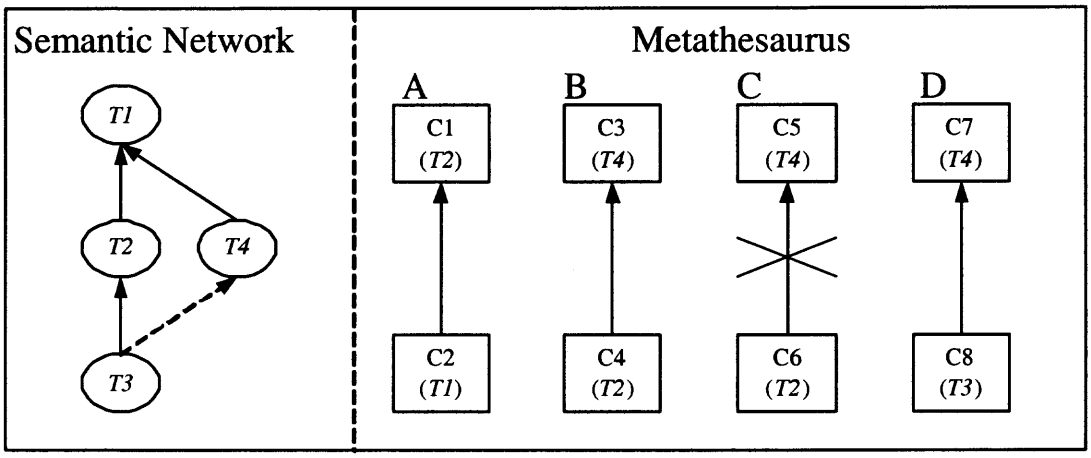


Figure 2.2 Examples of unexpected IS-A relationships between concepts. In A, the semantic type (*T2*) of the parent concept is a *descendant-of* the semantic type (*T1*) of the child concept, suggesting that either the parent concept's semantic type is too specific or the child concept's semantic type is too general. In B, the semantic types of the two concepts are neither the same nor in an *ancestor-descendant* relationship, suggesting that one or the other concepts is missing a semantic type (e.g., C3 might be missing *T1* or *T2*, or C4 might be missing *T4*). C also has unrelated semantic types between two concepts, but in this case the explanation is that the IS-A relationship between C5 and C6 is incorrect. D also has unrelated semantic types in two concepts, but in this case the explanation is that *T4* is conceptually an *ancestor-of T3* (drawn as a dotted arrow) but is not included in the UMLS Semantic Network.

Each unexpected IS-A relationship can be explained by one or more of the following six causes:

1 *Parent-Too-Specific*: the semantic type of the parent concept is a *descendant-of* the semantic type of the child concept; if the parent concept was assigned a less specific semantic type, the IS-A between concepts would become expected (see Figure 2.2A).

2 *Child-Too-General*: the semantic type of the child concept is an *ancestor-of* the semantic type of the parent concept; if the child concept was assigned a more specific semantic type, the IS-A between concepts would become expected (see Figure 2.2A).

3 *Parent-Type-Missing*: if the parent concept were to be assigned an additional semantic type, the IS-A between concepts would become expected (see Figure 2.2B).

4 *Child-Type-Missing*: if the child concept were to be assigned an additional semantic type, the IS-A between concepts would become expected (see Figure 2.2B).

5 *Wrong-Is-A*: the IS-A between the concepts is incorrect (see Figure 2.2C).

6 *Missing-Ancestor-Descendant-link*: if an *ancestor-descendant* link was added to the SN, the IS-A between the concepts would become expected (see Figure 2.2D).

While automated methods can be used to detect inconsistencies, automatic determination of the specific reason for each case is not generally possible. For example, if the semantic type of the child concept is an *ancestor-of* the semantic type of the parent concept, there is no way to automatically determine whether the problem is Parent-Too-Specific or Child-Too-General without human review. This review, in turn, depends on the definitions of the semantic types and (where available) the definitions of the concepts.

After using the automated methods to detect inconsistencies, domain experts need to determine the specific reasons for each case, depending on the definitions of the semantic types and concepts.

To conduct the review, all records having the relationship CHD (a CHILD-OF relationship) and the relationship attribute "IS-A" are extracted from the file MRREL. These records contain two CUIs, CUI1 and CUI2, for which the relationship is CUI1 IS-A CUI2. The preferred English name for each concept is obtained from the file MRCON.

All semantic types associated with each of the CUIs are obtained from the file MRSTY, and the concept pairs are aggregated into "relationship sets" based on the semantic types of the parent and child concepts. Relationships involving concepts with multiple semantic types were aggregated into multiple relationship sets. The names and tree addresses of each semantic type are obtained from the file SRDEF [65]. For example, the concept ANATOMIC STRUCTURES (C0700276) is the parent of HUMAN BODY (C0242821) in the META. They are assigned to sematnic types *Anatomical Structure* (T017 A1.2) and *Human* (T016 A1.1.7.2.5.1), respectively. This concept pair is in the *Anatomical Structure* and *Human* relationship set.

Once the relationship sets are obtained, those that represent expected IS-A relationships are identified. These were cases where the semantic type of the parent concept was either identical to, or an *ancestor-of*, the semantic type of the child concept. It is determined

by examining the tree addresses. For example, the tree address for semantic type *Entity* (T071) is "A", and the tree address for semantic type *Intellectual Product* (T170) is "A2.4". Since "A2.4" has the prefix "A," it implies that *Intellectual product is-descendant-of Entity* in the SN; therefore, the set of relationships from MRREL in which the parent concepts have the semantic type *Entity* and the child concepts have the semantic type *Intellectual Product* is expected. Conversely, since "A" has no prefix "A2.4", the set of relationships from MRREL in which the parent concepts have the semantic type *Intellectual Product* and the child concepts have the semantic type *Entity* is unexpected (see Figure 2.2A). Then domain experts manually examined the unexpected IS-A relationship sets to try to determine the reason why they were occurring (that is, which of the above six causes listed applied).

2.3 Results

In the January 2002 release of the UMLS, there were 10,417,419 records in MRREL; 654,292 records had the relationship "CHD." Of these, 69,991 records had IS-A relationship attributes. These records involved 20,442 unique parent codes and 67,453 unique children codes, with 67,589 unique codes overall (since most parent concepts were also children). These concepts had a total of 68,192 semantic type instances in MRSTY. After merging concepts pairs into relationship sets based on their semantic types and excluding expected IS-A relationship sets, there remained 17,022 relationships in 246 relationship sets. The largest relationship sets, containing over 30 concept pairs, are shown in Figure 2.3. These 34 largest relationship sets represent 13.8% of the 246 relationship sets and account for 16,256 (95.5%) of the 17,022 concept pairs.

Semantic Type of Parent Concept			Semantic Type of Children Concepts			Pairs	Cause
T121	A1.4.1.1.1	*Pharmacologic Substance*	T200	A1.3.3	*Clinical Drug*	9296	Wrong-Is-A
T109	A1.4.1.2.1	*Organic Chemical*	T200	A1.3.3	*Clinical Drug*	723	Wrong-Is-A
T029	A2.1.5.2	*Body Location or Region*	T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	721	Missing-Type
T127	A1.4.1.1.3.4	*Vitamin*	T200	A1.3.3	*Clinical Drug*	582	Wrong-Is-A
T121	A1.4.1.1.1	*Pharmacologic Substance*	T074	A1.3.1	*Medical Device*	412	Wrong-Is-A
T195	A1.4.1.1.1.1	*Antibiotic*	T200	A1.3.3	*Clinical Drug*	362	Wrong-Is-A
T125	A1.4.1.1.3.2	*Hormone*	T200	A1.3.3	*Clinical Drug*	331	Wrong-Is-A
T130	A1.4.1.1.4	*Indicator, Reagent, or Diagnostic Aid*	T200	A1.3.3	*Clinical Drug*	286	Wrong-Is-A
T104	A1.4.1.2	*Chemical Viewed Structurally*	T200	A1.3.3	*Clinical Drug*	281	Wrong-Is-A
T030	A2.1.5.1	*Body Space or Junction*	T029	A2.1.5.2	*Body Location or Region*	261	Missing-Type
T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	T029	A2.1.5.2	*Body Location or Region*	230	Missing-Type
T029	A2.1.5.2	*Body Location or Region*	T030	A2.1.5.1	*Body Space or Junction*	228	Missing-Type
T122	A1.4.1.1.2	*Biomedical or Dental Material*	T200	A1.3.3	*Clinical Drug*	226	Wrong-Is-A
T123	A1.4.1.1.3	*Biologically Active Substance*	T200	A1.3.3	*Clinical Drug*	219	Wrong-Is-A
T196	A1.4.1.2.3	*Element, Ion, or Isotope*	T200	A1.3.3	*Clinical Drug*	219	Wrong-Is-A
T110	A1.4.1.2.1.9.1	*Steroid*	T200	A1.3.3	*Clinical Drug*	197	Wrong-Is-A
T082	A2.1.5	*Spatial Concept*	T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	178	Missing-Type
T024	A1.2.3.2	*Tissue*	T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	174	Missing-Type
T116	A1.4.1.2.1.7	*Amino Acid, Peptide, or Protein*	T200	A1.3.3	*Clinical Drug*	158	Wrong-Is-A
T197	A1.4.1.2.2	*Inorganic Chemical*	T200	A1.3.3	*Clinical Drug*	153	Wrong-Is-A
T118	A1.4.1.2.1.8	*Carbohydrate*	T200	A1.3.3	*Clinical Drug*	148	Wrong-Is-A
T129	A1.4.1.1.3.5	*Immunologic Factor*	T200	A1.3.3	*Clinical Drug*	131	Wrong-Is-A
T130	A1.4.1.1.4	*Indicator, Reagent, or Diagnostic Aid*	T074	A1.3.1	*Medical Device*	122	Wrong-Is-A
T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	T024	A1.2.3.2	*Tissue*	102	Missing-Type
T030	A2.1.5.1	*Body Space or Junction*	T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	96	Missing-Type
T022	A2.1.4.1	*Body System*	T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	86	Missing-Type
T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	T030	A2.1.5.1	*Body Space or Junction*	68	Missing-Type
T120	A1.4.1.1	*Chemical Viewed Functionally*	T200	A1.3.3	*Clinical Drug*	51	Wrong-Is-A
T077	A2	*Conceptual Entity*	T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	43	Missing-Type
T129	A1.4.1.1.3.5	*Immunologic Factor*	T074	A1.3.1	*Medical Device*	43	Wrong-Is-A
T047	B2.2.1.2.1	*Disease or Syndrome*	T046	B2.2.1.2	*Pathologic Function*	33	Child-Too-General
T116	A1.4.1.2.1.7	*Amino Acid, Peptide, or Protein*	T074	A1.3.1	*Medical Device*	33	Wrong-Is-A
T023	A1.2.3.1	*Body Part, Organ, or Organ Component*	T022	A2.1.4.1	*Body System*	32	Missing-Type
T125	A1.4.1.1.3.2	*Hormone*	T074	A1.3.1	*Medical Device*	31	Wrong-Is-A

Figure 2.3 Unexplained relationship sets with greater than 30 concept pairs.

2.3.1 *Clinical Drug* Relationship Sets

The largest unexplained relationship set involves parent concepts of semantic type *Pharmacologic Substance* (T121, A1.4.1.1.1) and child concepts of semantic type *Clinical Drug* (T200, A1.3.3). *Pharmacologic Substance* is defined as "A substance used in the treatment or prevention of pathologic disorders" while *Clinical Drug* is defined as "A pharmaceutical preparation as produced by the manufacturer". This unexplained relationship set contains 9296 concept pairs, accounting for 54.6% of the total unexplained relationship sets. Figure 2.4 shows one example, ANTIFUNGAL AGENTS (C0003308) and its child FLUCONAZOLE 100 MG ORAL TABLET (C0688874). *Clinical Drug* is defined as "a pharmaceutical preparation as produced by the manufacturer" and is an immediate *descendant-of Manufactured Object* (T073, A1.3) in the SN. Each member of this relationship set is an example of the presence of a Wrong-IS-A in MRREL.

The above *Pharmacological Substance/Clinical Drug* relationship set is the largest of 20 relationship sets in which the parent concepts have semantic types that are *descendant-of Chemical* (T103, A1.1.4) and the child concepts have the semantic type *Clinical Drug*. The other 19 unexplained relationship sets (15 of which have over 30 concept pairs and are listed in Figure 2.3) involve an additional 4,123 concept pairs. These sets also represent cases of Wrong-IS-A. An alternative possible explanation for these relationship sets is Parent-Type-Missing; correcting this cause would require assigning the semantic type *Clinical Drug* to concepts such as ANTIFUNGAL AGENTS. Another possible cause is Child-Type-Missing; correcting this cause would require assigning some semantic type from the *Chemical* subtree of the SN to concepts such as FLUCONAZOLE 100 MG ORAL TABLET. The fourth possibility is Missing-Ancestor-Descendant-link; correction would require adding a *descendant-of* relationship between *Clinical Drug* and twenty different *descendants-of Chemical*. Each of these solutions would be a violation of the UMLS's definition of *Clinical Drug*. The information in META supports this view, since the majority of the 81,165 *Clinical Drug* concepts in META are not involved in these unexplained relationships.

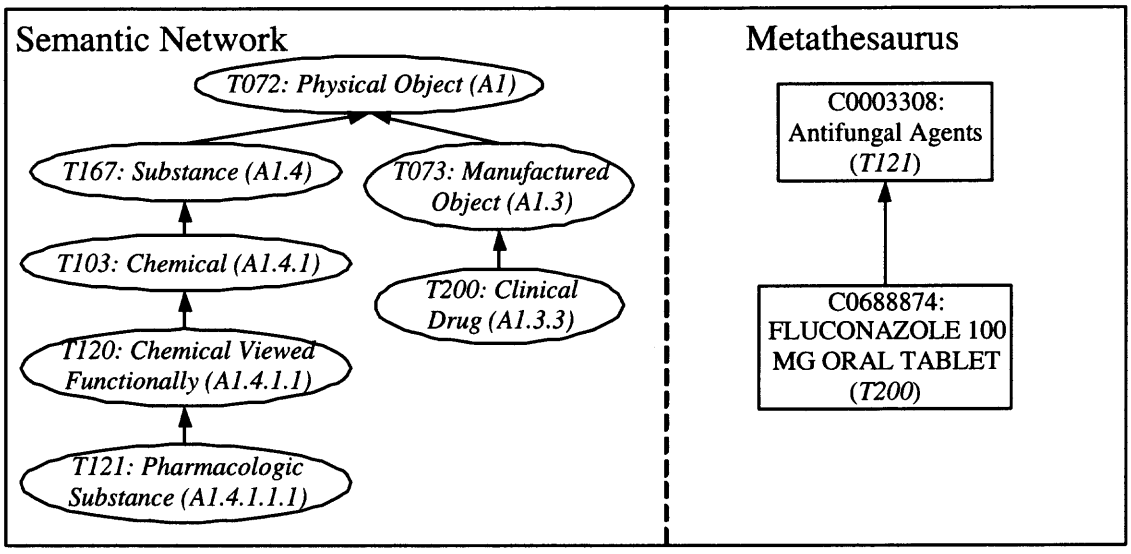


Figure 2.4 One example of an unexpected IS-A relationship found in the META. The parent concept, ANTIFUNGAL AGENTS, has semantic type *Pharmacological Substance*, and the child concept, FLUCONAZOLE 100 MG ORAL TABLET, has semantic type *Clinical Drug*. There is no *ancestor-descendant* relationship between these two semantic types, as shown in the hierarchy at left. See the text for a discussion of possible explanations.

2.3.2 *Medical Device* Relationship Sets

Like *Clinical Drug*, the semantic type *Medical Device* (T074, A1.3.1) is an immediate *descendant-of Manufactured Object*. As with *Clinical Drug*, many concepts with the semantic type *Medical Device* have parent concepts that have a semantic type in the *Chemical* subtree of the SN. There were 667 such concept pairs that were contained in eleven relationship sets (five sets had over 30 concept pairs and are shown in Figure 2.3). These represent cases of Wrong-IS-A, too.

2.3.3 *Body Part, Organ or Organ Component* Relationship Sets

There are 14 unexplained relationship sets (four shown in Figure 2.3) containing 485 concept pairs, in which the parent concepts have the semantic type *Body Part, Organ or Organ Component* (T023, A1.2.3.1). An additional 11 unexplained relationship sets (six shown in Figure 2.3), containing 1,336 concept pairs, have child concepts with the semantic type *Body Part, Organ or Organ Component*. Most of the unexplained concept pairs are

cases of Parent-Type-Missing or Child-Type-Missing; the review of these 25 relationship sets supports this view.

For example, CAPILLARY BED (C0489802) has the semantic type *Body Part, Organ, or Organ Component* and is the parent of SYSTEMIC CAPILLARY BED (C0923301), with semantic type *Body System*. SYSTEMIC CAPILLARY BED should also have the semantic type *Body Part, Organ, or Organ Component* (Child-Type-Missing).

In another example, CARDIAC VENOUS TREE (C0923573) has the semantic type *Body System* (T022, A2.1.4.1) and is the parent of SMALLEST CARDIAC VEINS (C0226663), with semantic type *Body Part, Organ, or Organ Component* (T023, A1.2.3.1). CARDIAC VENOUS TREE should also have the semantic type *Body Part, Organ, or Organ Component* (Parent-Type-Missing).

There are some cases where the IS-A relationship between concepts appears to be wrong. For example, SKELETAL SYSTEM OF UPPER LIMB (C081854), a *Body System*, is listed as a parent of BONY PELVIC GIRDLE (C0934859), a *Body Part, Organ, or Organ Component*. No changes of semantic type assignments will make BONY PELVIC GIRDLE IS-A SKELETAL SYSTEM OF UPPER LIMB a correct IS-A relationship (Wrong-IS-A).

Two of the relationship sets in which the parent concepts have semantic type *Body Part, Organ, or Organ Component* (T023, A1.2.3.1) are special cases. One set has 22 concept pairs in which the child concepts have the semantic type *Fully Formed Anatomical Structure* (T021, A1.2.3); an example is RIGHT BIG TOE (C0930961) IS-A HALLUX (C0018534). The other set has one concept pair in which the child concept has the semantic type *Anatomical Structure* (T017, A1.2): EXTERNAL RECTAL VENOUS PLEXUS (C058-0083) IS-A RECTAL VENOUS PLEXUS (C0580081). Because the tree address of *Body Part, Organ, or Organ Component* has as prefix the tree addresses of the other two semantic types, it is a descendant of *Fully Formed Anatomical Structure* and *Anatomical Structure*, similar to Figure 2.2A. Both these sets can be resolved by changing the semantic type of the children (e.g., RIGHT BIG TOE and EXTERNAL RECTAL VENOUS PLEXUS) from

Fully Formed Anatomical Structure to *Body Part, Organ, or Organ Component*—cases of Child-Too-General.

2.3.4 *Body Location or Region and Body Space or Junction* Relationship Sets

One unexplained relationship set has 228 concept pairs in which the semantic type of the parent concepts is *Body Location or Region* (T029, A2.1.5.2) and the semantic type of the child concepts is *Body Space or Junction* (T030, A2.1.5.1). A second relationship set has 261 concept pairs that have the opposite semantic type assignments. For example, RIGHT INGUINAL CANAL (C0459928), with semantic type *Body Space or Junction*, IS-A INGUINAL CANAL (C0021445), with semantic type *Body Location or Region*. Conversely, MIDDLE ETHMOIDAL CELL (C0928857), with semantic type *Body Location or Region*, IS-A SINUS (C0030471), with semantic type *Body Space or Junction*. All the concepts in these two sets should have both semantic types (Parent-Type-Missing and Child-Type-Missing).

2.3.5 *Disease or Syndrome and Pathologic Function* Relationship Set

The previous four categories account for 33 of the 34 large relationship sets shown in Figure 2.3. The remaining relationship set contains 33 concept pairs in which the parent concepts have semantic type *Disease or Syndrome* (T047, B2.2.1.2.1) and the child concepts have the semantic type *Pathologic Function* (T046, B2.2.1.2). For example, INFERTILITY, MALE (C0021364) IS-A INFERTILITY (C0021359). INFERTILITY, MALE and the other 32 children in the set should have their semantic types changed from *Pathologic Function* to *Disease or Syndrome*—cases of Child-Too-General.

2.3.6 Small Unexplained Relationship Sets

The above five categories cover the 34 relationship sets in Figure 2.3 and 25 additional relationship sets (24.0% of the unexplained 246 relationship sets). Together, these sets

cover 16,429 (96.5%) of the concept pairs. The remaining 593 concept pairs are grouped into 187 relationship sets. Table 2.1 shows the results of the analysis of 100 randomly selected concept pairs from this remaining group.

Table 2.1 Reasons for Unexplained Relationship Sets from a Sample of Small Sets

Cause	No. of relationships
Child-Missing-Type	66
Parent-Missing-Type	18
Wrong-IS-A	6
Missing-Ancestor-Descendant-link	4
Child-Too-General	4
Parent-Too-Specific	2

One systematic way to evaluate these sets is to identify those in which the semantic type of the parents *is-descendent-of* the semantic type of the children (as was done for the *Body Part, Organ, or Organ Component, Fully Formed Anatomical Structure and Disease or Syndrome/Pathologic Function* relationship sets described above) to determine if the cause is Parent-Too-Specific or Child-Too-General. Eighteen of the remaining relationship sets, containing 95 concept pairs, meet this criterion. Twelve of the relationship sets, containing 63 concept pairs, are caused by Child-Too-General; for example, all 37 children with semantic type *Spatial Concept* (T082, A2.1.5) should have the semantic type of their parents [*Body Location or Region* (T029, A2.1.5.2) in 29 cases and *Spatial Concept* (T082, A2.1.5) in eight cases].

The remaining six of the above 18 relationship sets, containing 22 concepts, along with a random sample of the final 169 small relationship sets (summarized in Table 2.1), containing 498 concept pairs, were due to a variety of causes, including Parent-Too-Specific, Parent-Type-Missing, Child-Type-Missing, and Wrong-IS-A. Specific counts of each cause

are difficult to produce, however. Ambiguity in the meaning of the semantic types and concepts, as well as the intent of *IS-A* in the SN and IS-A relationship in the META all contribute to this difficulty. Take, for example, the IS-A relationship between ARTERIOVENOUS MALFORMATION (C003857), with semantic type *Congential Abnormality* - (T019, A1.2.2.1), and its child ARTERIOVENOUS FISTULA (C0003855), with semantic type *Anatomical Abnormality* (T190, A1.2.2). Certainly arteriovenous fistulae are malformations of the arteriovenous system; some of them are congenital, but others are not, such as those that are created surgically [8]. But the term "arteriovenous malformation" is also used to refer to a very specific congenital abnormality. The meaning of "arteriovenous malformation" must be known before the cause of this unexplained relationship can be resolved. If both meanings are intended, then the ambiguous concept should be split into two concepts, for example, CONGENTIAL ARTERIOVENOUS MALFORMATION and CONGENITAL OR ACQUIRED ARTERIOVENOUS MALFORMATION. The former would have an IS-A to the latter, and the original IS-A would be preserved as ARTERIOVENOUS FISTULA IS-A CONGENITAL OR ACQUIRED ARTERIOVENOUS MALFORMATION.

2.3.7 Missing-Ancestor-Descendant-Link

The structure of the SN is of particular interest [31, 67]. In fact, this study is undertaken in part to seek evidence that the IS-A relationships in META might support the addition or deletion of *ancestor-descendant* relationships in the SN. In the review of the results of this study, several relationship sets are successfully found. They seem to be due to the cause Missing-Ancestor-Descendant-link. The largest of these, with nine concept pairs, has child concepts with semantic type *Injury or Poisoning* (T037, B2.3) and parent concepts with semantic type *Disease or Syndrome* (T047, B2.2.1.2.1). One example pair is INERT GAS NARCOSIS IS-A OCCUPATIONAL DISEASE. It is believed that the semantic types of both concepts are correct and that the IS-A relationship between them is also correct. The only remaining explanation, then, is the inference that *Injury or Poisoning is-descendant-of*

Disease or Syndrome should be added to the SN (Missing-Ancestor-Descendant). This set of nine concept pairs may seem to be scant supporting evidence; however, an additional 2,186 PARENT-CHILD relationships between concepts of these types can be found in MRREL. Although the relationship type is null, many of these may represent additional IS-A pairs if the relationships types were to be made explicit. As a result, the NLM is suggested to consider the addition of *Injury or Poisoning is-descendant-of Disease or Syndrome* to the SN.

2.4 Discussion

The majority of problems uncovered by the method were incorrect IS-A relationships in the META hierarchy. Correction of such hierarchical errors is an important part of UMLS maintenance, since many users rely on this knowledge for classification purposes. For example, a user who wishes to search for articles about disease of the SKELETAL SYSTEM OF UPPER LIMB, and uses META to help with an "explode" function, may retrieve articles discussing the BONY PELVIC GIRDLE.

The addition of missing semantic type assignments, as well as the removal of incorrect assignments, is also of great importance to users of the UMLS, who depend on such information for understanding how concepts from disparate terminologies are integrated into the META. Consider, for example, a case in which a UMLS user is constructing a list of prostheses from the META. Since there is no semantic type "Prosthesis" in the SN, such concepts are categorized with the semantic types *Medical Device* and *Body Part, Organ and Organ Component.* Thus, a query of META for concepts with both types will miss terms such as HEART, ARTIFICIAL.

The method described in this chapter is intended to provide a way for the UMLS developers to identify quickly one kind of inconsistency in their knowledge sources. It shows that 24.3% of the relationships in MRREL are unexplained. However, since the analysis was restricted to PARENT-CHILD IS-A relationships, this represents only 2.6%

of all parent-child (PAR/CHD) relationships and about 0.3% of all the relationships in MRREL.[3] The application of the method to other kinds of relationships in MRREL will depend upon clarification of the semantics represented by the relationships. The vast majority (555,594 or 84.9%) of the PAR-CHD relationships are not further specified. If one assumes that the default PARENT-CHILD relationship is also IS-A, then the method could be extended to cover a much larger proportion (about 13%) of MRREL.

The method is automated insofar as it identifies unexplained relationships, but it then requires manual review to identify the specific cause for each instance. The results of the manual review suggest several ways in which the method could be extended to further the automated process and reduce the burden of manual review. For example, by knowing that the semantic type of a child concept is the *ancestor-of* the semantic type of the parent concept (as in Figure 2.2A), likely causes can be narrowed down to Parent-Too-Specific and Child-Too-General. Since the semantic typing in the UMLS is supposed to be as specific as possible, the most likely solution in each case will probably be to simply replace the semantic type of the child concept with the (more specific) type of the parent concept. Manual review then only needs to be done to confirm the appropriateness of each type assignment to the child concepts, as shown in Section 2.3. This reduces to four the number of possible causes for the remaining relationships.

Another way to simplify the review of unexplained relationships is to examine the relationship sets to determine if they are evidence for Missing-Ancestor-Descendant-link relationships in the SN. This requires analysis of only the semantic type pairs, not the concept pairs, in the relationship sets. In those sets where an *ancestor-descendant* relationship is missing from the SN, its addition will provide an explanation for all members of the set. In the remaining cases, the possible causes will now be reduced to three.

There may also be a way to automate the detection of Wrong-IS-A. Previous work has shown that some semantic types are mutually exclusive [17]. By considering this

[3]Most relationships in MRREL are reciprocal, so the denominator is about half of the 10,147,419 records.

restriction, users of the method can automatically tell when no amount of addition of semantic types to the parent or child concepts will result in a correct IS-A between them. For example, if *Clinical Drug* concepts are considered as manufactured objects, then such concepts could never be classified as any semantic type in the *Chemical* subtree of the SN.

Thus, addition of other methods may automatically reduce most human review to deciding between Parent-Type-Missing and Child-Type-Missing. In such cases, the missing type is often simply the type of the other concept, making the correction of these unexplained IS-A relationships relatively easy.

It was found that 17,022 IS-A relationships in META are unexplained by the semantic types of the concepts involved. A desirable result would be to extend Figure 2.3 to show the numbers of concept pairs for each of the six causes in each of the 246 relationship sets. Unfortunately, this effort is extremely difficult. Even if each of the concept pairs were manually analyzed, there would be many cases where no resolution is possible without clarification from the NLM (e.g., the pair ARTERIOVENOUS FISTULA IS-A ARTERIOVENOUS MALFORMATION). However, if the NLM were to apply this methods, they might easily resolve many of the results through editorial decisions. For example, if the NLM were to decide that, as a general editorial principle, concepts with the semantic types *Clinical Drug* or *Medical Device* should not have IS-A relationships to concepts with semantic types in the SN's *Chemical* subtree, the causes for 14,086 (82.8%) of the unexplained IS-A relationships would be resolved.

Regardless of whether or not the unexplained relationships can be resolved unequivocally, this method detects those that are inconsistent with respect to the semantic types of the concepts. The review suggests that the majority of these inconsistent IS-A relationships are wrong and should be deleted. Therefore, it is believed that the NLM can improve the UMLS by adding this method to the lexical [6] and semantic [5,32] auditing methods they are already using in order to identify problematic parts of META and SN that are deserving of human review.

2.5 Conclusions

The UMLS contains an enormous body of knowledge about terminologies, and its developers are expending great effort to make it coherent, consistent, and correct. Automated methods can help to focus human review on problem areas. The method easily identifies inconsistencies in one part of the UMLS—the PARENT-CHILD IS-A relationships between concepts in META, as compared to the *ancestor-descendant* relationships between their corresponding semantic types, where almost one quarter are in need of correction. This method, combined with other methods, can be applied using the UMLS developers' editorial authority to effect the necessary corrections.

CHAPTER 3

AUDITING CONCEPT CATEGORIZATIONS IN THE UMLS USING A META SCHEMA OF SN

3.1 Background: Metaschema of the SN

In the SN, the semantic types are nodes and the relationships between them are the links. The semantic types are arranged in a strict hierarchy through hierarchical IS-A links. In addition, the semantic types are related through non-hierarchical semantic relationships. The semantic relationships defined for a semantic type are generally inherited via the IS-A links by all the children of this semantic type, unless the inheritance is explicitly blocked.

The process of generating a metaschema of the Semantic Network begins with partitioning. Since each semantic type has a set of relationships that are either defined for it directly or are inherited from the parent, one can partition the SN based on the distribution of the relationships among the semantic types. All semantic types exhibiting the exact same set of relationships are grouped together [14]. The set of relationships that is shared by all semantic types in a group is the structure of those semantic types and their group. Such a group is called a structural group. Every semantic type is assigned to one and only one structural group. Therefore, all structural groups are pairwise disjoint and their union yields all the semantic types of the SN. The partition of the SN into structural groups is called the structural partition.

However, in the structural partition of the SN, there are cases of structural groups with multiple roots. For an effective partition of the SN, each group should not just be structurally uniform, but also semantically uniform. For this, a group needs to have a unique root, i.e., one semantic type that all other semantic types in the group are descendants of the unique root. In order to obtain semantically uniform groups, rules need to be developed to

transform those structural groups with multiple roots into (cohesive) semantic type groups, each with a unique root. For a detailed explanation, see [54].

Another problem with the structural partition is its large number of leaf singletons. Note that a singleton is a group of one semantic type. A semantic type without children is called a leaf. To avoid it, some rules [54] need to be developed to add a leaf singleton to its parent's structural group. After applying rules to the structural partition, the cohesive partition is obtained. It consists of groups, called semantic type collections, with unique roots. Some of these collections are structural groups; others are semi-structural groups (see [54] for details).

From the cohesive partition, the cohesive metaschema of the UMLS is generated. Each semantic type collection is represented in the metaschema as a node, called a meta-semantic type. The meta-semantic type is named after the unique root of the corresponding semantic-type collection. The meta-semantic types in the metaschema are connected by two kinds of links, the meta-child-of hierarchical relationships and the semantic meta-relationships. The hierarchical meta-child-of relationships are induced from the IS-A relationships in the SN. The meta-relationships are induced from the semantic relationships. The meta-child-of hierarchy in the metaschema supports the inheritance of the meta- relationships among meta-semantic types. The cohesive metaschema of the SN consists of 28 meta-semantic types (see Figure 3.1 for the metaschema hierarchy). It provides an abstract, compact view of the SN.

3.2 Auditing Methodology

In the META, each concept is assigned to one or more semantic types, each of which in turn is associated with one meta-semantic type. For example, the concept RETROVIRUS VECTOR LN is assigned to the three semantic types: *Virus; Pharmacologic Substance; and Indicator, Reagent, or Diagnostic Aid*, which are partitioned in the metaschema into three

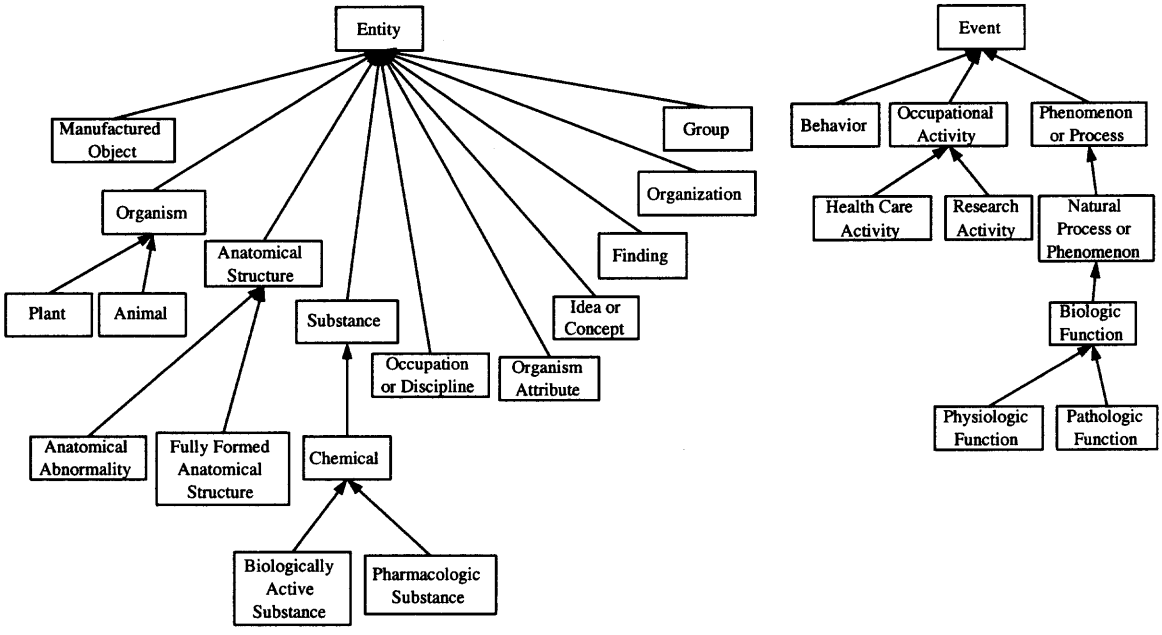


Figure 3.1 The cohesive metaschema hierarchy of the UMLS Semantic Network.

meta-semantic types **Organism, Pharmacologic Substance**[4], and **Chemical**, respectively. Therefore, the concept RETROVIRUS VECTOR LN is associated with those three meta-semantic types.

However, a concept that is assigned to two or more semantic types is not necessarily associated with two or more meta-semantic types since multiple semantic types may be grouped into one meta-semantic type. For example, the concepts PULSUS BIGEMINUS, HYPOXEMIA, DNA MARKER, GENETIC MARKERS, ANOXEMIA, CHROMOSOME MARKERS, and RNA MARKER are assigned to the semantic types *Laboratory or Test Result* and *Sign or Symptom*, which are grouped together into one meta-semantic type **Finding** in the metaschema. Thus, all those seven concepts are associated with only one meta-semantic type **Finding**.

The first hypothesis is that the probability of a concept being erroneously assigned to multiple semantic types from different meta-semantic types is higher than that of being erroneously assigned to multiple semantic types of the same meta-semantic type. The

[4]Bold font will be used for meta-semantic types in this chapter.

reason is that closely related semantic types are grouped together into one meta-semantic type in the metaschema. The chance of a concept being assigned correctly to two closely related semantic types is higher than that being assigned correctly to two semantic types that are not closely related, as is expected for two semantic types of two different meta-semantic types.

This hypothesis leads to the idea of concentrating the auditing effort on concepts that are associated with multiple meta-semantic types, since such concepts may be more error-prone than concepts with a single meta-semantic type. A few definitions are needed to describe the auditing method.

3.2.1 Intersection of Semantic Types

DEFINITION (INTERSECTION OF SEMANTIC TYPES): An intersection of two or more semantic types is a non-empty set of concepts that are assigned to each of these semantic types and only to them.

Figure 3.2 shows the intersection of the semantic types C and D. The concept A is assigned to only two semantic types C and D. So A is in the intersection of C and D, denoted $A \in C \bigcap D$. The notation of an intersection uses the mathematical intersection symbol $\bigcap$. As an example from the UMLS, RETROVIRUS VECTOR LN $\in$ *Virus* $\bigcap$ *Pharmacologic Substance* $\bigcap$ *Indicator, Reagent, or Diagnostic Aid.*

According to the definition of intersection of semantic types, each concept in the META will be in at most one intersection of semantic types. Thus, all intersections of semantic types are disjoint. For example, the concept RETROVIRUS VECTOR LN in the previous example will not be a concept in any one of the following three binary intersections: *Virus* $\bigcap$ *Pharmacologic Substance; Pharmacologic Substance* $\bigcap$ *Indicator, Reagent, or Diagnostic Aid; or Virus* $\bigcap$ *Indicator, Reagent, or Diagnostic Aid.* The reason is that the concept RETROVIRUS VECTOR LN is assigned to three of these semantic types, not two. Thus, it can only be a concept in the intersection of all those three semantic types.

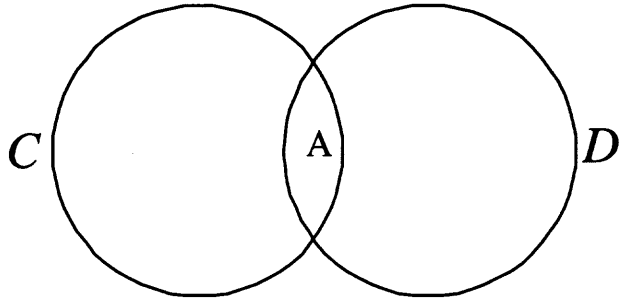


Figure 3.2 Example of the intersection of semantic types.

3.2.2 Meta-semantic Type Association

DEFINITION (META-SEMANTIC TYPE ASSOCIATION): A concept is called associated with a meta-semantic type if it is assigned to at least one of the semantic types in this meta-semantic type.

For example, the semantic types *C* and *D* in Figure 3.2 and the semantic types *E* and *F* are all grouped into one meta-semantic type **A** (see Figure 3.3(a)). As mentioned before, the concept A is assigned to only two semantic types *C* and *D*. The concept B is assigned only to the semantic type *E* and the concept C is assigned only to the semantic type *F*. Thus, all three concepts A, B, and, C are associated with the meta-semantic type **A** (see Figure 3.3(b)).

However, since each concept can be assigned to more than one semantic type, it may also be associated with more than one meta-semantic type if the assigned semantic types are partitioned into different meta-semantic types. For example, the concept ENZYMES is assigned to two semantic types *Organic Chemical* and *Enzyme*. The semantic types *Organic Chemical* and *Enzyme* reside in two meta-semantic types **Chemical** and **Biologically Active Substance**, respectively. Therefore, the concept ENZYMES is associated with those two meta-semantic types.

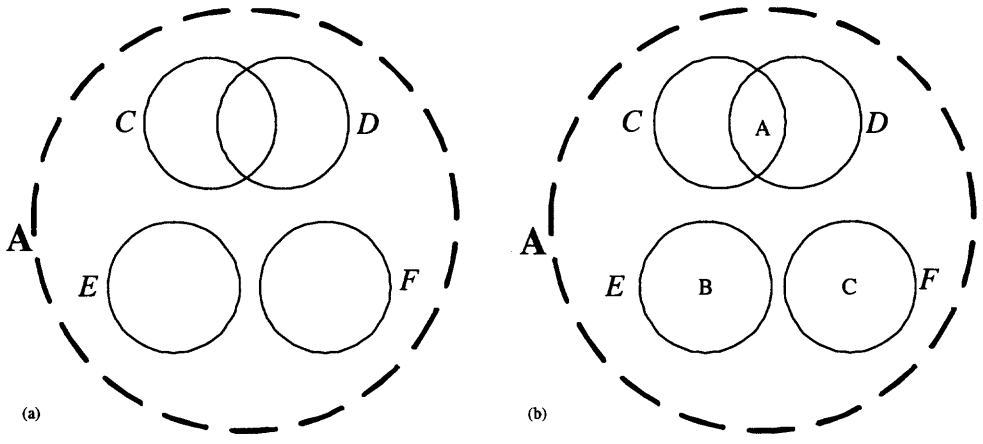


Figure 3.3 Example of meta-semantic type association (semantic types are represented by circles and meta-semantic types are represented by bold dash circles).

3.2.3 Intersection of Meta-semantic Types

DEFINITION (INTERSECTION OF META-SEMANTIC TYPES): An intersection of two or more meta-semantic types is a non-empty set of concepts that are associated with each of these meta-semantic types and only with them.

Figure 3.4 shows the intersection of meta-semantic types **A** and **B**. The concepts U, W, X, Y, and, Z are all associated with the meta-semantic types **A** and **B**. Thus, all of them are in the intersection of the **A** and **B**.

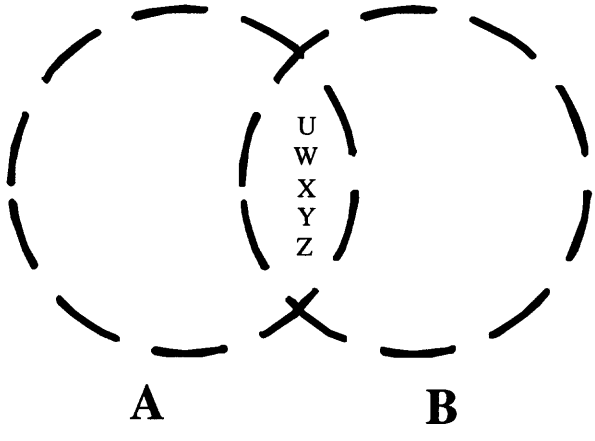


Figure 3.4 Example of the intersection of meta-semantic types.

The same notation is used for the intersection of meta-semantic types. For example, ENZYMES $\in$ **Chemical** $\cap$ **Biologically Active Substance**.

As with the intersection of semantic types, each concept can only be in at most one intersection of meta-semantic types. Therefore, all the intersections of meta-semantic types are disjoint.

A concept in the intersection of meta-semantic types must be in an intersection of semantic types. However, a concept in one intersection of semantic types may not necessarily be in any intersection of meta-semantic types. The reason is that the intersected semantic types may be grouped into the same meta-semantic type. In a previous example, the seven concepts PULSUS BIGEMINUS, HYPOXEMIA, DNA MARKER, GENETIC MARKERS, ANOXEMIA, CHROMOSOME MARKERS, and RNA MARKER are in the intersection of semantic types *Laboratory or Test Result* $\cap$ *Sign or Symptom*. But both semantic types *Laboratory or Test Result* and *Sign or Symptom* are grouped together into one meta-semantic type **Finding** in the metaschema. In this case, all seven concepts are just in the meta-semantic type **Finding**, not in any intersections of meta-semantic types. Therefore, not all intersections of semantic types are intersections of meta-semantic types. Thus, the effort to review the intersections of meta-semantic types should be smaller than the effort of reviewing all the intersections of semantic types.

Note that concepts in one intersection of meta-semantic types are not necessarily in one intersection of semantic types. An intersection of meta-semantic types may consists of several intersections of semantic types. When the domain expert examines the concepts in one such intersection of meta-semantic types, it is difficult for the expert to analyze all those concepts together since they belong to different intersections of semantic types and thus have different compound semantics [29]. This makes the task of reviewing more complicated. In order to make the auditing job simpler and more efficient, each one of the intersections of meta-semantic types needs to be partitioned into multiple pure intersections, defined as follows.

3.2.4 Pure Intersection of Meta-semantic Types

DEFINITION (PURE INTERSECTION OF META-SEMANTIC TYPES): A pure intersection of meta-semantic types is a subset of the intersection of the corresponding meta-semantic types, containing all concepts in one intersection of semantic types.

According to the definition, all pure intersections of one intersection of meta-semantic types are disjoint and their union yields the intersection of the meta-semantic types. In other words, the collection of all pure intersections of an intersection of meta-semantic types is a partition of the intersection of meta-semantic types.

The graphical representation of Figures 3.2–3.4 uses the standard Venn Diagram [25]. However, the graphical representation of the pure intersections is not straightforward as the intersections of semantic types and the intersections of meta-semantic types. The reason is that each pure intersection involves three kinds of entities: meta-semantic types, semantic types, and concepts. From the semantic type point of view, all meta-semantic types should be disjoint since each semantic type is in only one meta-semantic type. But from the concept point of view, some meta-semantic types are not disjoint because some concepts can be associated with multiple meta-semantic types. Therefore, in order to capture all these details, another way is used to represent the pure intersections. Figure 3.5 shows five pure intersections of the intersection of meta-semantic types **A** $\bigcap$ **B** of Figure 3.4. The intersection of meta-semantic types appears as the bold dash oval that is connected to the meta-semantic types with bold lines, while the pure intersections appear as ovals inside the intersection of meta-semantic types and are connected to the semantic types drawn as circles inside their meta-semantic types.

Figure 3.5 shows that the concepts U, W, X, Y, and Z are all in the same intersection of meta-semantic types **A** $\bigcap$ **B**. The meta-semantic type **A** consists of four semantic types *C, D, E,* and *F*, while the meta-semantic type **B** consists of two semantic types, *M* and *N*. However, each of the concepts U, W, X, Y, and Z is in a different intersection of semantic types. The concept U is in the semantic type intersection $C \bigcap D \bigcap M$; the concept W is in

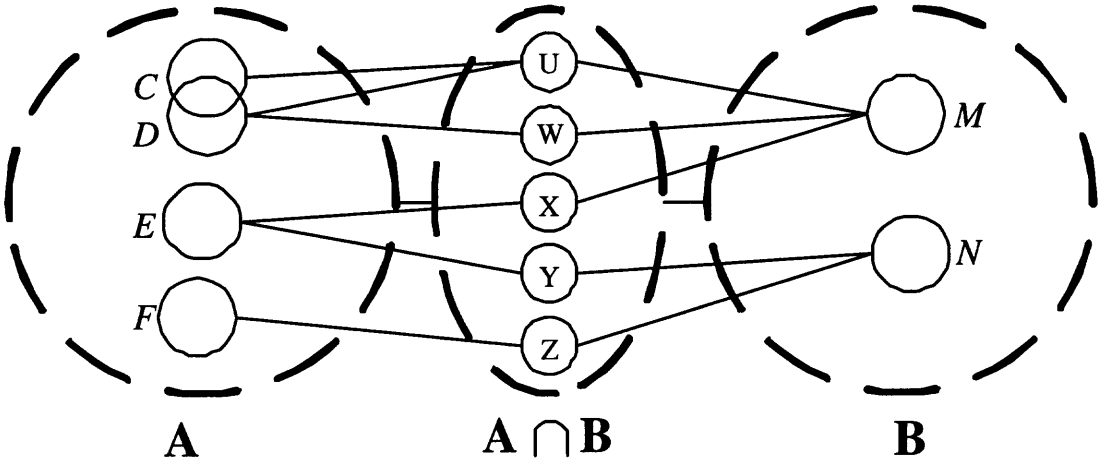


Figure 3.5 Example of pure intersections (the bold dash oval represents the intersection of meta-semantic types and the ovals inside the bold dash oval represent pure intersections).

the semantic type intersection $D \bigcap M$; the concept X is in the semantic type intersection $E \bigcap M$; the concept Y is in the semantic type intersection $E \bigcap N$; and finally the concept Z is in the semantic type intersection $F \bigcap N$.

The notation for a pure intersection is the list of names of each meta-semantic type followed by its corresponding semantic type (or intersection of semantic types) in curly brackets, where the intersection symbol $\bigcap$ appears between any two meta-semantic types in the intersection list. The intersection of meta-semantic types **A** $\bigcap$ **B** in Figure 3.4 is partitioned into five pure intersections in Figure 3.5. They are $\mathbf{A}\{C \bigcap D\} \bigcap \mathbf{B}\{M\}$, $\mathbf{A}\{D\} \bigcap \mathbf{B}\{M\}$, $\mathbf{A}\{E\} \bigcap \mathbf{B}\{M\}$, $\mathbf{A}\{E\} \bigcap \mathbf{B}\{N\}$, and $\mathbf{A}\{F\} \bigcap \mathbf{B}\{N\}$. The union of all five pure intersections is **A** $\bigcap$ **B**.

Consider the following example. In the metaschema, the semantic types *Event, Activity, Daily or Recreation Activity*, and *Machine Activity* are grouped into the meta-semantic type **Event**; while the semantic types *Idea or Concept, Functional Concept, Temporal Concept, Qualitative Concept, Quantitative Concept, Spatial Concept*, and some others are grouped into the meta-semantic type **Idea or Concept**. The concepts STRESSFUL EVENTS, HOUSEHOLD CONSUMPTION, and WITHDRAWING CARE are all in the intersection of meta-semantic types **Event** $\bigcap$ **Idea or Concept**. However, they are in different

pure intersections. The concept STRESSFUL EVENTS is in the pure intersection **Event**{*Event*} ⋂ **Idea or Concept** {*Qualitative Concept*}; HOUSEHOLD CONSUMPTION is in **Event** {*Activity*} ⋂ **Idea or Concept**{*Quantitative Concept*}; and WITHDRAWING CARE is in **Event**{*Activity*} ⋂ **Idea or Concept**{*Idea or Concept*}.

After the pure intersections of the meta-semantic types are generated, the domain expert can now review all the concepts in one pure intersection, which is easier since they all have the same compound semantics as expressed by the associated semantic types [29].

An effective auditing process should expose many errors with limited efforts. With this in mind, the auditor's review is concentrated on the pure intersections of meta-semantic types containing very few concepts. The second hypothesis is that the likelihood of a mistake for a small pure intersection is higher than in the case of a large pure intersection. The reason is that if a combination of semantic types makes sense semantically, then there would probably be quite a few, or at least several, concepts associated with it. For example, the pure intersection **Chemical**{*Organic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} is a reasonable combination, since many drugs are composed of organic chemicals. This pure intersection contains the largest number of concepts (70,436) among all pure intersections. On the other hand, the case where a pure intersection contains only one or two concepts may indicate an erroneous categorization, where no concepts should be associated with such a combination.

The process of the auditing is as follows. First, all intersections of meta-semantic types of the metaschema are identified. All those intersections are refined to generate the pure intersections. Now, a "divide and conquer" approach is applied in order to limit the number of concepts reviewed by the domain expert while at the same time covering concepts with high likelihood of wrong categorizations. Hence, this auditing technique minimizes the effort while trying to maximize the impact of the audit. On one side, a domain expert reviews concepts of each pure intersection containing a relatively small number of concepts. The total number of concepts reviewed is limited due to the low cardi-

nality of the pure intersections considered. On the other hand, a domain expert reviews the semantic soundness of the intersected semantic types of all medium and large-size pure intersections, looking for combinations of semantic types that are not semantically sound. There is one review per pure intersection independent of the number of concepts assigned to this intersection. Only for those unlikely pure intersections, the expert will review their concepts independent of their number. This way the number of concepts reviewed is limited due to the small number of semantically unsound pure intersections, and their likelihood to have erroneous categorization is high due to their unsound compound semantics.

3.3 Results

First, all intersections of meta-semantic types, which contain a total of 170,179 concepts, are identified in the auditing process. Then, each one of them was partitioned into pure intersections to create 874 pure intersections. Table 3.1 describes the distribution of the number of concepts for all pure intersections. Reviewing Table 3.1, one finds that most of the pure intersections are small sets of one or two concepts. For example, there are 332 pure intersections containing only one concept and 113 pure intersections containing only two concepts. On the other extreme, the pure intersection that contains the largest set of concepts has 70,436 concepts. The average number and median number of concepts for a pure intersection are 195 and 2, respectively. Note that the median is small due to the large number of very small pure intersections. On the other hand, the weighted median number of concepts is 27,002 due to the size of the two largest pure intersections.

3.3.1 Analysis of Small Pure Intersections

A high percentage of incorrect categorizations are found by examining all pure intersections containing one to ten concepts (covering a total of 657 pure intersections and 1680 concepts). They can be divided into four categories: (1) polysemy, (2) inconsistency, (3)

Table 3.1 Distribution of Number of Concepts for Pure Intersections

No. of concepts	No. of pure intersections	No. of concepts	No. of pure intersections	No. of concepts	No. of pure intersections	No. of concepts	No. of pure intersections
1	332	39	1	113	1	440	1
2	113	40	3	116	1	453	1
3	64	42	1	118	1	466	1
4	35	43	2	119	1	484	1
5	28	47	2	120	1	522	1
6	25	48	2	122	1	534	1
7	18	49	2	125	1	541	1
8	17	50	1	127	2	543	1
9	17	51	2	128	1	549	1
10	8	52	1	130	1	568	1
11	9	53	2	131	1	587	1
12	8	54	1	135	1	603	1
13	3	55	2	142	1	648	1
14	12	56	1	148	1	649	1
15	4	57	2	150	1	678	1
16	5	60	1	154	1	688	1
17	6	62	1	161	1	787	1
18	4	67	1	169	1	815	1
19	3	68	1	176	1	880	1
20	6	69	1	185	1	883	1
21	1	70	1	197	1	1096	1
22	3	74	1	213	1	1187	1
23	4	76	1	230	1	1219	1
24	3	77	1	234	1	1290	1
26	2	80	1	242	1	1460	1
28	1	85	1	247	1	2339	1
29	1	87	1	279	1	3074	1
30	3	88	1	287	1	3126	1
32	2	93	2	296	1	4937	1
33	1	96	1	304	1	8061	1
35	2	98	2	328	1	10407	1
36	2	106	1	339	1	27002	1
37	1	107	1	341	1	70436	1
38	2	111	1	354	1		

miscategorization, and (4) redundant categorization. The redundant categorizations were discussed in [52]. The other three categories are discussed in this section.

Some examples of those three kinds of incorrect categorizations are shown in Table 3.2. The first indication of a polysemy error is an intersection of semantic types that is not semantically sound. For example, the concept TALIPES CAVUS is the only concept of the pure intersection **Anatomical Abnormality**{*Congenital Abnormality* $\bigcap$ *Acquired Abnormality*} $\bigcap$ **Finding**{*Sign or Symptom*}. How can a congenital abnormality be an acquired abnormality at the same time? These two semantic types are mutually exclusive siblings in the SN. In order to disambiguate a polysemous concept, it can be replaced with several new concepts according to the different intersecting semantic types, and let each of the new concepts be associated with only one of the semantic types. In the above case, one possible solution is to create two alternative concepts, TALIPES CAVUS <1>[5] that belongs to the pure intersection **Anatomical Abnormality**{*Congenital Abnormality*} $\bigcap$ **Finding**{*Sign or Symptom*}, and TALIPES CAVUS <2> that belongs to the pure intersection **Anatomical Abnormality**{*Acquired Abnormality*} $\bigcap$ **Finding**{*Sign or Symptom*}. Another possible solution, instead of creating two concepts for TALIPES CAVUS, is to re-categorize this concept with the parent semantic type *Anatomical Abnormality* of the two semantic types currently assigned to it. This is consistent with the representation of this concept in the source terminologies of the UMLS.

Similarly, the concept TOXICODENDROM (POISON IVY) has been assigned to the pure intersection **Plant**{*Plant*} $\bigcap$ **Pathologic Function**{*Disease or Syndrome*}. This is also a polysemy error. The same concept is used for the plant and for the disease caused by the plant. To resolve this polysemy, the current version of the UMLS contains two concepts: TOXICODENDROM (POISON IVY) <1> that is a plant, and TOXICODENDROM

[5]Following the UMLS notation, the different meanings of a polysemous concept are denoted by <1> and <2>

Table 3.2 Examples of Various Types of Incorrect Categorizations

Concepts	Pure intersections
Polysemy:	
Talipes Cavus	**Abnormality**{*Congenital Abnormality* ∩ *Acquired Abnormality*} ∩ **Finding**{*Sign or Symptom*}
Toxicodendrom(Poison Ivy)	**Plant**{*Plant*} ∩ **Pathologic**{*Disease or Syndrome*}
Inconsistency:	
Mussels Prawns Scallop, NOS	**Animal**{*Invertebrate*} ∩ **Substance**{*Food*}
Thirsty Physical Exhaustion	**Physiologic Function**{*Physiologic Function*} ∩ **Finding**{*Sign and Symptom*}
Miscategorization:	
Cytarabine	**Chemical**{*Nucleic Acid, Nucleoside, or Nucleotide* ∩ *Biomedical or Dental Material*}∩ **Pharmacologic Substance**{*Pharmacologic Substance*}
Marine Tents Marine Algaes	**Plant**{*Alga*} ∩ **Pharmacologic Substance**{*Pharmacologic Substance*}
Diphtheria-Tetanus Vaccine	**Health Care Activity**{*Therapeutic or Preventive Procedure*} ∩ **Natural Phenomenon or Process** {*Natural Phenomenon or Process*}
Biliscan Glucose Random	**Chemical**{*Chemical*} ∩ **Substance**{*Body Substance*}
Support Hospital-Patient Relations Facility-Patient Relations	**Behavior**{*Social Behavior*} ∩ **Health Care Activity**{*Health Care Activity*}
Polysemy+Inconsistency:	
Adulthood Old-Age	**Idea or Concept**{*Temporal Concept*} ∩ **Group**{*Age Group*}
Polysemy+Miscategorization:	
Rhogam Screen Total Body Clearance Rate Specific Gravity Measurement Oxygen Measurement, Partial Pressure, Arterial	**Finding**{*Lab or Test Result*} ∩ **Health Care Activity**{*Laboratory Procedure*}

(POISON IVY) <2> that is a disease caused by the plant. Both above intersections then disappear with the change in the categorization of the polysemous concept.

Table 3.2 shows some examples of inconsistent categorization. The concepts MUSSELS; SCALLOP, NOS; and PRAWNS are the only three concepts of the pure intersection **Animal**{*Invertebrate*} ⋂ **Substance**{*Food*}. However, they are not the only invertebrates that are food. Many others, e.g., SHRIMP, LOBSTER, and OCTOPUS are food as well. But they are assigned only to *Invertebrate*, not *Food*. Thus, if those three concepts are categorized as food, some other invertebrates such as shrimp, lobster, and octopus should also be categorized as food. This is an inconsistent categorization case. Note that in this case, one semantic type (e.g., *Invertebrate*) represents a sort type while the other (e.g., *Food*) represents a role type. Another example occurs with the concepts THIRSTY and PHYSICAL EXHAUSTION that are the only two concepts of the pure intersection **Physiologic Function**{*Physiologic Function*} ⋂ **Finding**{*Sign and Symptom*}. As in the previous example, other concepts, e.g., STARVATION and DEHYDRATION should have also been in this pure intersection. However, they are categorized only as *Sign and Symptom*, not *Physiologic Function*. Some examples of miscategorization are also listed in Table 3.2.

Case 1. All concepts in one pure intersection are categorized incorrectly.

Case 1.1. The concepts should be categorized only to some of the intersecting semantic types, not all of them.

For example, the concept CYTARABINE is the only concept in the pure intersection **Chemical**{*Nucleic Acid, Nucleoside, or Nucleotide* ⋂ *Biomedical or Dental Material*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}. However, CYTARABINE is a nucleoside analog. It is a pharmacologic substance and potentially a hazardous substance. It should be assigned to neither the semantic type *Nucleic Acid, Nucleoside, or Nucleotide* nor the semantic type *Biomedical or Dental Material*. Thus, this pure intersection will not exist after this miscategorization is resolved. In another example, the pure intersection **Pharmacologic Substance**{*Pharmacologic Substance*} ⋂ **Plant**{*Alga*} contains

two concepts LAMINARIA TENTS and MARINE ALGAES. However, LAMINARIA TENTS is a pharmacologic substance produced from a type of marine algae. It should not be assigned to the semantic type *Alga*, while the concept MARINE ALGAES should be assigned only to the semantic type *Alga*. Hence, after correcting the errors, there will be no such pure intersection.

Case 1.2. All concepts should not been categorized as any one of the intersecting semantic types.

For example, the pure intersection **Health Care Activity**{*Therapeutic or Preventive Procedure*} $\cap$ **Natural Phenomenon or Process**{*Natural Phenomenon or Process*} contains only one concept DIPHTHERIA-TETANUS VACCINE. However, a vaccine is a pharmacologic substance and immunologic factor. It is neither a procedure nor a process. So it should not be assigned to any one of those two semantic types. This intersection becomes empty.

In another example, the concepts BILISCAN and GLUCOSE RANDOM are the only two concepts of the pure intersection **Substance**{*Body Substance*} $\cap$ **Chemical**{*Chemical*}. However, the concepts BILISCAN and GLUCOSE RANDOM are neither body substances nor chemicals. They are just laboratory procedures. Thus, both concepts should not be assigned to any one of those semantic types. Again, the intersection becomes empty.

Case 2. Some of the concepts in the pure intersection are categorized incorrectly.

For example, the concepts SUPPORT, HOSPITAL-PATIENT RELATIONS, and FACILITY-PATIENT RELATIONS are the only concepts in the pure intersection **Behavior**{*Social Behavior*} $\cap$ **Health Care Activity**{*Health Care Activity*}. The concepts HOSPITAL-PATIENT RELATIONS and FACILITY-PATIENT RELATIONS are categorized correctly. However, the concept SUPPORT is not necessarily related to health care. Thus, it should not be assigned to *Health Care Activity*.

However, not all concepts in one pure intersection always demonstrate the same kind of errors. Sometimes, different kinds of errors are found for various concepts in the same pure intersection.

Mixed Case 1. Polysemy + inconsistency

For example, consider the concepts ADULTHOOD and OLD-AGE, which are the only two concepts of the pure intersection **Idea or Concept**{*Temporal Concept*} $\bigcap$ **Group**{*Age group*}. This is an inconsistent categorization since many other concepts, e.g., the concepts CHILDHOOD, JUVENILE, and YOUNG ADULTS should have also been in this pure intersection. However, they are assigned only to *Age Group*. Also, this is a polysemy error because each of those two concepts refers to two different concepts, one is the state of age, and the other is the group of people in that state. To disambiguate these concepts, the concept ADULTHOOD is replaced by ADULTHOOD <1>, a temporal concept, and ADULTHOOD <2>, an age group. Similarly, OLD-AGE is replaced by OLD-AGE <1>, a temporal concept, and OLD-AGE <2>, an age group. Again, the intersection becomes empty, while the polysemy and inconsistency are resolved.

Mixed Case 2. Polysemy + miscategorization

The example in this case occurs with the pure intersection **Health Care Activity**{*Laboratory Procedure*} $\bigcap$ **Finding**{*Lab or Test Result*}, which contains four concepts: TOTAL BODY CLEARANCE RATE; SPECIFIC GRAVITY MEASUREMENT; OXYGEN MEASUREMENT, PARTIAL PRESSURE, ARTERIAL; and RHOGAM SCREEN. However, the concept TOTAL BODY CLEARANCE RATE is found to be polysemous because it refers to two concepts, one is a laboratory procedure and the other is its result. To disambiguate this polysemous concept, two concepts have to be created. One is assigned to *Laboratory Procedure* and the other is assigned to *Lab or Test Result*. The concepts SPECIFIC GRAVITY MEASUREMENT; OXYGEN MEASUREMENT, PARTIAL PRESSURE, ARTERIAL; and RHOGAM SCREEN are laboratory procedures and should not be assigned to the semantic type *Lab or Test Result*. Therefore, the intersection will become empty.

Table 3.3 Analysis of Errors in Small Pure Intersections

No. of concept	No.of pure intersections	No. of pure intersections with errors	Percentage of intersections with errors	Total No. of concepts	No. of erroneous concepts	Percentage of erroreous concepts
1	332	120	34	332	120	34
2	113	56	50	226	105	46
3	64	27	42	192	75	39
4	35	13	37	140	43	31
5	29	13	45	145	55	48
6	25	9	36	150	44	29
7	18	2	11	126	13	10
8	17	3	18	136	11	8
9	17	4	23	153	13	8
10	8	0	0	80	0	0
Total	658	247	38	1680	479	29

Table 3.3 lists the results of the analysis for pure intersections containing one to ten concepts. This table presents the error percentages for both erroneous pure intersections (i.e., with some incorrectly categorized concepts) and incorrectly categorized concepts. The percentage of erroneous pure intersections and the percentage of incorrect categorizations are quite high for the pure intersections containing small numbers of concepts up to the intersections containing six concepts. It decreases when the size increases above six concepts. For all the pure intersections with up to ten concepts, 38% contain erroneous categorizations and 29% of the concepts have incorrect categorizations.

3.3.2 Analysis of Large Pure Intersections

The domain expert will not review concepts of large and medium-sized pure intersections. The domain expert will just check the semantic soundness of medium to large-size pure intersections. Analysis of the concepts is limited only to the pure intersections judged semantically suspicious. The domain experts reviewed the pure intersections containing more than ten concepts. There are 217 of them. Almost all are semantically sound. For example, the pure intersection **Chemical**{*Organical Chemical*} $\cap$ **Pharmacologic Substance**{*Pharmacologic Substance*}, which contains the largest number of concepts

(70,436), is a reasonable combination since many drugs are also organic chemicals. The same is true of the pure intersection **Chemical**{*Amino Acid, Peptide or Protein*} ⋂ **Biologically Active Substance**{*Enzyme*} that contains 27,002 concepts.

Table 3.4 lists the 16 largest pure intersections that are associated with the metasemantic type **Chemical** and the 16 largest pure intersections that are not associated with it. The 16th pure intersection associated with **Chemical** is an interesting case. As a matter of fact, it is a case of redundant categorization [52]. All these 883 concepts should not be categorized as **Organic Chemical**. After removing this redundant categorization, those 883 concepts should join the 9th pure intersection **Chemical**{*Carbohydrate*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} in the left column of Table 3.4. A similar situation exists for the 10th pure intersection **Chemical**{*Organic Chemical* ⋂ *Amino Acid, Peptide, or Protein*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} in the left column of Table 3.4. The concepts actually belong to the second largest pure intersection in this column. All others are semantically sound.

A few of the pure intersections are semantically suspicious. For example, **Manufactured Object**{*Manufactured Object*} ⋂ **Organization**{*Organization*} contains 70 concepts. However, no concept can be a manufactured object as well an organization simultaneously. Basically, the semantic types *Manufactured Object* and *Organization* are mutually exclusive. Therefore, **Manufactured Object**{*Manufactured Object*} ⋂ **Organization**{*Organization*} is semantically suspicious and probably should not exist. All 70 concepts were reviewed and found polysemous. For example, DAY CARE CENTERS FOR CHILDREN is in this pure intersection. However, it refers to two concepts. One is an organization and the other is a manufactured object that includes buildings and facilities in day-care centers. All other concepts in this pure intersection such as PRIMARY SCHOOLS, LABORATORIES, INFORMATION CENTER, etc., have the same polysemy error. To disambiguate these polysemous concepts, two concepts are created for each polysemous concept. The original one is assigned to the semantic type *Organization* and the other with the word "building" added

Table 3.4 Largest Pure Intersections and Their Cardinalities

Pure intersection
Chemical{*Organic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (70,436)
Chemical{*Amino Acid, Peptide, or Protein*} ⋂ **Biologically Active Substance**{*Enzyme*} (27,002)
Chemical{*Amino Acid, Peptide, or Protein*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (10,407)
Chemical{*Amino Acid, Peptide, or Protein*} ⋂ **Biologically Active Substance**{*Biologically Active Substance*} (8061)
Chemical{*Amino Acid, Peptide, or Protein*} ⋂ **Biologically Active Substance**{*Immunologic Factor*} (4937)
Chemical{*Amino Acid, Peptide, or Protein*} ⋂ **Biologically Active Substance**{*Receptor*}(4299)
Chemical{*Nucleic Acid, Nucleoside, or Nucleotide*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (3126)
Chemical{*Steroid*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (3074)
Chemical{*Carbohydrate*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (2339)
Chemical{*Organic Chemical* ⋂ *Amino Acid, Peptide, or Protein*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (2132)
Chemical{*Lipid*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (1460)
Chemical{*Organic Chemical*} ⋂ **Pharmacologic Substance**{*Antibiotic*} (1364)
Chemical{*Inorganic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (1290)
Chemical{*Amino Acid, Peptide, or Protein*} ⋂ **Pharmacologic Substance** {*Pharmacologic Substance*} ⋂ **Biologically Active Substance**{*Immunologic Factor*} (1219)
Chemical{*Organophosphorus Compound*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (1187)
Chemical{*Organic Chemical* ⋂ *Carbohydrate*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} (883)
Anatomical Abnormality{*Congenital Abnormality*} ⋂ **Pathologic Function**{*Disease or Syndrome*} (1,096)
Anatomical Abnormality{*Acquired Abnormality*} ⋂ **Pathologic Function**{*Disease or Syndrome*} (815)
Finding{*Finding*} ⋂ **Physiologic Function**{*Organ or Tissue Function*} (787)
Anatomical Abnormality{*Anatomical Abnormality*} ⋂ **Pathologic Function**{*Disease or Syndrome*} (603)
Health Care Activity{*Therapeutic or Preventive Procedure*} ⋂ **Occupational Activity**{*Educational Activity*} (549)
Finding{*Finding*} ⋂ **Pathologic Function**{*Pathologic Function*} (534)
Finding{*Finding*} ⋂ **Pathologic Function**{*Disease or Syndrome*} (339)
Pathologic Function{*Disease or Syndrome*} ⋂ **Finding**{*Sign or Symptom*} (328)
Event{*Daily or Recreational Activity*} ⋂ **Health Care Activity**{*Therapeutic or Preventive Procedure*} (243)
Health Care Activity{*Health Care Activity*} ⋂ **Occupational Activity**{*Educational Activity*}(197)
Manufactured Object{*Manufactured Object*} ⋂ **Entity**{*Intellectual Product*} (176)
Pathologic Function{*Pathologic Function*} ⋂ **Finding**{*Sign or Symptom*} (161)
Organism Attribute{*Organism Attribute*} ⋂ **Finding**{*Finding*} (135)
Plant{*Plant*} ⋂ **Substance**{*Food*} (125)
Health Care Activity{*Diagnostic Procedure*} ⋂ **Entity**{*Intellectual Product*} (111)
Physiologic Function{*Cell Function*} ⋂ **Pathologic Function**{*Cell or Molecular Dysfunction*} (106)

is assigned to *Manufactured Object*. After disambiguating these polysemous concepts, the original pure intersection will not contain any concepts and should not exist. Similarly, since the semantic types *Inorganic Chemical* and *Organic Chemical* are mutually exclusive, the pure intersection **Chemical**{*Organic Chemical* ⋂ *Inorganic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*} should not exist. The 247 concepts in this pure intersection were reviewed. All of them should not be categorized as inorganic chemical. Thus, they are in the largest pure intersection **Chemical**{*Organic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}.

Among the 217 pure intersections containing more than ten concepts, only six medium-sized are judged semantically suspicious. After all 405 concepts in these six semantically suspicious pure intersections have been reviewed, only two pure intersections should

exist. One is **Physiologic Function**{*Organ or Tissue Function*} ∩ **Pathologic Function**-{*Pathologic Function*}. Despite the suspicious semantic type combination, all twelve concepts in it are categorized correctly. For example, the concepts MESIAL MOVEMENT OF TEETH, SKIN WRINKLING, and OSTEOLYSIS are organ or tissue functions, but they are pathologic functions as well. The other semantically suspicious pure intersection that should exist is **Chemical**{*Amino Acid,Peptide, or Protein* ∩ *Element, Ion or Isotope*} ∩ **Biologically Active Substance** {*Immunologic Factor*} ∩ **Pharmacologic Substance** {*Pharmacologic Substance*}. For example, the concepts IODINE I 131 MONOCLONAL ANTIBODY 3F8, IODINE I 131 MONOCLONAL ANTIBODY ANTI-B1, and IODINE I 131 MONOCLONAL ANTIBODY G-250 are in this pure intersection. It is true that an antibody cannot be an element, ion or isotope. However, each concept in this pure intersection is not an antibody produced naturally but rather an antibody engineered for therapeutic purposes, coupled with a radioactive substance in order to selectively target the tissue to which the antibody is directed. Therefore, each such concept is categorized as both *Immunologic Factor* and *Element, ion, or isotope*.

However, all other 377 concepts in the other four semantically suspicious pure intersections are erroneously assigned to some semantic types. Table 3.5 lists semantically suspicious pure intersections, the number of concepts categorized to them, and the pure intersections to which the concepts should belong.

Out of the 405 concepts in the six pure intersections reviewed, 377 concepts, about 93%, have erroneous categorizations. Note that these reviews are much easier than those of the small pure intersections, since all of the concepts of a large or medium-size pure intersection typically share the same semantics and have the same categorization error. Hence, the method of auditing medium to large pure intersections is an example of a successful auditing process: finding many errors with a limited review effort.

Table 3.5 Semantically Suspicious Medium-size Pure Intersections

Semantically suspicious pure intersections	No. of concepts	Correct semantic types of pure intersections
Physiologic Function{*Organ or Tissue Function*}⋂**Pathologic Function**{*Pathologic Function*}	12	Same as original
Chemical{*Amino Acid, Peptide, or Protein*⋂*Element, Ion or Isotope*} ⋂ **Biologically Active Substance** {*Immunologic Factor*}⋂**Pharmacologic Substance**{*Pharmacologic Substance*}	16	Same as original
Chemical{*Inorganic Chemical* ⋂ *Organic Chemical* ⋂ *Amino Acid, Peptide, or Protein*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}	20	**Chemical**{*Organic Chemical* ⋂ *Amino Acid, Peptide, or Protein*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}
Chemical{*Organic Chemical* ⋂ *Element, Ion or Isotope*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}	40	**Chemical**{*Organic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}
Manufactured Object {*Manufactured Object*}⋂ **Organization**{*Organization*}	70	**Organization**{*Organization*}
Chemical{*Organic Chemical* ⋂ *Inorganic Chemical*} ⋂ **Pharmacologic Substance** {*Pharmacologic Substance*}	247	**Chemical**{*Organic Chemical*} ⋂ **Pharmacologic Substance**{*Pharmacologic Substance*}

3.4 Discussion

The task of checking the correctness of all the concepts, and their related data, in a large terminology is overwhelming. Usually, there are not enough resources for such a task. Furthermore, the tendency of terminology designers is to invest most of the available resources in extending the terminology.

However, the accuracy of a terminology is critical for its mission in overcoming terminological differences among various health care information systems. Thus, auditing techniques for terminologies, similar to auditing techniques in other fields, e.g., finance, are designed in an effort to expose as many errors as possible with a limited effort.

The auditing technique is designed in the same approach by checking only a limited number of concepts such that their probability to be erroneous is high. The technique

is based on the two hypotheses. The first hypothesis is that the probability of a concept being incorrectly assigned to a pure intersection is higher than that of being incorrectly assigned to an intersection of semantic types that are in one meta-semantic type. In order to confirm this hypothesis, 100 concepts are checked. These concepts are in the intersections, containing one to six concepts of semantic types, from the same meta-semantic type. The error percentage is about 20% versus about 40% in the method. This result confirms that the first approach using intersections of meta-semantic types is an effective auditing approach.

The approach also differentiates between the treatment of small intersections and large intersections. This approach is based on the second hypothesis that the probability of incorrect categorizations is high for small pure intersections. The results in Table 3.3 confirm the second hypothesis. The percentage of erroneous categorizations for pure intersections of up to six concepts, about 40%, is high. The percentage decreases for medium-size pure intersections containing seven to ten concepts, and is further reduced for large ones where most pure intersections are judged as semantically sound. This observation confirms the second hypothesis and shows that concentrating on the concept-based analysis of small pure intersections is justified.

These results support auditing as a divide and conquer technique applying different processing to small intersections and large intersections.

An intersection of two semantic types assigned to two different meta-semantic types does not automatically imply that there is an error in the categorization of the concepts of this intersection. The interdisciplinary nature of medicine implies that medical knowledge is also interdisciplinary. Thus, it is quite natural for a concept to be categorized as several semantic types. Such semantic types that are assigned the same concepts may or may not be closely related and thus may or may not be in the same meta-semantic type.

Thus, it does not imply that all concepts in pure intersections are erroneously categorized. Even for the small pure intersections, 60% of the concepts are properly categorized. Actually, the fact that many concepts are assigned to a specific combination of semantic

types, i.e., to their intersection, supports the idea that this combination is semantically sound in spite of the fact that the intersecting semantic types are assigned to different meta-semantic types.

Table 3.4 lists the 16 largest pure intersections of semantic types that are descendants of *Chemical* and 16 largest pure intersections that are not descendants of *Chemical*. The reason for this distinction is the dominance of the first kind among the largest pure intersections. The soundness of the combination of the semantic types in these pure intersection is straightforward. The two exceptions are the 10th and 16th pure intersections in the left column, which are redundant categorization cases.

As reported in Table 3.5, only four pure intersections of medium size out of the six suspicious ones are actually found semantically unsound, and the categorizations of their concepts need some modification.

As a matter of fact, one can apply the auditing approach for a partition of the SN rather than of a metaschema. While every metaschema is based on a partition, not every partition is appropriate for the construction of a metaschema. For example, in the partition called "semantic partition" [7, 44] not all 15 groups consist of a connected subtree of the SN, a necessary condition for constructing a metaschema.

Hence, applying the auditing technique to a partition rather than a metaschema can broaden its usefulness. Only 5 out of the 32 pure intersections of Table 3.4 would be pure intersections when applying the technique to the partition of [7, 44] instead of the metaschema. All five are in the right column of Table 3.4. Hence, part of the expert work of checking the semantic soundness of the pure intersections is saved when using the semantic partition. On the other hand, if the semantic partition is used, then some of the erroneous small pure intersections would not be detected. For example, the errors in the categorizations of the concepts TALIPES CAVUS and CYTARABINE (see Table 3.2) as well as three of the four semantically unsound medium-sized pure intersections (see Table 3.5) would not be detected.

There seems to be a trade off between the recall and the precision. It is interesting to note that one of the principles underlying the semantic partition is exclusivity, which minimizes the number of concepts associated with different groups. The exclusivity and proximity qualities, coupled with the flexibility regarding the connectivity of the groups, enable to avoid detecting many of the large pure intersections that are actually semantically sound. The cohesive metaschema does not share these qualities. Using it in the audit generates the large pure intersections that are semantically sound. On the other hand, the cohesive metaschema helps us uncover many erroneous small pure intersections that would be missed if the semantic partition were to be used instead.

Finally, in the design of a metaschema and its underlying partition, one can choose various granularities, resulting in different numbers of meta-semantic types. The choice of granularity seems to influence the trade off between recall and precision. The emphasis on recall is more important. One reason is that the effort for checking the soundness of a pure intersection is independent of the size of the intersection and is easier than checking the categorizations of a concept since it is done with the broad categorizations of semantic types.

For some errors exposed in the auditing, the actual error was not due to categorization of a concept to two semantic types in different meta-semantic types but to two semantic types in the same meta-semantic type. Examples of exclusive pairs of semantic types are (*Congenital Abnormality; Acquired Abnormality*) and (*Organic Chemical; Inorganic Chemical*). Each of the two examples is a pair of siblings, where a concept should not be assigned to both, due to their semantic incompatibility. However, a pair involving *Inorganic Chemical* and any descendant of *Organic Chemical* will also be an exclusive pair. The audit technique does not consider intersections of exclusive pairs of semantic types. But this is a natural potential extension to complement the technique. One could enumerate all exclusive pairs of semantic types and check their intersections. For every concept assigned to two exclusive semantic types, one should consider whether it is just a categorization error

or a case of polysemous categorization. In the later case, one can create two concepts with different meanings, one for each semantic type, or re-categorize the polysemous concept to the parent semantic type of the two exclusive sibling semantic types to preserve consistency with the source terminology. The later case was demonstrated earlier for TALIPES CAVUS.

Three kinds of incorrectly categorized concepts are identified: polysemy, inconsistency, and miscategorization. For some concepts, there is a combination of various kinds of errors. Typically, errors of the first kind stem from polysemy in the terminology used by health care workers in verbal communication. Humans overcome such polysemy cases due to the context in which concepts are used. However, a concept entry in the META should be unambiguous. In some cases of inconsistency, one semantic type represents a sort type while the other represents a role type. The sort type categorization seems to be the consistent one while the role type appears only in some of the cases. See, for example, the intersection of *Invertebrate* and *Food* in Table 3.2. For such cases of inconsistent categorization, a decision is needed whether to add the missing categorization to all other qualified concepts or to remove the extra categorization for the concepts that had it. Either way will lead to a consistent categorization.

3.5 Conclusion

The UMLS integrated many biomedical terminologies. During the integration, each concept was assigned to at least one semantic type. However, due to the size and complexity of the UMLS, it is unavoidable that some incorrect associations have been generated. To find and correct such incorrect associations, the notion of intersection semantic types was introduced in [29]. The more complex concepts, those with compound semantics [29], are associated with intersection semantic types. These are concepts that are likely to have errors in their modeling or categorization. Hence, the review of these concepts will provide effective auditing. However, the number of such concepts is quite large and only a small

sample of them were reviewed in [29] to provide a proof of concept. The comprehensive review of all such concepts is an overwhelming task.

An effective auditing technique is designed to review a substantial portion of the concepts of intersection semantic types, those which are more likely to have erroneous categorizations. For this purpose, an efficient auditing technique has been developed based on the pure intersections of meta-semantic types of the metaschema. The divide and conquer approach treats small and large pure intersections differently. The review of the concepts of small pure intersections led to the recognition of different kinds of incorrect assignments. The results of analysis for the pure intersections containing between one to ten concepts were presented. On the other hand, the combinations of all pure intersections containing more than ten concepts were reviewed to check their semantic soundness. The list of semantically suspicious pure intersections containing more than ten concepts was presented and all their concepts were reviewed. The results confirm the two hypotheses, which were the basis for the auditing technique.

Due to the divide and conquer approach, only a limited number of concepts were actually reviewed. A meaningful portion of them were found to have erroneous categorizations. Hence, the technique provides an effective auditing method; domain experts do not review intersections of semantic types associated with the same meta-semantic type. More errors are expected there, but their likelihood is lower than in this chapter.

CHAPTER 4

AUDITING AS PART OF THE TERMINOLOGY DESIGN LIFE CYCLE

4.1 Background: NCI Thesaurus

The National Cancer Institute Thesaurus (NCIT) was designed in response to a need for a consistent, shared vocabulary for the various projects and initiatives at the NCI, as well as in the broader cancer research community. The NCIT covers clinical and basic research as well as administrative terminology.

The NCIT's design is based on description logic. It has a tool for automatic classification couched in this model. The NCIT has defined and inferred versions. The defined version is the one containing the assertions made about each concept by the editors. The inferred version includes in addition assertions and tree placements inherited during DL classification. In this chapter, the inferred version of the NCIT is used to do the analysis.

The data model of the NCIT uses four basic elements: concepts, kinds, roles, and properties [9]. The foundational unit of information is the concept. The NCIT contains 42,404 concepts organized in 21 disjoint hierarchies, covering different subject areas such as Biological Process, Experimental Organism Diagnoses, Genes, Gene Products, and Property or Attribute. Each hierarchy consists of IS-A relationships between child and parent concepts forming a directed acyclic graph (DAG). The largest hierarchy, Diseases Disorders and Findings, contains 9,613 concepts. Roles describe semantic (non IS-A) relationships between concepts and are inherited by a child concept from a parent concept along the IS-A hierarchy. For example, the concept MALIGNANT BREAST NEOPLASM has the role *located in*[6], connecting it to the concept BREAST. Since the concept BREAST DUCTAL CARCINOMA IS-A *Malignant Breast Neoplasm*, it inherits the *located in* role to the concept BREAST.

[6]Role names will be italicized and start with a lowercase letter.

Each concept is associated with exactly one of 21 disjoint sets called *kinds*, representing major subdivisions in the NCIT, e.g., Biological Process Kind and Gene Kind. Properties are used to describe a concept, examples include: definition, preferred name, synonyms, and semantic type.

4.2 Methods

4.2.1 Dividing a Terminology into Areas

The terminology dividing methodology is based on the notions of *area*, *structure* [54], and *root*, defined as follows.

Definition 1 (Area): An area is a group of all concepts that have exactly the same roles. □

Definition 2 (Structure): The structure of a concept (and an area) is the set of its roles. □

Hence, An area of a terminology is structurally uniform.

Definition 3 (Root of an Area): A concept X in an area A is called a root of A if no parents of X are in A. □

A terminology is divided by its areas, meaning that each concept belongs to one and only one area according to its structure. An area is named by listing its roles inside braces. The area with no roles is named $\emptyset$. Figure 4.1(a) shows the division of a sample abstract terminology into four areas. Concepts B and C are grouped into the area $\{r_1\}$ since both of them have only the role r_1. Similarly, D, E, and F are in the area $\{r_1, r_2\}$ and G, H, I, and J are in the area $\{r_3\}$. Concept A is in the area $\emptyset$ because it has no roles at all. The symbols "*" and "+" will be explained later. In Figure 4.1(a), D and E are the two roots of $\{r_1, r_2\}$. Concepts A and B are the roots of $\emptyset$ and $\{r_1\}$. G and J are the roots of $\{r_3\}$.

All the descendant concepts along a path down from an area's root until a concept where a new role is introduced share the same structure and thus belong to the same area. This is due to the inheritance of roles along the IS-A hierarchy which enables the structural division into areas. In case inheritance may be interrupted by blocking mechanism as

is the case of the UMLS Semantic Network [43], the division into groups of identical relationships is more problematic [53].

For example, in the NCIT, 38 concepts are grouped into the area {*has associated location*} in the Biological Process hierarchy. The concepts CELLULAR PROCESS, VIRUS-CELL MEMBRANE INTERACTION, BLOOD CIRCULATION, DIGESTION, URINATION, RESPIRATION, NEUROLOGIC PROCESS, UTERINE SWELLING, and ANGIOGENIC INHIBITION are the nine roots of {*has associated location*}. The rest of the area's concepts are descendants of one of these roots. For example, the root CELLULAR PROCESS has 18 descendants, such as CELL AGING and CELL VIABILITY PROCESS.

The division based on areas lends itself to a useful kind of abstraction diagram called *area taxonomy* (*AT*). The AT serves as a compact representation of the terminology and provides a high-level view of the distribution of its roles. Specifically, the AT is a DAG of area nodes (Figure 4.1(b)). A node is labeled with the list of roles exhibited by all its constituent concepts. A node can be connected to another node in the AT via a *child-of* relationship. These *child-of* relationships serve as abstractions of the underlying IS-As in the terminology. One area node X is *child-of* a another area node Y if the root of X has an IS-A relationship to some concept in Y. Note that the concept in Y need not be a root. The area $\{r_1\}$ is a *child-of* the area $\emptyset$ in the AT (drawn as a bold arrow in Figure 4.1) because the root of $\{r_1\}$, concept B, IS-A concept A in $\emptyset$. For example, in the NCIT's Biological Process hierarchy, {*has associated location, has initiator process*} is a *child-of* {*has associated location*}, because the root of the first area, APOPTOSIS, IS-A CELL DEATH, which resides in the second area.

Definition 4 (Introducing concept): A concept at which one or more new roles are introduced into the terminology is called an *introducing concept*. □

For example, the concept APOPTOSIS is an introducing concept for the role *has initiator process*. In every area, the root is, by definition, an introducing concept because it introduces one or more new roles. As a matter of fact, the reason why this concept is

actually a root of its area—and is not in its parent's area—is that it introduces new roles. Other roles may be inherited.

For example, in the area {*has associated location, has initiator process*}, the role *has initiator process* is introduced at APOPTOSIS, which is the root. Thus, this area is considered an introduction point for this role. Role-introduction points are highlighted by placing a "*" next to the name of any role introduced at a root of the particular area. Hence, the above *child-of* relationship is actually from {*has associated location, has initiator process**} to {*has associated location**}, which in turn is a child of ∅.

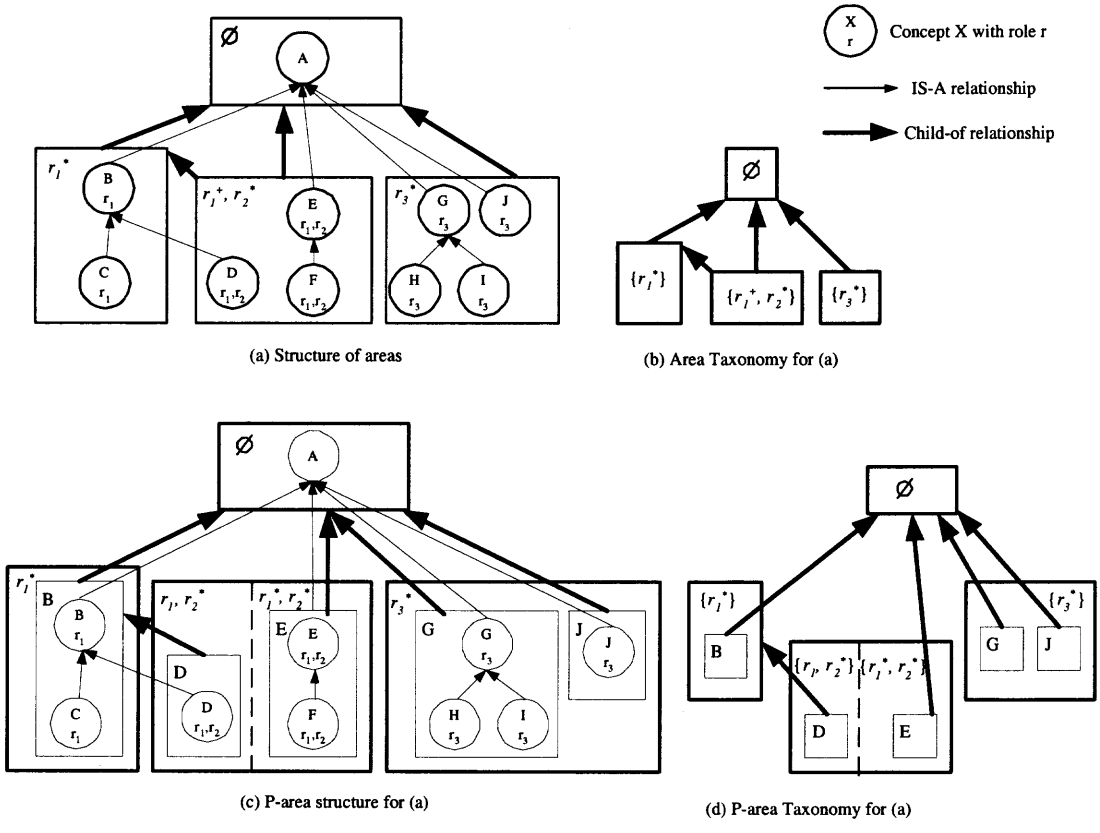


Figure 4.1 Examples of an area taxonomy and p-area taxonomy.

Some areas have several introduction patterns for the same structure. In Figure 4.1(a), the role r_1 in $\{r_1, r_2\}$ is inherited by root D but introduced by root E. In such a case of varying introduction patterns for a role, the role is marked with "+" instead of "*" in the area name. See, e.g., $\{r_1^+, r_2^*\}$ in Figure 4.1(a). In the area {*has result biological*

*process**, *is part of process*$^{+}$}, the first role is introduced at all its roots, but the second is introduced by some and inherited by others.

4.2.2 Dividing an Area into P-Areas

An area of a terminology is by definition structurally uniform. However, an area might not be semantically uniform in the sense of having a unique root concept that generalizes all its descendant concepts in the area. A unique root can convey the semantics of the whole group. For example, the unique root PATHOGENESIS of {*has initiator process**, *has result process**} containing 15 concepts conveys the general semantics of all concepts in the area. When a role is allowed to be introduced at multiple points in the IS-A hierarchy, as it is in the NCIT, then an area may have multiple roots. The area {*has associated location**} has among its concepts a group of 19 rooted at CELLULAR PROCESS and another group of eight rooted at NEUROLOGIC PROCESS. Obviously, each of these two groups is semantically uniform, but the area is not uniform. Therefore, areas are further divided into concept groups, called *partial areas (p-areas)*, which are both structurally uniform and singly-rooted. A p-area is named after its unique root concept since the root generalizes all the concepts of the p-area.

Definition 5 (P-area): A p-area in an area A is a group of concepts that contains only one root X and all descendants of X in A. □

As in the area taxonomy, a "*" is used to indicate the p-area where the role is introduced. The lack of a "*" means the role is inherited. The areas in Figure 4.1(a) can be further divide into six p-areas according to the roots A, B, D, E, G, and J (see Figure 4.1(c)). In the previous example, the area {*has associated location**} is further divided into nine p-areas because it has nine roots: CELLULAR PROCESS(19), VIRUS-CELL MEMBRANE INTERACTION(5), BLOOD CIRCULATION(1), DIGESTION(1), URINATION(1), RESPIRATION(1), NEUROLOGIC PROCESS(8), UTERINE SWELLING(1) and ANGIOGENIC INHIBITION(1)[7]. The number in parentheses represents the number

[7]P-areas name will be in "small caps" font and follows by a number.

of concepts in the respective p-area, including the root. Each of these nine p-areas has a uniform semantics captured by its name.

The division of areas into p-areas leads to an expanded, two-level AT called as the *p-area taxonomy* (*PAT*). The PAT, similar to the AT, is a DAG, with p-areas represented as nodes and connected to other p-areas via *child-of* relationships. To capture the additional level of division, p-areas are grouped into areas of the AT. In a PAT diagram, p-areas are drawn as boxes inside their respective area boxes (Figure 4.1(d)). A p-area box is labeled with the name of its root concept, which conveys the essence (semantics) of the group. Recall that all p-areas within a given area exhibit the exact same roles, so role names do not distinguish p-areas as they do for areas. The PAT offers a view that provides both relationship distribution information across the entire terminology and further hierarchical grouping information within areas.

Note that the "+" in the AT is disambiguated in the PAT. Each root of a p-area has a distinct introduction pattern. Areas with a "+" in their names are divided into several parts of a specific introduction pattern, separated from one another in the PAT diagram by a dashed line. Each of these parts will include the p-areas of the corresponding introduction pattern. Figure 4.1(b) shows that the area $\{r_1^+, r_2^*\}$ is separated (by the dashed line) into two parts: $\{r_1, r_2^*\}$ for the p-area D and $\{r_1^*, r_2^*\}$ for the p-area E. In the NCIT, $\{$*has associated location*$^+$, *is part of process*$^+\}$ is separated into three different parts: $\{$*has associated location*, *is part of process*$^*\}$ with 19 p-areas; $\{$*has associated location**, *is part of process*$\}$ with 22 p-areas; and $\{$*has associated location**, *is part of process*$^*\}$ with two p-areas, whose roots introduce both roles since their parents are in $\emptyset$ (having no roles).

An advantage of the AT and PAT is in providing groupings of similar concepts into small collections. Furthermore, the taxonomies guide the auditor to concentrate on groups of concepts with higher likelihood of errors, as discussed below. Sometimes an indented hierarchy is also used to display all concepts in a p-area to aid in the auditing process. After

all, auditing must involve the review of actual concepts. As a result, the auditor can move among three levels of display as necessary to support the auditing process.

4.2.3 Auditing Methodology

The auditing methodology is based on a "divide and conquer" approach, where one first divides the terminology into areas and then further divides the areas into p-areas, as described above. Then the conquer phase utilizes these p-areas to expose errors, otherwise buried undetected in the large complex knowledge base. The AT and PAT are typically smaller than the original terminology's concept network. These compact views of the terminology allow the terminologies' designers, developers, and auditors to see it in a new light, different from the view used in its design. These views can help in the orientation to and navigation of the terminology needed in the auditing process. Looking at the knowledge from this angle, where concepts are grouped according to their structure and an associated root, can help in exposing problems undetected in the design process.

In particular, the PAT serves as the basis of the auditing methodology. In the first part of the manual auditing phase, one utilizes the notion of "concept similarity" to help in identifying omissions and misplacements of concepts. Two concepts are *structurally similar* if they share the same set of roles and are thus in the same area. Furthermore, two concepts in the same area are called *semantically similar* if they share an ancestor in the area. Hence, they would also have the same root and be in the same p-area. Therefore, each p-area in the PAT contains concepts of similar structure and semantics. If one finds that two concepts, similar in their essence, are in different p-areas in the PAT, (or, worse, in different areas), there may be some inconsistencies or errors for some of these concepts. For example, in the NCIT, INHALING and RESPIRATION (see Figures 4.4 and 4.5) are similar in essence, but are in different areas. Also, if a concept similar in its essence to concepts of a p-area is missing from that p-area, either being in a different p-area or not in the terminology at all, this may indicate an unjustified absence. For example, the concept

EXHALING, related to INHALING, is missing from the NCIT. It is easier and more effective for an auditor to detect irregularities when reviewing relatively small areas and p-areas of similar concepts, due to the limited capacity of human comprehension and memory.

Due to the limited resources available for auditing and the desire to optimize their impact, the methodology is intended to check a limited number of concepts whose probability of being erroneous is high. For comparison, in quality assurance (QA) of new software systems, the QA professional's task is to expose errors in the new system. Experienced QA professionals know how to pick subsystems with high likelihood of errors and focus their in-depth checking efforts. Similarly, the techniques are designed to use automated means to identify groups of concepts with high likelihoods of errors, where the manual review is to be concentrated.

The second part of the audit phase follows this approach and focuses on "small" p-areas, having very few concepts. Previous experience [26, 28, 29] suggests that whenever there are small groups of "similar" concepts, there is a high likelihood that these groups represent irregularities that are manifestations of errors more severe than omissions and misplacements of concepts. The reason for this is as follows. If a p-area exists due to its legitimate structure and semantics, then there would probably be quite a few, or at least several concepts, in it. For example, in the NCIT, the p-area SUBCELLULAR PROCESS(87) (see Figure 4.4) contains the largest number of concepts. It is a legitimate p-area since many biological processes are at the subcellular level. On the other hand, a p-area containing only one or two concepts may indicate an error where no concepts at all should be grouped in the particular manner. For example, in the p-areas INHALING(1) and EJACULATION(1) (see Figures 4.4 and 4.5), the concepts are missing a role and therefore end up in erroneous p-areas by themselves. In the auditing methodology, special attention must be paid to all concepts in the PAT's small p-areas.

The following two measures with respect to areas and p-areas are needed. In particular, ways to denote the number of p-areas within areas and concepts within p-areas are required.

Definition 6 (Cardinality): The cardinality of an area is its number of p-areas. □

Definition 7 (Size): The size of a p-area (area) is its number of concepts. □

Note that the size of p-areas is ignored when one defines the cardinality of an area.

Definition 8 ($\bar{k}$-p-area): A $\bar{k}$-p-area is a p-area of size k or less. □

Note that the overline $\bar{k}$ is used to indicate integers and differentiate them from p(partial). Example: A $\bar{3}$-p-area is a p-area that has 1, 2, or 3 concepts.

Definition 9 ($\bar{m}$-area): An $\bar{m}$-area is an area of cardinality m or less. □

Definition 10 ($\bar{m}$-$\bar{k}$-area): An $\bar{m}$-$\bar{k}$-area is an $\bar{m}$-area that consists of $\bar{k}$-p-areas only. □

In later sections, $\bar{3}$-$\bar{3}$-areas and $\bar{5}$-$\bar{3}$-areas are used to test the following two hypotheses.

Hypothesis 1: The probability of erroneous concepts is higher for $\bar{k}$-p-areas with small k than for $\bar{k}$-p-areas with large k.

Hypothesis 2: The likelihood of errors in concepts of a $\bar{k}$-p-area with small k is higher in an $\bar{m}$-area with low m than in an $\bar{m}$-area of high m.

In Hypothesis 2, one further differentiates between small p-areas in areas of high cardinality and low cardinality. In the first case, there are many concepts sharing the same structure and being hierarchically independent of one another, which is a likely configuration. An example of such an area is {*has associated location*$^+$, *is part of process*$^+$} (see Figure 4.3), which has 43 p-areas, 33 of which have only one or two concepts. Only one error was discovered in the 124 concepts of this area.

In the second case, there is one or very few hierarchically related concepts with a unique combination of roles. The rare occurrence of the structure of this p-area may indicate an error. Consider, for example, the single concept TRANSCRIPTION INITIATION in its p-area (see Figure 4.3). This p-area is the only one in {*has associated location*, *has result biological process**, *has result chemical or drug**, *is part of process*}. As a matter of fact, the role with the target TRANSCRIPTION should be *is part of process* instead of *has result biological process* (as is the case in the new release of the NCIT). After this change,

this p-area will belong to {*has associated location*$^{+}$, *has result chemical or drug**, *is part of process*}, which already has nine p-areas (see Figure 4.3).

Following the two hypotheses, the auditing methodology concentrates the typically limited time available for an expert's manual review on the p-areas with a relatively high likelihood of errors. To test these two hypotheses, an extensive audit of one of the NCIT's hierarchies is conducted, including all its p-areas, small and large.

4.3 Results

4.3.1 AT and PAT for a NCIT Hierarchy

The Biological Process hierarchy of the NCIT has been chosen to demonstrate both the AT and PAT. Figure 4.2 shows the AT organized by levels according to the number of roles. There are seven roles, numbered from zero to six, which can be defined for the concepts of the Biological Process hierarchy. The levels of the AT are labeled 0 to 5 according to the number of roles in each. The 589 concepts in the Biological Process hierarchy are grouped into 37 areas (see Figure 4.2). For example, 38 concepts are grouped into {*has associated location**}. For each area, the cardinality (i.e., its number of p-areas) is listed.

Figure 4.3 shows the PAT for the Biological Process hierarchy. Due to the lack of space, the figure does not show the p-areas of some areas. Those p-areas are shown in Figure 4.4 in a subhierarchy of the PAT used later as part of the auditing demonstration. The numbers in parentheses are the total numbers of concepts in the respective p-areas. The previously mentioned area {*has associated location**} (see Figure 4.4) is further divided into nine p-areas: CELLULAR PROCESS(19), VIRUS-CELL MEMBRANE INTERACTION(5), BLOOD CIRCULATION(1), DIGESTION(1), URINATION(1), RESPIRATION(1), NEUROLOGIC PROCESS(8), UTERINE SWELLING(1) and ANGIOGENIC INHIBITION(1).

Figure 4.5 shows ORGANISMAL PROCESS and all its 40 descendants. Figure 4.5(a) displays them in an indented hierarchy format, provided by the NCIT interface [51], and

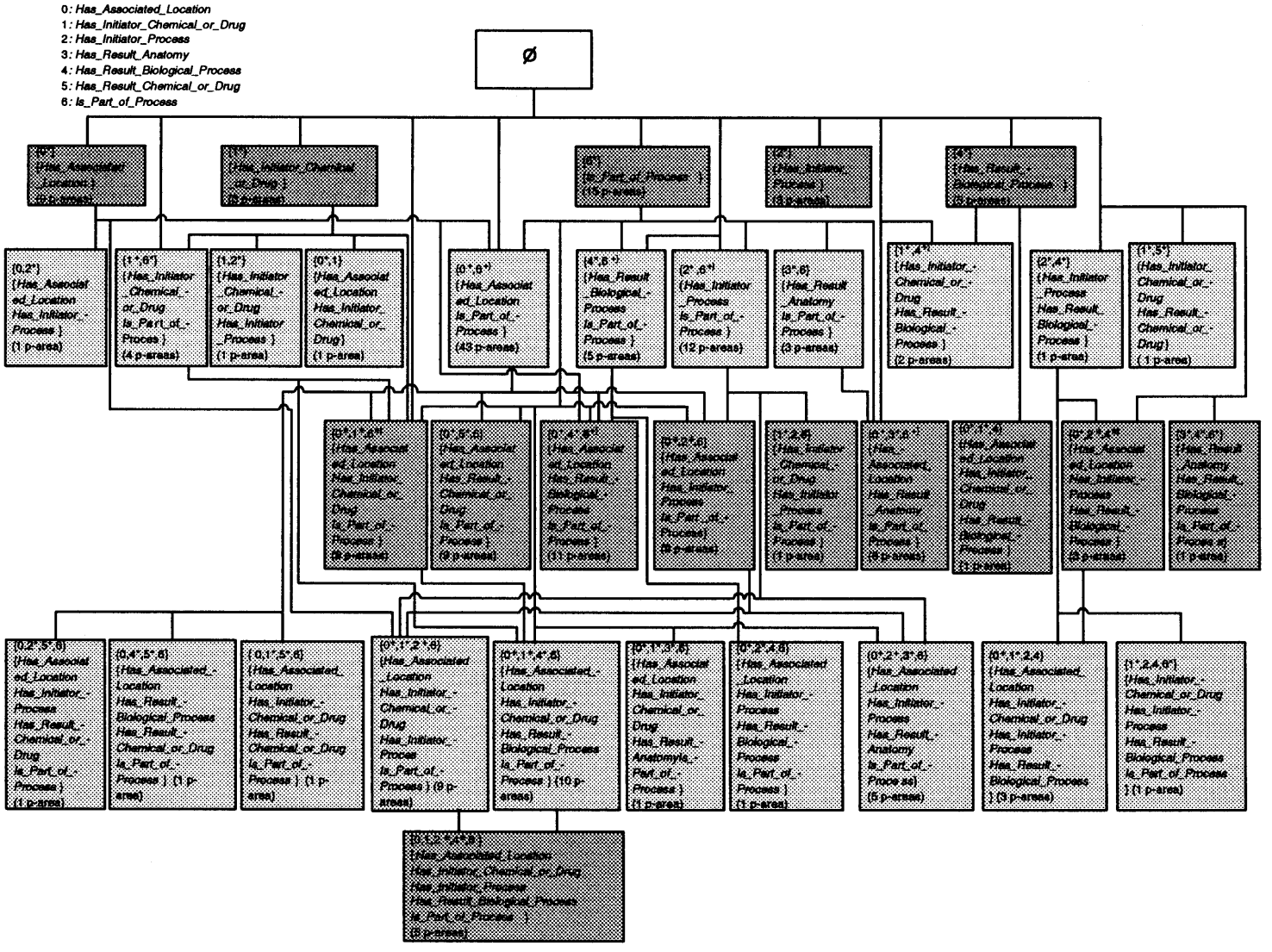


Figure 4.2 AT for the Biological Process hierarchy.

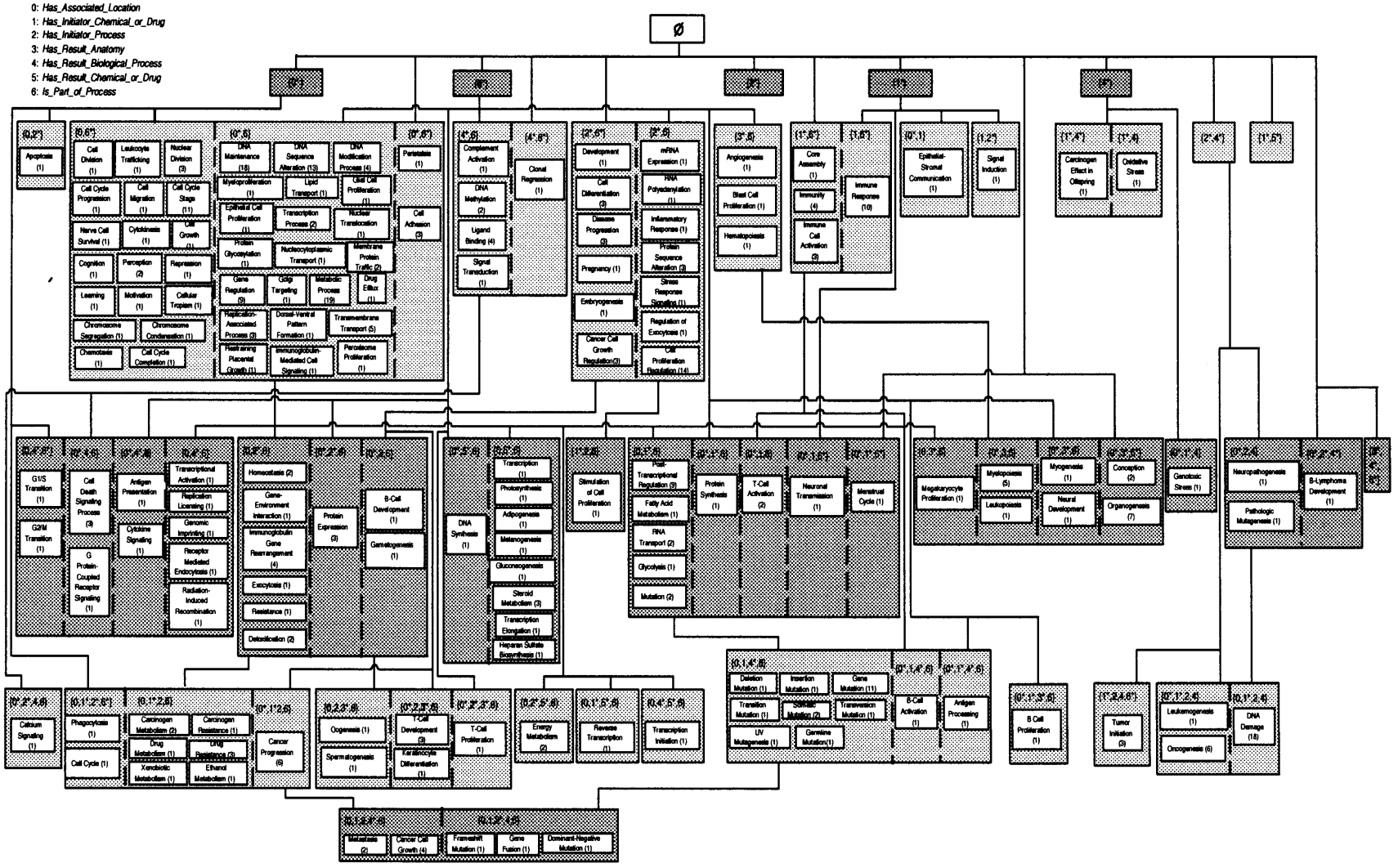


Figure 4.3 PAT for the Biological Process hierarchy.

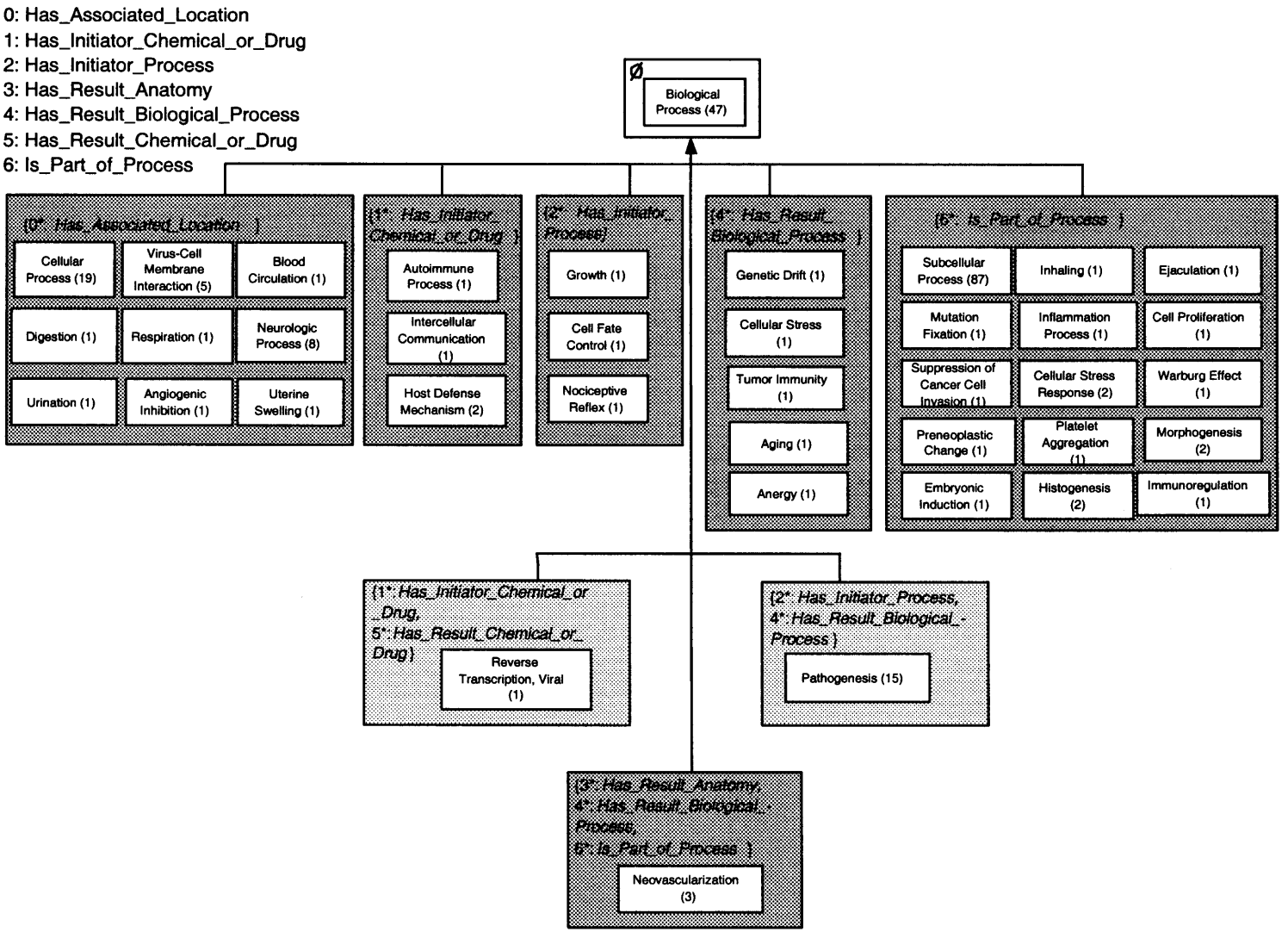


Figure 4.4 A portion of the PAT for the Biological Process hierarchy.

Figure 4.5(b) displays them as a collection of 19 p-areas grouped into nine areas. In the next section, this figure will be utilized to demonstrate different kinds of errors that have been found with the use of the auditing methodology. In the root area $\emptyset$ of the PAT, the "..." denotes the fact that only concepts that are descendants of ORGANISMAL PROCESS, not all the concepts, are listed. Note also that the other areas in Figure 4.5 may be incomplete since some of their concepts are not descendants of ORGANISMAL PROCESS.

Various fonts are used in Figure 4.5 to highlight concepts that are different from the rest of the concepts in the same p-area. The same font is also used to highlight groups of concepts similar in essence but in different areas, e.g., INHALING and RESPIRATION. These differences and similarities in fonts will help to highlight errors described in the next subsection.

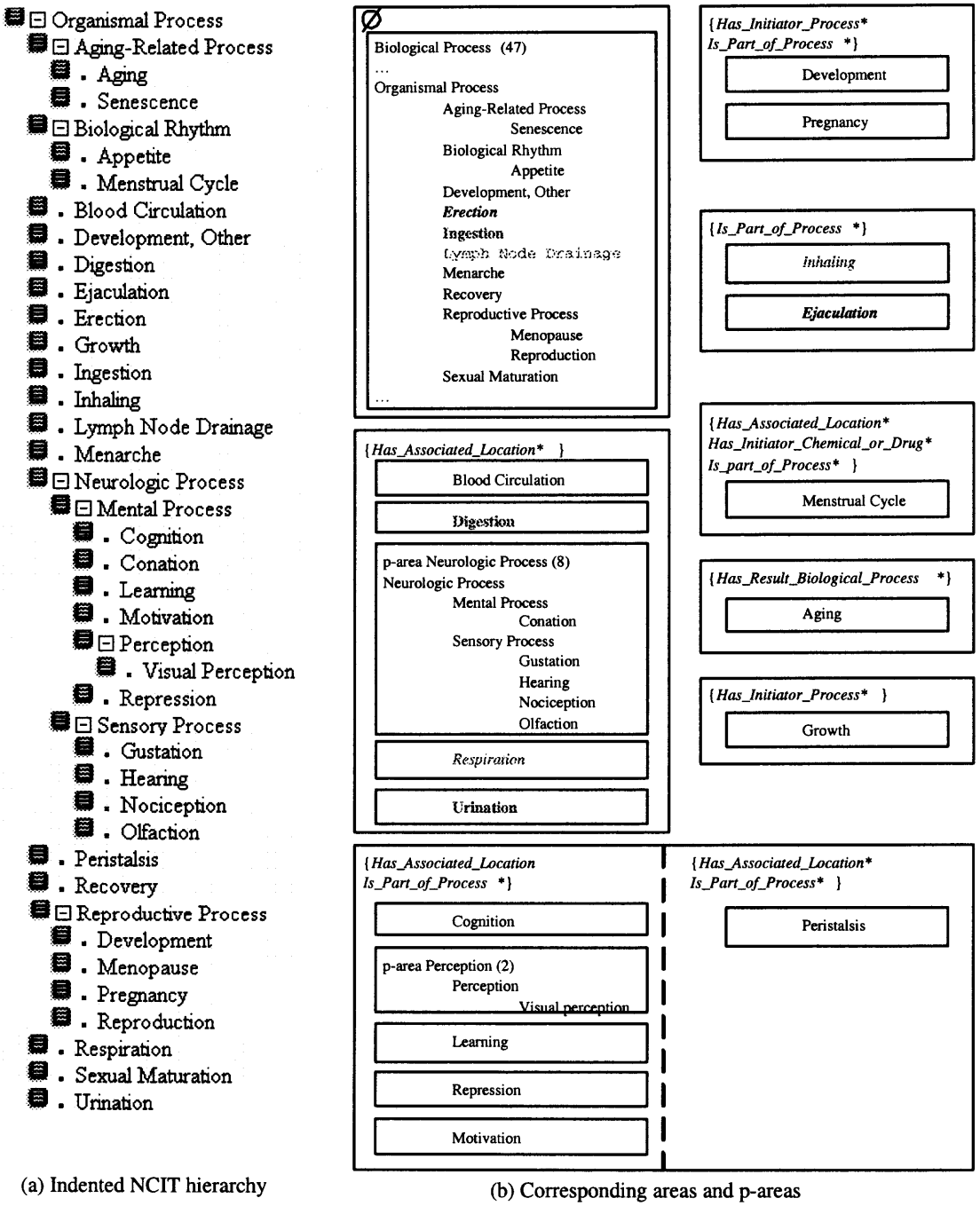
4.3.2 Errors Found in P-areas

Various kinds of modeling errors exposed by small groups of similar concepts, represented by the AT and PAT, will be demonstrated. As mentioned in the previous section, it is easier for auditors to find missing concepts, missing roles, or erroneous concepts by comparing groups of structurally and semantically similar concepts. Furthermore, auditors can easily find inconsistencies among concepts if concepts, similar in their essence, are not in the same area or p-area. If for one concept a role exists while for a similar concept it does not, this may suggest that the latter is missing that role.

Much of the demonstration is directed at the p-areas and concepts in Figure 4.5. However, in order to test the hypotheses, all p-areas, large and small, of all areas have been reviewed. The cases of $\bar{3}$-$\bar{3}$-areas are emphasized explicitly.

Missing Roles:

From the PAT (Figure 4.4), it is found that INHALING(1) in {*is part of process**} contains only one concept INHALING. The same is true for RESPIRATION(1) in {*has associated location**} (see Figure 4.5(b) highlighted with italics). RESPIRATION has the



(a) Indented NCIT hierarchy

(b) Corresponding areas and p-areas

Figure 4.5 Descendants of ORGANISMAL PROCESS in (a) NCIT hierarchy indented format, and (b) as selected areas and p-areas in the diagram format.

role *has associated location* to LUNG. These two related concepts INHALING and RESPIRATION are in different areas. As noted, this may indicate some inconsistency or error. Inhaling is a part of the process of respiration. However, INHALING is missing the role *has associated location* to LUNG, which is the target of this role for RESPIRATION. INHALING will have two roles after this new role is added to it. Since its parent ORGANISMAL PROCESS does not have any roles, both roles should have been introduced at INHALING. The concept INHALING should thus be moved from its original area to {*is part of process**, *has associated location**}.

Another p-area in {*is part of process**} contains only one concept **EJACULATION** which is part of REPRODUCTION. But **EJACULATION** is also missing the role *has associated location*. After moving these two concepts to {*is part of process**, *has associated location**}, the area {*is part of process**} in Figure 4.5(b) becomes empty and does not appear in the revised Figure 4.6, reflecting the changes. It should be noted that this area will still exist in the AT and the PAT due to other p-areas.

From the PAT, it is found that seven concepts in four $\bar{3}$-p-areas, CELLULAR STRESS(1) in {*has result biological process**}, CELLULAR STRESS RESPONSE(2) in {*is part of process**}, CANCER CELL GROWTH REGULATION(3) in {*has initiator biological process*, is part of process**}, and OXIDATIVE STRESS(1) in {*has initiator chemical or drug*, has result biological process*}, are missing the role *has associated location* with the value CELL. Two concepts, CANCER CELL GROWTH and CELLULAR INFILTRATION, in $\emptyset$ have the same kind of mistake.

Missing Synonyms:

The above concept INHALING does not have any synonyms. However, inhaling is part of the concept RESPIRATION in {*has associated location**}. Thus INSPIRATION, obviously referring to the same part of respiration, should be a legitimate synonym of INHALING.

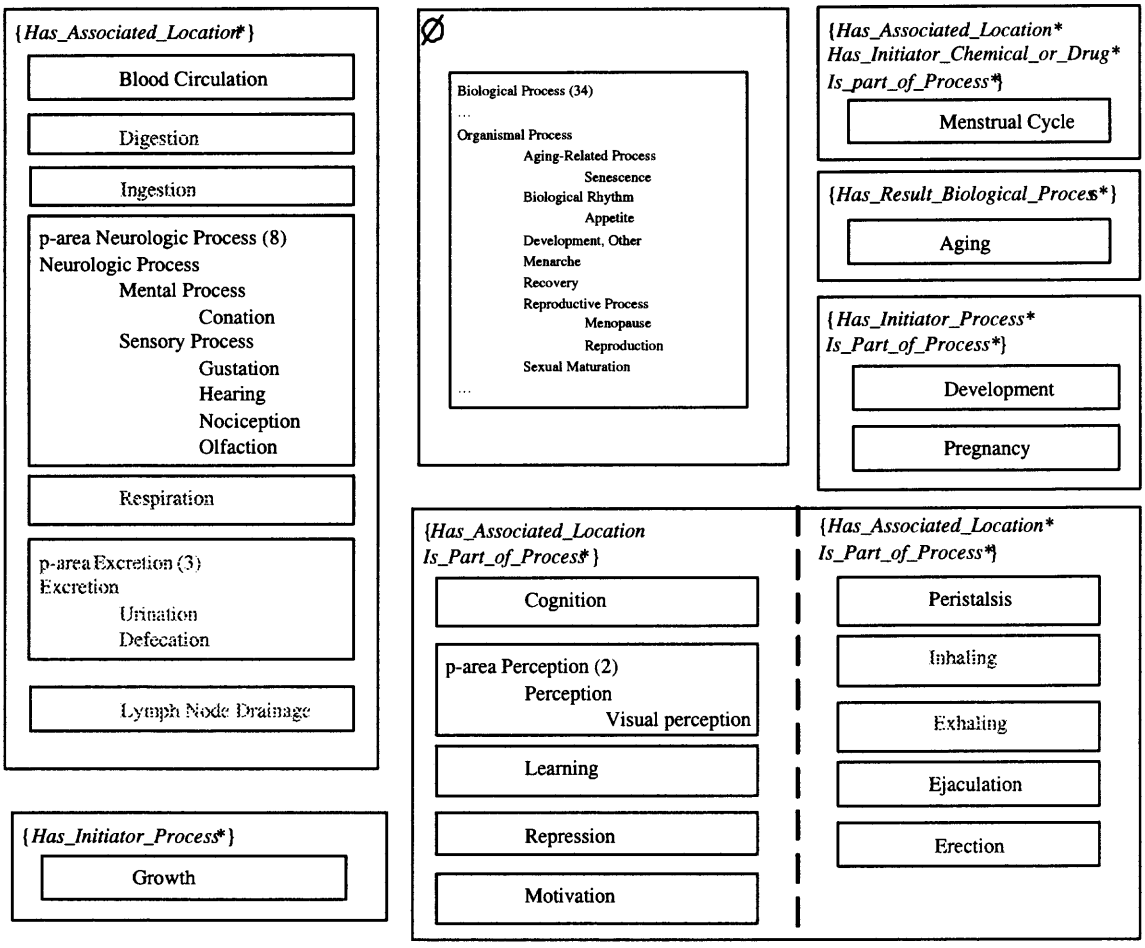


Figure 4.6 Areas and p-areas of the Organismal Process subhierarchy after corrections.

This example was brought up since it is related to a previously discussed missing role error, and because it exposes an inconsistency in the choice of names for concepts. Altogether, 70 missing synonyms are found for the Biological Process hierarchy of the NCIT. However, those are not counted as errors, and not included in the error analysis tables in Section 4.4. To mention just one other example, concepts G1 PHASE, G2 PHASE, and INTERPHASE in the p-area CELL CYCLE STAGE(11) are missing G1 PERIOD, G2 PERIOD, and RESTING PHASE as synonyms, respectively.

Missing Concepts:

Respiration consists of two parts, inspiration and expiration, with the corresponding synonyms INHALING and EXHALING, respectively. These two concepts should be in one area since they are similar in essence. From the PAT, it is found that EXHALING(1) is missing from the area that INHALING(1) is located in, and in fact from the NCIT altogether. The concept EXHALING with the synonym EXPIRATION should also be added as part of respiration to the same area as INHALING[8].

As another example, the cell cycle includes interphase (which can be divided into four steps: G0 phase, G1 phase, S phase, and G2 phase) and cell division phase. After all concepts in the p-area CELL CYCLE STAGE(11) are examined, G0 PHASE is found to be missing from the NCIT[8]. As with synonyms, the missing concepts are not included in the error analysis tables of Section 4.4.

Concept Redundancy:

The following redundancy error and missing synonyms, which are not from the Biological Process hierarchy, are mentioned due to their critical importance to NCI interests. In the Properties or Attributes hierarchy of the NCIT, there are two concepts, BENIGN and NON-MALIGNANT, listed as children of DISEASE MORPHOLOGY MODIFIER. They are synonyms, as both of them have an identical definition: "not cancerous." So only one concept should appear. The other one should be a synonym. Furthermore, NOT CANCEROUS and NONCANCEROUS should appear as synonyms, too. As a matter of fact, there is in the NCIT a concept MOUSE NONCANCEROUS CONDITIONS whose name contains such an extra synonym. Note that if a cancer researcher searches for all benign diagnoses, all those listed as NON-MALIGNANT, NOT CANCEROUS, and NONCANCEROUS will be missed.

Missing Parent:

[8]This error was corrected by Dr. Nicole Thomas, an NCIT editor, following the report

In the Biological Process hierarchy, there are only four concepts with multiple parents in the following p-areas: LEUKOCYTE TRAFFICKING(1) in {*has associated location, is part of process**}, TUMOR IMMUNITY(1) in {*has result biological process**}, INFLAMMATION PROCESS(1) in {*is part of process**}, and the root of CANCER CELL GROWTH REGULATION(3) in {*has initiator process*, is part of process**}. As the model of the NCIT allows multiple parents, this low number raises concerns that there should probably be more concepts of this sort. This is especially true since the same process can have different aspects such as structural, functional, and clinical that can be reflected by the appropriate parents. For example, the parents of INFLAMMATION PROCESS are MULTICELLULAR PROCESS (structural) and PATHOLOGIC PROCESS (clinical). The parents of CANCER CELL GROWTH REGULATION are CELL PROLIFERATION REGULATION (functional) and PATHOLOGIC PROCESS (clinical). The fact that these two concepts have the same parent, PATHOLOGIC PROCESS, suggests that the siblings of these two concepts may have more than one parent as well.

After all children of PATHOLOGIC PROCESS were examined, it was found that the p-area AUTOIMMUNE PROCESS(1) in {*has initiator chemical or drug**} has one concept, AUTOIMMUNE PROCESS, which is an immune process (involved with the immune response) and should therefore also be a child of IMMUNE FUNCTION. Another example occurs with NECROSIS (in the p-area CELLULAR PROCESS(19) of {*has associated location**}), which is a descendant of CELLULAR PROCESS. Necrosis is a pathological process caused by the progressive degradative action of enzymes and is generally associated with severe cellular trauma. Therefore, it is missing PATHOLOGIC PROCESS as another parent. For an alternative modeling approach, see Section 4.4.

Incorrect IS-A:

SENILE CORNEAL CHANGE, in the root area of the Biological Process hierarchy, is a child of PATHOLOGIC PROCESS; but this is incorrect. Senile corneal change is part of the normal aging process. It is neither abnormal nor pathologic (a manifestation of disease).

The correct placement of SENILE CORNEAL CHANGE is as a child of AGING-RELATED PROCESS (and as a sibling of AGING) in the same area [8].

The parent of TUMORIGENESIS in the p-area ONCOGENESIS(6) is ONCOGENESIS, in the area {*has associated location*, has initiator chemical or drug*, has initiator process, has result biological process*} (Figure 4.3). This represents an incorrect IS-A relationship because TUMORIGENESIS has ONCOGENESIS as a synonym [8].

Redundant Target:

The concept PHAGOCYTOSIS of the p-area PHAGOCYTOSIS(1) is in the area {*has associated location, has initiator chemical or drug*, has initiator process*, is part of process**}, with only two p-areas. It has two target values for the role *has associated location*. One is CELL and the other is PHAGOCYTIC CELL. The first target CELL should be removed from this role since the other target, PHAGOCYTIC CELL, is more specific. This and some other errors were corrected in later release of NCIT independent of the work. Some of the reported errors are still under consideration[9].

4.3.3 Testing the Hypotheses

Two hypotheses concerning the concentration of errors in specific kinds of p-areas were formulated. To test these hypotheses, all p-areas, small and large, of the Biologic Process hierarchy were audited. The analysis was concentrated on $\bar{3}$-$\bar{3}$-areas, which often represent some kind of irregularity. Altogether, there are 174 $\bar{3}$-p-areas in the PAT. Of these, there are only 27 $\bar{3}$-p-areas in 18 $\bar{3}$-$\bar{3}$-areas consisting of 33 concepts in total. The results of the analysis are given in Tables 4.1, 4.2, and 4.3.

Table 4.1 gives a breakdown of the p-areas according to their size. For each size, the number of concepts and the number and percentage of errors are listed. Table 4.2 presents a breakdown of the areas by their cardinality. The areas with only $\bar{3}$-p-areas are further

[9]Nicole Thomas, personal communication.

Table 4.1 Analysis of Errors by P-areas Sizes

P-area size	# P-areas	Total # Concepts	Erroneous Concepts	Percentage of Errors
1	141	141	18	13%
2	18	36	3	8%
3	15	45	6	13%
4–6	10	47	1	2%
7–15	10	112	0	0%
16–20	4	74	1	1%
21–50	1	47	14	30%
more than 50	1	87	1	1%
Total:	200	589	44	7%

Table 4.2 Distribution of Areas by Their Cardinality and Number of $\bar{3}$-p-areas

Area cardinality m	# Areas	# $\bar{m}$-$\bar{3}$-areas	# Concepts	# Errors	% of Errors	Other Areas	# Concepts	# Errors	% of Errors
1	15	13	18	7	39%	2	62	14	23%
2	1	1	2	1	50%	0	0	0	0%
3	5	4	13	2	15%	1	25	1	4%
4	1	0	0	0	0%	1	18	0	0%
5	4	2	12	2	17%	2	18	0	0%
6-10	7	1	11	0	0%	6	130	4	3%
11-15	3	1	13	1	8%	2	138	12	9%
16-45	1	0	0	0	0%	1	129	0	0%
Total:	37	22	69	13	19%	15	520	31	6%

distinguished from other areas. For each kind, the number of concepts, number of errors, and error percentage are listed.

In Table 4.3, only $\bar{3}$-p-areas are presented. Thus, the last row in Table 4.3, which shows the information regarding all such p-areas, reflects the sums of the first three rows of Table 4.1. In Table 4.3, it presents the distribution of these p-areas, their concepts and errors, between two kinds, according to their numbers in their respective areas. In the first row, only $\bar{3}$-$\bar{3}$-areas are considered. There are 27 such p-areas in 18 $\bar{3}$-$\bar{3}$-areas (see first three rows in Table 4.2, 13+1+4=18) consisting of a total of 33 concepts, ten of which (30%) are erroneous. In the second row, all other areas are considered. That is, cases where an area's cardinality is larger than three (e.g., the area $\{$*has result biological process*$\}$ contains five $\bar{3}$-

Table 4.3 Analysis of Errors in $\bar{3}$-p-areas of Different Kinds of Areas

# $\bar{3}$-p-areas in Area	# P-areas	# Concepts	Erroneous Concepts	Percentage of Errors
in $\bar{3}$-$\bar{3}$-areas	27	33	10	30%
Others	147	189	17	9%
Total	174	222	27	12%

p-areas, see Figure 4.4) or cases where an area contains $\bar{k}$-p-areas with k larger than three (e.g., the area {*has associated location*$^{+}$, *has initiator chemical or drug**, *has initiator process, has result biological process*} has three p-areas, but one is a $\bar{6}$-p-area and another is an $\bar{18}$-p-area. See Figure 4.3). There are 147 $\bar{3}$-p-areas with 189 concepts, 17 of which (9%) are erroneous.

4.4 Discussion

The division methodology relies on structural and semantic similarity of concepts, and it groups all concepts into areas and p-areas, accordingly. For example, the resulting division of the Biological Process hierarchy of NCIT contains 37 areas and 200 p-areas. Based on this, the AT and PAT, providing compact, abstract views of the terminology, were derived. The two diagrams help in comprehending and managing the terminology.

Auditing a whole terminology or even substantial parts of it, is an overwhelming task due to its size and complexity. Also, auditing resources are typically limited. Thus, the auditing methodology is designed to focus the available limited resources for manual editing, on relatively small parts of the terminology with high likelihood of errors. The purpose of such a focus is to maximize the impact of a limited auditing effort. This approach of the methodology is expressed by the two hypotheses of Section 4.2, discussed below.

The first hypothesis was that the probability of erroneous classifications and incorrect or incomplete modeling is higher for small p-areas than for large p-areas. As seen in

Table 4.1, the percentage of erroneous concepts for $\bar{3}$-p-areas (about 12%) is high. The percentage decreases for medium-sized p-areas (2%) and large p-areas (1%). (The one exception is discussed below.) These results support experimentally the interpretation that small p-areas, confirming to the statement of Hypothesis 1 are those with up to three concepts. The results of Table 4.1 support the first hypothesis and show that for effective and economical auditing. The effort should be concentrated on smaller p-areas, where most of the errors are likely to be.

One exception is the top-level, singly rooted area $\emptyset$ (47 concepts) with an error rate of 30%. This area contains concepts with no roles at all. However, 13 out of the 47 concepts (three of which are highlighted in Figure 4.5(b)) are missing roles. After adding the missing roles, these concepts are moved to other areas, leaving this area with 34 concepts (Figure 4.6) and one error. The error percentage of this area is thus reduced to 3%. It should be noted that there is very little semantic similarity among concepts in this area because almost all concepts located at the top levels of the hierarchy are grouped into this area. One can say that it is a very special area, which contains many unrelated concepts, since there is no unifying structure to make them similar. That is, although technically all concepts of the $\emptyset$ area share the same empty set of roles, the lack of common specific roles causes the lack of a unifying structure. One can gather from this that the auditing methodology should be augmented and special attention paid to the root area $\emptyset$ during the auditing process.

Table 4.2 gives the number of concepts and errors as a function of the cardinality of an area and the size of its p-areas. It should be noted that the likelihood of errors is higher for areas with relatively small cardinality and small p-areas versus all other areas. (Note that the 14 errors in the first row are the exceptions that were just discussed.) The combination of the two factors is considered in the following discussion of Hypothesis 2.

The second hypothesis was motivated by the intention to further prioritize the auditing of concepts of small p-areas. Such priority is important when there are not enough resources

to manually audit all the small p-areas. For example, in the case, the $\bar{3}$-p-areas add up to 222 concepts (last row of Table 4.3) which is almost 38% of the concepts in the hierarchy. Hypothesis 2 means that one expects a higher likelihood of errors in $\bar{m}$-$\bar{k}$-areas, for small m and k values. As a consequence of the results in Table 4.1, the interpretation for this hierarchy is that small p-areas are $\bar{3}$-p-areas. For testing Hypothesis 2, all $\bar{3}$-p-areas were studied. Table 4.3 compares the percentages of errors for $\bar{3}$-$\bar{3}$-areas (consisting of only $\bar{3}$-p-areas) versus $\bar{3}$-p-areas in areas with larger cardinality or with larger p-areas. As can be seen from Table 4.3, by just checking 33 concepts of the $\bar{3}$-$\bar{3}$-areas (about 15%) of the 222 concepts of $\bar{3}$-p-areas, about 37% of the 27 errors can be found in those concepts.

However, reviewing Table 4.2, one can take a less strict interpretation of what is a small cardinality. Table 4.4 presents the results where $\bar{5}$-$\bar{3}$-areas are considered, as the cardinality of an area was modified to five. For the price of reviewing 12 more concepts, two more errors are exposed. With this interpretation of Hypothesis 2, by just checking 45 concepts of $\bar{5}$-$\bar{3}$-areas (about 20% of 222 concepts in $\bar{3}$-p-areas), about 44% of the 27 errors can be exposed. Hence, there is a trade-off in choosing the number of concepts reviewed (33 versus 45) between the recall (37% versus 44%) and the precision (30% versus 27%), where erroneous concepts are considered relevant.

Table 4.4 Analysis of Errors in $\bar{3}$-p-areas of Different Kinds of Areas

# $\bar{3}$-p-areas in Area	# p-areas	Total # Concepts	Erroneous Concepts	Percentage of Errors
in $\bar{5}$-$\bar{3}$-areas	37	45	12	27%
Others	137	177	15	8%
Total	174	222	27	12%

To demonstrate the impact of the correction of the errors, Figure 4.6 shows the division of the descendants of ORGANISMAL PROCESS into areas and p-areas reflecting their structure after the corrections. It should be noted that compared to Figure 4.5(b), the number of areas was reduced from eight to seven. Furthermore, the number of $\bar{2}$-areas,

which was six in Figure 4.5(b), was reduced to five in Figure 4.6. These changes reflect simplifications of the AT and PAT following the correction of errors. Another change is the reduction of size in the AT root area $\emptyset$ from 47 to 34, due to the discovery of missing roles.

It was also found that only four concepts have more than one parent in the Biological Process hierarchy. This may be a reason for the relatively low number of errors were found in this hierarchy. Typically, many concepts with multiple parents are more complex due to the compound nature of the concepts and the multiple inheritance of roles from the different parents. Thus, one expects to find more errors in a hierarchy with more complex concepts. For comparison, the techniques were used to search for missing roles in the Experimental Organism Diagnosis hierarchy of the NCIT consisting of 1,097 concepts. Of these, 237 concepts have two parents and five have three parents. By using the methodology, 640 missing roles in 578 concepts were found, a much higher rate than in the Biological Process hierarchy where only 38 missing roles (the most common kind of error) were found.

A philosophical difference with the designers of the NCIT should be noted. As a design policy of the NCIT, all functions/processes in the Biological Process hierarchy that are not categorized as a PATHOLOGIC PROCESS are to be understood as normal biological processes. Hence, parents of concepts in the Pathologic Process sub-hierarchy can only be concepts that are categorized as pathologic processes. In other words, any normal biological process is not an appropriate parent for the descendants of PATHOLOGIC PROCESS. According to this philosophy, instead of adding more multiple parents, the NCIT modeling team modified these four concepts to have only one parent. While the NCIT approach is respected, this dissertation respectfully suggests an alternative.

In order to solve this modeling problem, it is suggested that new concepts be created as children of both PATHOLOGIC PROCESS and another normal process. These new concepts and their descendants can inherit roles from both the pathologic and normal processes. For example, a new concept called CELLULAR PATHOLOGIC PROCESS that is a child of both PATHOLOGIC PROCESS and CELLULAR PROCESS would be created. Then, the concepts

CANCER CELL GROWTH REGULATION (with its two children) and B-LYMPHOMA DEVELOPMENT would be added as children of CELLULAR PATHOLOGIC PROCESS. These concepts will inherit roles from both PATHOLOGIC PROCESS and CELLULAR PROCESS as necessary. This modeling is according to the polyhierarchy characteristic of the desiderata [16].

4.5 Conclusions

A methodology has been developed to divide a hierarchy of a medical terminology, satisfying systematic inheritance, into groups called areas and then further divide areas into p-areas. Two abstraction taxonomies, the AT and PAT, were obtained from these divisions. These taxonomies can help audit the terminology since they highlight groups of concepts with potential errors. When the auditing methodology was applied to a hierarchy of the 2004 release of the NCIT, different kinds of errors, e.g., missing roles, missing concepts, incorrect IS-As, etc. were encountered. The results of the audit show that 12% of the concepts in small p-areas have errors. Furthermore, the percentage of errors in areas with few small p-areas is high (30%). The results support both the hypotheses that direct the auditing methodology to focus the available limited resources for manual editing, on relatively small parts of the terminology with high likelihood of errors. At the same time, the need to include auditing as an integral part of the terminology design life cycle, following similar actions taken in software engineering and knowledge-based systems [62, 66] has been demonstrated with the errors exposed.

CHAPTER 5

STRUCTURAL AUDITING TECHNIQUES FOR GENE HIERARCHY IN NCI THESAURUS

5.1 Background

5.1.1 Structural Characteristics of the NCIT Gene Hierarchy

There are 1,786 concepts in the Gene hierarchy of the NCIT (2004 version). There are 1,554 concepts located at the leaves of this hierarchy, i.e., these concepts have no children. They are the actual gene concepts. The 232 internal concepts serve to classify the genes into categories. The Gene hierarchy is different from other NCIT hierarchies in that the internal concepts are not gene concepts, just categories of genes. In contrast, an internal concept of the Biological Process hierarchy can be a process with more refined processes as children. For example, the CANCER PROGRESSION internal concept describes a process, and has 12 descendants.

There are only 42 concepts with two parents, and all are gene concepts. Examples includes GRB7 GENE, MADD GENE, and MAGED1 GENE. Only one, SMARCC2 GENE, has a child. (The issue of a gene concept with a child is discussed in Section 5.4.) The Gene hierarchy has eight levels. An example of a longest path of eight concepts, each one more specific than the previous one, is GENE, ENZYME GENE, HYDROLASE GENE, PHOSPHATASE FAMILY GENE, PROTEIN PHOSPHATASE GENE, PROTEIN SERINE-THREONINE PHOSPHATASE GENE, PROTEIN PHOSPHATASE 2A SUBUNIT GENE, PPP2R5D GENE.

The number of children (called the *degree*) for internal, category nodes varies from 1 to 116 (see Table 5.1). For example, PROTEIN PHOSPHATASE GENE has 29 descendants, which are thus comprising the concepts satisfying the protein phosphatase category. Among them, there are five children of PROTEIN PHOSPHATASE GENE, three of which are category

Table 5.1 Degree Distribution for Category Concepts of the Gene Hierarchy

Degree	# internal nodes	Degree	# internal nodes	Degree	# internal nodes
1	56	12	2	31	1
2	40	13	1	35	2
3	19	14	3	36	3
4	17	15	2	38	1
5	14	16	1	47	1
6	12	17	2	50	1
7	10	20	1	51	1
8	9	21	2	67	1
9	9	22	2	74	1
10	6	23	1	89	1
11	8	24	1	116	1

Table 5.2 Characteristics of the Gene Hierarchy Levels

Level	Internal Category Concepts	Mixed Category Concepts	Terminal Category Concepts	Total # Category Concepts	Gene Concepts	Total # Concepts
0	0	1	0	1	0	1
1	2	20	8	30	6	36
2	4	24	45	73	500	573
3	4	18	46	68	409	477
4	1	7	31	39	379	418
5	1	4	10	15	126	141
6	0	0	6	6	102	108
7	0	0	0	0	32	32
Total	12	74	146	232	1554	1786

concepts and two are gene concepts. The distribution of the concepts separated into category and gene concepts among the levels is presented in Table 5.2. About 83% of the nodes are located on a few middle levels: 73 internal nodes and 500 leaves on level 2; 68 internal nodes and 409 leaves on level 3; and 39 internal nodes and 379 leaves on level 4. It can distinguish between three kinds of category concepts: (1) *Terminal category concepts* where all children are genes; (2) *Internal category concepts* where all children are also category concepts; and (3) *Mixed category concepts* having both kinds of children. An example of the latter is PROTEIN PHOSPHATASE GENE. Table 5.2 shows that the majority (63%) of categories are terminal, while just a few (5%) are Internal. The remaining 32%

are mixed. The 86 internal and mixed category concepts form the "skeleton" of the Gene hierarchy, providing a compact view of the kinds of genes that exist in the hierarchy.

5.1.2 Importance of Auditing Gene Hierarchy in the NCIT

One of the fastest expanding areas of biomedical research concerns the knowledge of genes and genomes. The Human Genome Project (HGP) gathered knowledge from human DNA strands. Obtaining a comprehensive human genome sequence has strongly impacted areas of biomedical research and medicine [20]. For example, the identification of a gene permits the development of diagnostic tests that reveal potential health problems before they manifest themselves as symptoms [39]. Knowing a patient's genetic makeup may allow physicians to minimize certain disease risks [35]. The results produced by the genome project enhance the understanding of human heredity. Overall, it supports the study of carcinogenesis, the design of antimicrobial drugs, gene therapy, and fundamental biomedical research.

Ongoing and rapid advances in the understanding of genomic phenomena are becoming increasingly important for clinical research and medicine. The number of scholarly articles in biomedical research is growing at an astronomical rate [23]. A significant number of databases are now collecting and cataloging genomic data. An annual review of molecular biology databases, for example, lists several hundred databases relevant to the genomic domain [3]. The databases that will have the greatest impact are those able to link transparently to other closely related resources.

The Gene Ontology (GO) [24] provides a controlled terminology allowing researchers to report their results regarding genes and gene products. It continues to be an important resource in the molecular biology domain. Many of the model organism databases are devoting abundant resources to annotating the genes in their databases with GO codes. GO is composed of three disjoint components: cellular components, molecular functions, and

biological processes. As of December 2005, GO contained 1,681 component terms, 7,384 function terms, and 10,291 process terms.

The UMLS integration effort involving GO's concepts is reported in [41]. However, according to the design policy of the UMLS, only concepts and relationships from GO were incorporated. No relationships to the rest of the UMLS were added. The NCIT has added such relationships.

5.2 Auditing Methodology

A structural auditing methodology [48] has been applied to the Biological Process and Experimental Organism Diagnosis hierarchies of the NCIT. However, due to the special nature of the Gene hierarchy (see Section 5.1.1), this structural auditing methodology must be adapted for application to the Gene hierarchy. The required modifications are presented.

The auditing methodology for the Gene hierarchy is performed in two major parts as presented in Figure 5.1. First, it divides the current Gene hierarchy into areas and p-areas and constructs the corresponding taxonomies as described in Section 4.2. Second, auditing is performed in three steps that utilize the areas, p-areas, and the taxonomies to focus on certain groups of concepts.

Nearly all of the genes in the NCIT are derived from the DNA sequence data. Therefore, the originating organism and the location of the gene (chromosome and indices of introns) should be known. Also, many genes have known disease-associated alleles. In each such case, the gene should be assigned roles like *gene_associated_with_disease* (in short, "*disease* role"), *gene_found_in_organism*, and *gene_in_chromosomal_location*. Nevertheless, many gene concepts are missing such roles. In this work, It is not searching for general modeling errors in the Gene hierarchy. It just limits the attention to discovering role errors, e.g. missing roles, missing targets for existing roles, and wrong targets or redundant targets in the Gene hierarchy concepts, which are common errors in this hierarchy. Due to the special nature of the Gene hierarchy, It needs to tailor the auditing methodology to

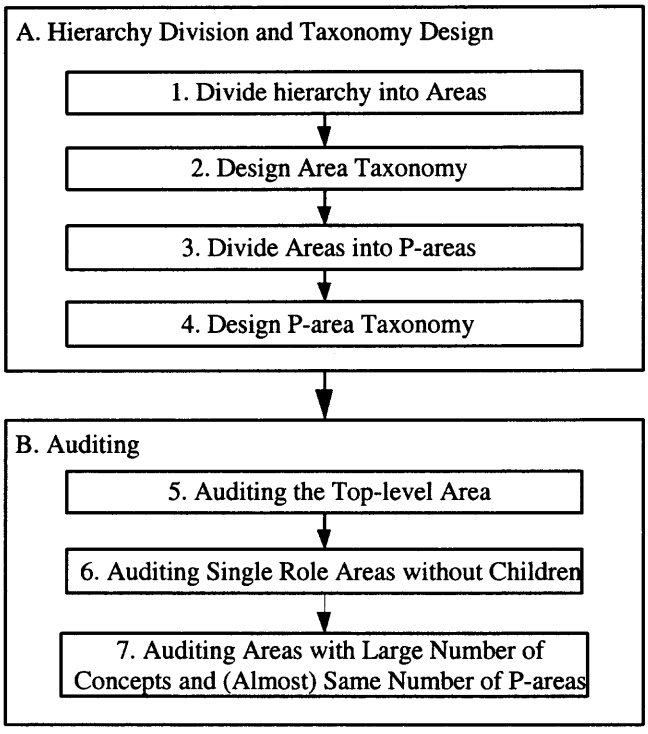


Figure 5.1 The flow chart of the auditing methodology.

this hierarchy's special structure discussed in Section 5.1.1. In particular, three different auditing methodologies are used for the Gene hierarchy.

The methodology is targeted for auditing specific areas and p-areas where the likelihood of finding errors is high. In the previous work of auditing the Biological Process hierarchy of [48], it did not focus on the concepts that are leaves. The reason is that one specific process may have another more refined process as its child. For example, VISUAL PERCEPTION is a child of PERCEPTION. But in the Gene hierarchy, no gene concept should be a child of another gene concept. All internal concepts in the Gene hierarchy should represent general categories of genes, e.g., CELL CYCLE GENE and REGULATORY GENE.

5.2.1 Review of the Top-level Area: ∅

In previous research work [48], the concepts in the top-level singly rooted area ∅ have shown a high percentage of errors. There is very little semantic similarity among its many concepts because there is no true unifying set of roles. There are in fact no roles.

5.2.2 Review First-level Areas having No Children

For the first-level areas having no children, p-areas with one concept inside these areas may have errors. This phenomenon implies that the concepts of such p-areas are gene concepts. Every concept located in the first-level areas has just one role. As explained before, information on other roles should be available for genes by their way of discovery. The situation of a gene concept with only one role is typically indicating that other roles are missing. On the other hand, if an area of just one role has children, then such an area may contain internal nodes representing categories rather than genes which are less likely to miss roles.

5.2.3 Review Large Areas with Number of P-areas Close to Number of Concepts

Continuing with the same kind of reasoning, one may look for areas with two or more roles which may be still missing other roles. The methodology concentrates on relatively large areas with all or almost all concepts being genes. Such areas are recognized by having (almost) as many p-areas as concepts. This is the situation due to many leaves introducing the same role. The reason for concentrating on such large areas is the expectation of finding many concepts missing the same role.

Another two kinds of errors considered are in cases where a role exists. One possible kind of error is that the target concept is wrong. For example, PPP2R5E GENE has the chromosomal location 7p12-p11.2 in the NCIT. However, the correct location should be 14q23.1. The other possible error is that the existing target(s) for the role are correct, but there should be more targets. As before, the methodology focuses on areas with many

concepts and (almost) as many p-areas indicating many similar gene concepts, to maximize the number of errors found, while minimizing the effort.

5.2.4 Hypotheses

The underlying assumptions of the methodology can be expressed by the following two hypotheses.

Hypothesis 1: The probability of role errors in concepts is higher for small p-areas than for large p-areas.

Hypothesis 2: The probability of role errors in concepts is higher in areas with a large number of concepts and (almost) the same number of p-areas than in other areas.

The reasoning for the first hypothesis is that gene concepts tend to appear in p-areas of one concept. Some roles appear just for genes and not for categories (e.g., chromosomal location), so they are introduced at the leaves. For other roles, e.g., *gene_plays_role_in_process* (in short, "*process* role") or *disease* roles, they are mainly introduced at the leaf level although not solely. The reasoning for the second hypothesis is that those are the areas with many p-areas of one concept.

5.3 Results

5.3.1 AT and PAT for the NCIT Gene Hierarchy

There are six roles (numbered from zero to five for convenience) defined for the concepts of the Gene hierarchy of the NCIT. They are:

0: gene_associated_with_disease
1: gene_found_in_organism
2: gene_in_chromosomal_location
3: gene_plays_role_in_process
4: gene_is_biomarker_type
5: gene_is_biomarker_of

The AT for the Gene hierarchy is displayed in Figure 5.2. The 1,786 concepts are divided into 27 areas. Each area is drawn as a large bold box. An area is named by listing its roles inside curly braces. As explained in Section 4.2, a "*" following a role name indicates the introduction of the role at this area while a "+" is used when the role is introduced at some roots and inherited at others. The number in parentheses following the name of an area is its number of p-areas. Areas with the same number of roles are located on the same level of the AT. There are seven levels labeled from 0 to 6 according to their number of roles. A *child-of* relationship is drawn as a bold arrow.

The 27 areas are further divided into 1583 p-areas. Only a portion of the PAT, consisting of the top two levels of the AT, is presented in Figure 5.3 due to lack of space. There are four areas from levels 0 and 1 in Figure 5.3. Each small box inside an area box represents one p-area. The number of concepts in a p-area is listed in parentheses following its name. A dashed-line box groups p-areas such that their roots have the same parent. For example, NEO GENE and LACZ GENE in the area $\{1^*\}$ share a common parent REPORTER GENE in $\emptyset$ (see Figure 5.3). The root of the p-area is connected to its parent by a thin arrow. The "..." in the area $\emptyset$ denotes the fact that only some of the concepts that are parents of the roots of p-areas in the four shown areas are listed. An indented hierarchy format is used to display concepts in $\emptyset$.

To demonstrate higher levels of the PAT, it will bring a small excerpt. For example, CYP1A1 GENE, KLK2 GENE, and ELF3 GENE have the same roles 1 through 5 and are grouped into the same area $\{1^+, 2^+, 3, 4^*, 5^+\}$ (Figure 5.4). The parent of CYP1A1 GENE, CYTOCHROME P450 FAMILY GENE, belongs to $\{3^*\}$. The parent of KLK2 GENE, KLK3 GENE, belongs to $\{1^*, 2^*, 3, 5^*\}$. The parent of ELF3 GENE, TRANSCRIPTION COACTIVATOR GENE, belongs to $\{1^*, 3^+\}$. Since the parents of these three concepts belong to other areas, they are roots of $\{1^+, 2^+, 3, 4^*, 5^+\}$. All three roots introduce the role 4, and so "*" follows 4 in the name. The role 3 is inherited from the parents of the three roots. The root CYP1A1 GENE introduces role 1. However, the other two roots

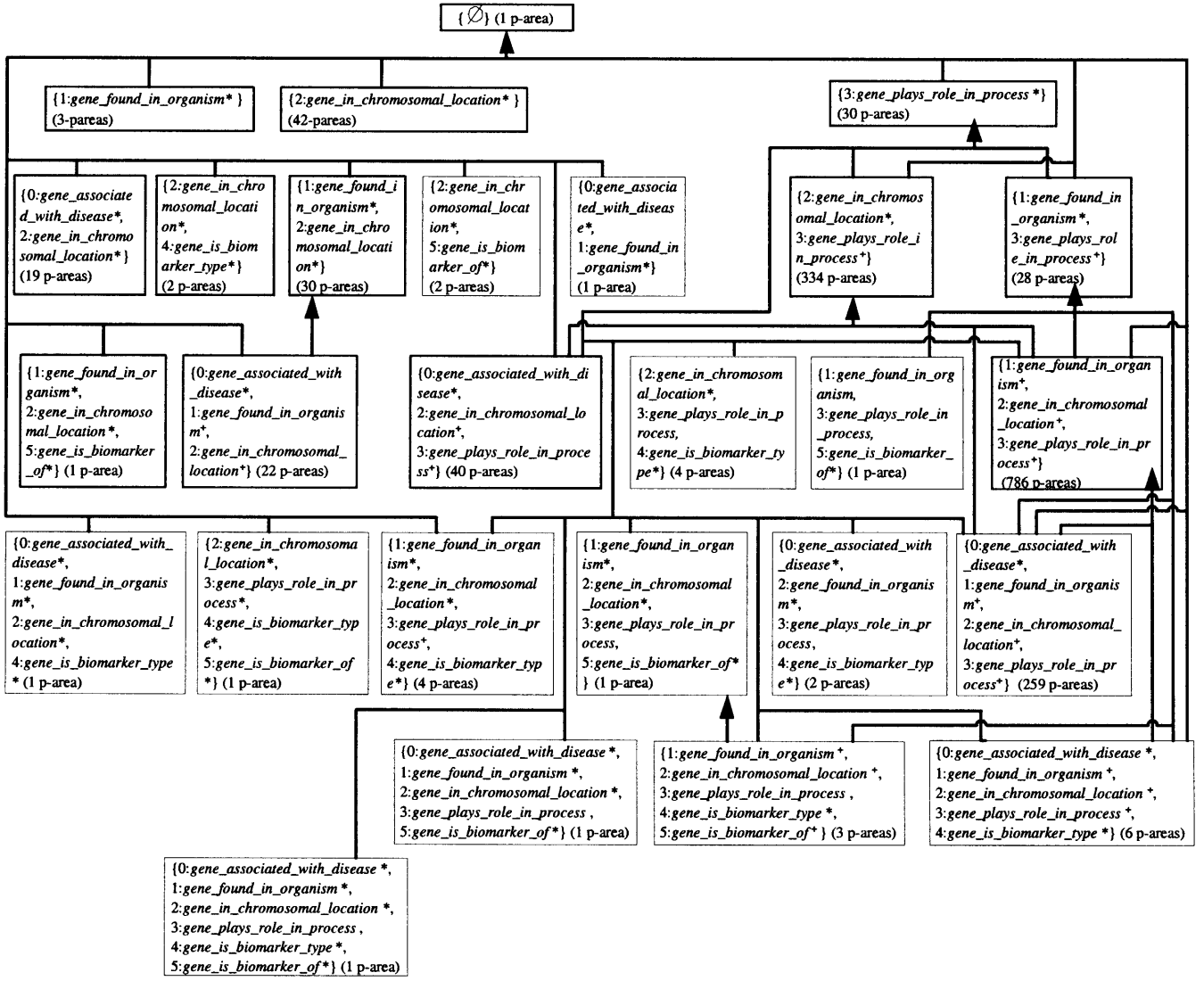


Figure 5.2 Area taxonomy for the Gene hierarchy.

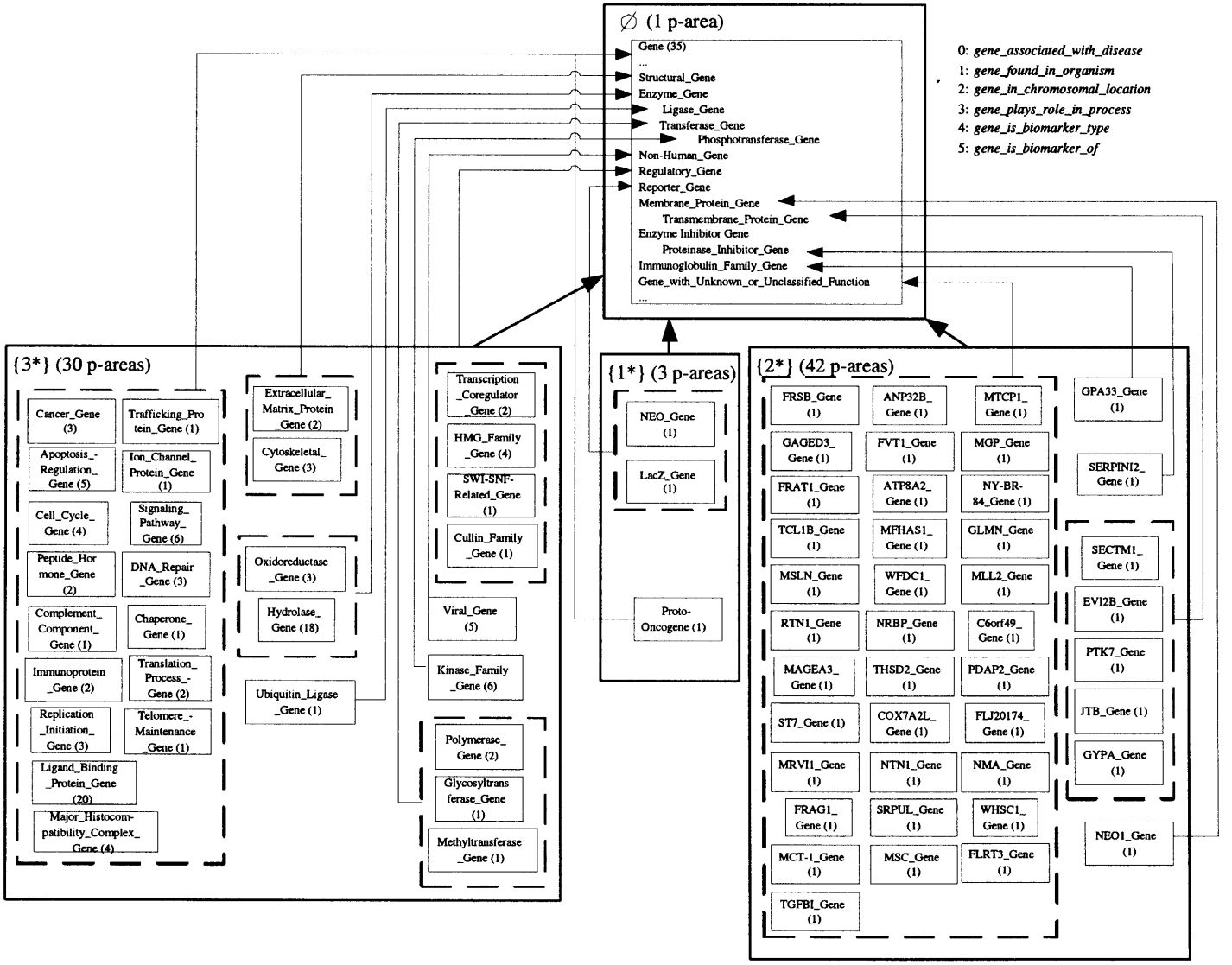


Figure 5.3 Excerpt of the p-area taxonomy for the Gene hierarchy.

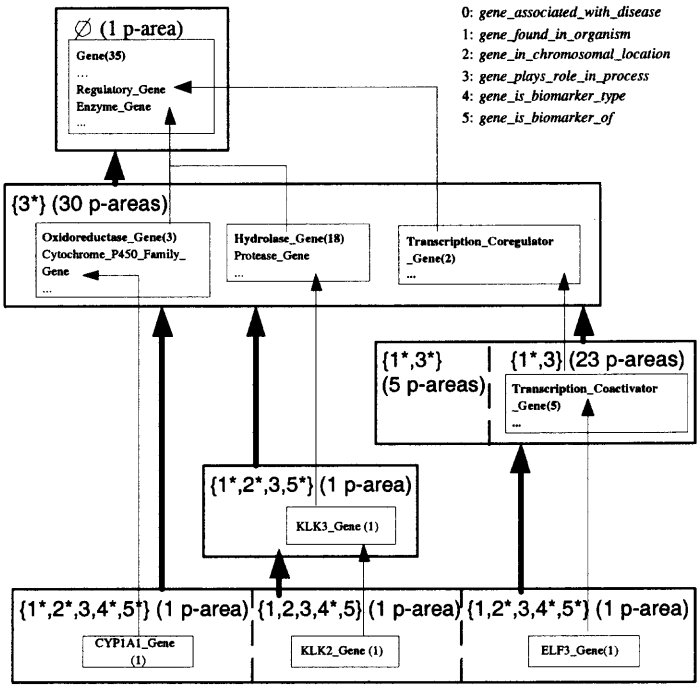


Figure 5.4 Another excerpt of the p-area taxonomy.

(KLK2 GENE and ELF3 GENE) inherit it. Thus, "+" follows 1 in the name. A similar situation happens for roles 2 and 5. The area is further divided into three p-areas due to its three roots. Those three p-areas have three different introduction patterns. Thus, they are in different parts of the area separated by dashed lines (Figure 5.4). Due to space limitations, all 28 p-areas of $\{1^*, 3^+\}$ are presented in Figure 5.5.

Table 5.3 Distribution of Areas, P-areas, and Concepts by Level

Level	# Areas	# P-areas	# Concepts
0	1	1	35
1	3	75	154
2	7	413	476
3	6	821	846
4	6	262	264
5	3	10	10
6	1	1	1
Total	**27**	**1,583**	**1,786**

The distributions of the areas, p-areas, and concepts in the AT levels are presented in Table 5.3. For example, level 1 contains three areas, 75 p-areas, and 154 concepts. In

Table 5.4 Areas with Their Numbers of Concepts and P-areas

Area	# Concepts	# P-areas
$\emptyset$	35	1
{1*}	3	3
{3*}	109	30
{2*}	42	42
{0*,1*}	1	1
{0*,2*}	19	19
{1*,2*}	32	30
{1*,3+}	81	28
{2*,3+}	339	331
{2*,4*}	2	2
{2*,5*}	2	2
{0*,1+,2+}	22	22
{0*,2+,3+}	40	40
{1+,2+,3+}	778	753
{1*,2*,5*}	1	1
{1,3,5*}	1	1
{2*,3,4*}	4	4
{0*,1+,2+,3+}	255	253
{0*,1*,2*,4*}	1	1
{0*,2*,3,4*}	2	2
{1*,2*,3,4*}	4	4
{1*,2*,3,5*}	1	1
{2*,3*,4*,5*}	1	1
{0*,1+,2+,3+,4*}	6	6
{0*,1*,2*,3,5*}	1	1
{1+,2+,3,4*,5+}	3	3
{0*,1*,2*,3,4*,5*}	1	1
Total	**1,786**	**1,583**

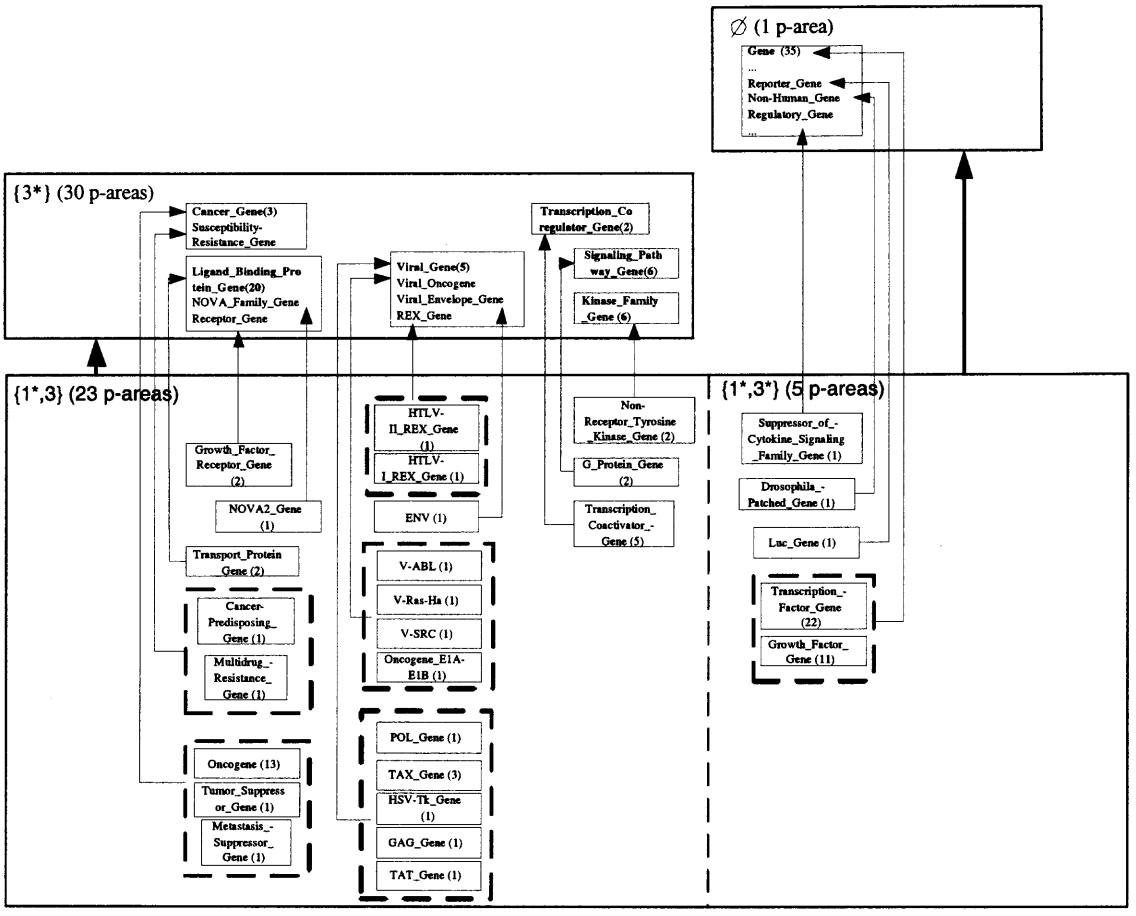


Figure 5.5 Another excerpt of the p-area taxonomy.

Figure 5.3, the three areas on level 1 introduce roles 1, 2, and 3, respectively. Level 3 contains the largest number of p-areas among the seven levels. All concepts without any role are grouped into $\emptyset$, the only area on level 0. It notes that same role can be introduced at different areas on different levels.

A list of all areas along with their numbers of concepts and p-areas is presented in Table 5.4. The area $\{1^+, 2^+, 3^+\}$ on level 3 is the largest. It contains 753 p-areas and 778 concepts. Out of the 1,583 p-areas, 1,526 (96%) have only one concept. There are 32 p-areas with two concepts. Table 5.5 presents the distribution of p-areas and concepts according to p-area size. It shows that the concepts tend to appear in very small p-areas and mainly in p-areas of one concept. Only 119 concepts appear in p-areas of size greater than ten.

Table 5.5 P-area Size Distribution

Size of p-areas	# p-areas	# concepts
1	1526	1526
2	32	64
3	9	27
4	3	12
5	4	20
6	3	18
11	1	11
13	1	13
18	1	18
20	1	20
22	1	22
35	1	35
Total	1,583	1,786

5.3.2 Role Errors Discovered

As discussed, the search for role errors focused on: (1) Area ∅; (2) areas having one role and no children; and (3) areas with a large number of concepts and (almost) the same number of p-areas (see Hypothesis 2). The auditing results are presented in the following subsections for each of these.

Top Area of the AT: ∅

There are 35 concepts in ∅. TK GENE and CAT GENE are the only two leaves. TK GENE is missing the role 3, while CAT GENE is missing roles 2 and 3. The category concept ENZYME INHIBITOR GENE is missing the *process* role with value ENZYME INHIBITION. INHIBITION IS-A CONCEPTUAL ENTITIES in the NCIT, but it should be a process as indicated by the definition and by its semantic type (Natural Phenomenon or Process). Similarly, the three descendants of ENZYME INHIBITOR GENE, PROTEINASE INHIBITOR GENE, CYSTEINE PROTEINASE INHIBITOR GENE, and CYSTATIN SUPERFAMILY GENE, are missing this role. The area ∅ has 15 out of 35 concepts (43%) with missing roles (see Table 5.6).

Table 5.6 Missing Roles for Concepts in $\emptyset$

Concept Name	Missing Role	Role Target Values
TK Gene	3	Phosphorylation
CAT Gene	2	11p13
	3	Detoxification, Acetylation
Enzyme Inhibitor Gene	3	Enzyme Inhibition
Proteinase Inhibitor Gene	3	Enzyme Inhibition
Cysteine Proteinase Inhibitor Gene	3	Enzyme Inhibition
Cystatin Superfamily Gene	3	Enzyme Inhibition
Enzyme Gene	3	Biochemical Reaction
Ligase Gene	3	Biochemical Reaction
Transferase Gene	3	Biochemical Reaction
Phosphotransferase Gene	3	Biochemical Reaction
Regulatory Gene	3	Biochemical Process
hGH Gene	3	Biochemical Process
Nucleosome Assembly Protein Gene	3	Biochemical Process
Immunoglobulin Gene	3	Host Defense Mechanism
CEA Family Gene	3	Host Defense Mechanism

First-level Areas without Children

There are two first-level areas without any children: {1*} and {2*} (Figure 5.2). For area {2*}, all 42 concepts are found to have missing roles. For example, ANP32B GENE is missing the role *gene_found_in_organism* with value HUMAN. (This role has been added to the current version of the NCIT). It should have two roles after the new role is added. Since its parent, GENE WITH UNKNOWN OR UNCLASSIFIED FUNCTION, has no role, both roles should have been introduced at this concept. The concept is thus moved from the original area to the area {1*, 2*}. Another example, MTCP1 GENE is missing the roles *gene_associated_with_disease* to the disease LEUKEMIA, *gene_found_in_organism* with value HUMAN, and *gene_plays_role_in_process* with the values CELL PROLIFERATION and REGULATION OF PROGRESSION THROUGH CELL CYCLE. After adding these three new roles, this concept is moved to the area {0*, 1^+, 2^+, 3^+}. All concepts in this area are missing one or more roles. After corrections, they are moved to five other areas. The original area in fact will disappear from the AT.

There are only three concepts in area {1*}. They are NEO GENE, LACZ GENE, and PROTO-ONCOGENE. LACZ GENE is missing role 3 (*gene_plays_role_in_process*). It should

move to the area $\{1^*, 3^+\}$. On the other hand, PROTO-ONCOGENE should not have the role 1. It is a category concept[8]. It should belong to the area $\emptyset$. Only one concept remains in $\{1^*\}$ after the corrections.

Interestingly, the other first-level area $\{3^*\}$, having children (Figure 5.2) and not being targeted for auditing, is of a different nature from the other two areas. Its concepts are not gene concepts but categories of genes (see Figure 5.3). As a matter of fact, no role errors were observed for any of the area's concept (Table 5.8).

Large Areas with (almost) all P-areas of One Concept

For Hypothesis 2, two criteria are applied to determine the areas targeted for auditing. The first requires a large number of concepts. The second criterion requires almost the same number of p-areas as concepts. It needs now to interpret those two quantitative measures. It selected the following interpretation for these two criteria: (1) a large area contains more than 30 concepts, and (2) the ratio of the number of p-areas to the number of concepts should be greater than 0.9. There are nine areas that meet Criterion 1. Three of these nine areas have a ratio less than 0.35. There are only six areas that meet both criteria. One of them ($\{2^*\}$) is discussed in the previous subsection. The remaining five are $\{1^*, 2^*\}$, $\{2^*, 3^+\}$, $\{0^*, 2^+, 3^+\}$, $\{1^+, 2^+, 3^+\}$, and $\{0^*, 1^+, 2^+, 3^+\}$ (as shown in Table 5.4).

The largest area $\{1^+, 2^+, 3^+\}$ contains 778 concepts and 753 p-areas. Among these, 364 concepts (located in 348 p-areas) have different kinds of errors. For example, the concept GATA1 GENE is missing the role *gene_associated_with_disease* with the value DYSERYTHROPOIETIC ANEMIA. Forty concepts have incorrect chromosomal location. For example, DTX1 GENE has chromosomal location 12q24.21. The correct location should be 12q24.13. A large number (41%) of the concepts in this area are found to be missing some target values for the role *gene_plays_role_in_process*. Although IL10RB GENE has the role *gene_plays_role_in_process* that connects it to INTERCELLULAR COMMUNICATION and RECEPTOR SIGNALING, it is still missing connections to BLOOD COAGU-

[8]Dr. Nicole Thomas, personal communication.

LATION and INFLAMMATORY RESPONSE. The areas {0*, 2^+, 3^+} and {0*, 1^+, 2^+, 3^+} also have high percentages (50% and 38%, respectively) of missing target values for *gene_plays_role_in_process* role.

The other two areas have just two roles and have high percentages of missing roles. For example, 305 out of the 339 concepts (90%) in {2*, 3^+} are missing *gene_found_in_organism*. It also has 19 concepts with the wrong chromosomal location. It further has 144 concepts missing target values for the process role. The last area, {1*, 2*}, has eight out of the 32 concepts (25%) with missing roles. It further has three concepts with the wrong chromosomal location.

5.3.3 Error Distributions in P-areas and Areas

The two hypotheses, in Section 5.2, express an expectation for higher probability of errors for some areas and p-areas. The auditing efforts should concentrate on these areas and p-areas. It now investigates error distributions in p-areas to check Hypothesis 1. Error distributions in areas are investigated to check Hypothesis 2.

There are 879 concepts detected by the auditing process with various types of role errors. Note that a concept may have more than one kind of role error. There are 377 concepts with missing roles, 598 concepts with missing role target values, and 80 concepts with the wrong target values. The distribution of erroneous concepts by p-area size is presented in Table 5.7. Among these concepts, 837 errors are in p-areas with one concept. The error percentage is as high as 55% for the concepts located in the p-areas with one concept. There are 31 errors in the p-areas with two concepts. The role-error percentage is as high as 48% for the concepts in these p-areas. (For another kind of error, see Section 5.4). The percentage of errors is reduced to single digits when the size of the p-areas is larger than two. No systematic trend appears for the error percentages for the p-areas with more than two concepts. However, when all errors are counted together for these p-areas, the

error percentage (6%) is significantly smaller than for p-areas with one or two concepts. This result confirms Hypothesis 1.

Table 5.7 Erroneous Concept Distributions by Size of P-areas

P-area size (# concepts)	# P-areas	# Concepts	Erroneous Concepts	Percentage of Errors
1	1526	1526	837	55%
2	32	64	31	48%
3	9	27	1	4%
4	3	12	–	0%
5	4	20	4	20%
6	3	18	6	33%
more than 7	6	119	–	0%
Total	1,583	1,786	864	49%

The error distribution among the areas is presented in Table 5.8. Besides the number of erroneous concepts (column 4) and percentage of erroneous concepts (column 5), the number of erroneous concepts and their percentages for each type of role error are also listed in the table. For the five areas selected by the second criterion as reported in Section 5.3.2, the percentages of erroneous concepts are as high as 34% for $\{1^*, 2^*\}$, 90% for $\{2^*, 3^+\}$, 50% for $\{0^*, 2^+, 3^+\}$, 45% for $\{1^+, 2^+, 3^+\}$, and 41% for $\{0^*, 1^+, 2^+, 3^+\}$ (Table 5.8). This result confirms Hypothesis 2.

For the other two areas, $\{1^*\}$ and $\{2^*\}$, with just one role and without any children, the error percentages are also high: 67% for $\{1^*\}$, and 100% for $\{2^*\}$. When combining these two areas, 44 out of 45 concepts (98%) have errors. All the concepts in these areas are gene concepts. This result confirms the viability of the auditing technique, presented in Section 5.3.2, regarding such concepts. The top-level area, $\emptyset$ has an error percentage as high as 43%. This result confirms the viability of focusing on this area (Section 5.3.2).

Table 5.8 Error Distribution of Areas and P-areas

Area	# concepts	# p-areas	# erroneous concepts	% of erroneous concepts	# p-areas with errors	% of p-areas with errors	# concepts with missing roles	% erroneous concepts with missing roles	# concepts with wrong chromosomal location	% of erroneous concepts with wrong chromosomal location	# concepts with missing targets	% of erroneous concepts with missing targets
$\emptyset$	35	1	15	43	1	100	15	43	–	–	–	–
{1*}	3	3	2	67	1	33	1	33	1	33	–	–
{2*}	42	42	42	100	42	100	42	100	3	7	1	2
{3*}	109	30	–	–	–	–	–	–	–	–	–	–
{0*,1*}	1	1	–	–	–	–	–	–	–	–	–	–
{0*,2*}	19	19	1	5	1	5	1	5	–	–	–	–
{1*,2*}	32	30	11	34	11	37	8	25	3	9	–	–
{1*,3+}	81	28	1	1	1	4	1	1	–	–	1	1
{2*,3+}	339	331	305	90	300	91	305	90	19	6	144	42
{2*,4*}	2	2	1	50	1	50	–	–	–	–	1	50
{2*,5*}	2	2	1	50	1	50	1	50	–	–	–	–
{0*,1+,2+}	22	22	3	14	3	14	2	9	1	5	–	–
{0*,2+,3+}	40	40	20	50	20	50	–	–	2	5	20	50
{1+,2+,3+}	778	753	364	47	348	46	1	0	40	5	325	42
{1*,2*,5*}	1	1	–	–	–	–	–	–	–	–	–	–
{1,3,5*}	1	1	–	–	–	–	–	–	–	–	–	–
{2*,3,4*}	4	4	1	25	1	25	–	–	–	–	1	25
{0*,1+,2+,3+}	255	253	105	41	105	42	–	–	11	4	98	38
{0*,1*,2*,4*}	1	1	–	–	–	–	–	–	–	–	–	–
{0*,2*,3,4*}	2	2	1	50	1	50	–	–	–	–	1	50
{1*,2*,3,4*}	4	4	3	75	3	75	–	–	–	–	3	75
{1*,2*,3,5*}	1	1	1	100	1	100	–	–	–	–	1	100
{2*,3*,4*,5*}	1	1	–	–	–	–	–	–	–	–	–	–
{0*,1+,2+,3+,4*}	6	6	2	33	2	33	–	–	–	–	2	33
{0*,1*,2*,3,5*}	1	1	–	–	–	–	–	–	–	–	–	–
{1+,2+,3,4*,5+}	3	3	–	–	–	–	–	–	–	–	–	–
{0*,1*,2*,3,4*,5*}	1	1	–	–	–	–	–	–	–	–	–	–
Total	1,786	1,583	879	49	843	53	377	21	80	4	598	33

5.4 Discussion

5.4.1 Advantages of the AT and PAT

The AT and PAT are derived from the divisions based on structural similarity of the concepts. There are 27 areas in the AT. It is two orders of magnitude smaller than the original terminology's concept network containing 1,786 concepts. However, there are 1,583 p-areas in the PAT, almost as many as the number of concepts. This is mainly due to roles introduced in the gene concepts at the leaves. Hence, the AT, not the PAT, is a compact abstraction network helping in the orientation and navigation of the structure of the terminology. The AT and PAT can be utilized to guide the auditing of the Gene hierarchy, focusing attention on areas with a high likelihood of errors.

Group-based auditing directs the review process to groups of similar concepts, rather than reviewing each concept independently. When a concept is reviewed in the context of other similar concepts, the review can help expose errors that may not be detected otherwise. The AT offers groups of concepts similar in their structure. Furthermore, the PAT directs the auditor to review groups that are both structurally and semantically similar, for the p-areas that have more than one concept. Hence, the AT and PAT support group-based auditing.

Many concepts in the Gene hierarchy are missing roles. For example, 42 concepts in the area {2*} have just the role 2. Each of them is missing at least one role. With so many roles missing from the Gene hierarchy, it took the approach of first dealing only with role errors. Only after correcting the roles, should auditing for other errors proceed. In this chapter, it reported only on this effort of detecting role errors.

Hypothesis 1 asserts that the probability of erroneous concepts is higher for small p-areas than for large p-areas. When interpreting small p-areas as having one or two concepts, the percentage is as high as 54% for erroneous concepts in small p-areas. The percentage decreases to 6% for larger p-areas, which supports Hypothesis 1. The auditing effort should

be concentrated on the small p-areas. The areas in which to find these p-areas are given by Hypothesis 2.

As a side remark, it also found in a number of p-areas of two concepts another kind error: incorrect IS-A relationships. In those cases, the p-area consists of two gene concepts, one IS-A of the other. This is a violation of the rule of the Gene hierarchy that all gene concepts are leaves and all internal nodes are category concepts. Thus, a gene concept can only be IS-A a category concept. Almost all the errors were corrected independently in a later release of the NCIT. One such error still appearing is IICER1 GENE IS-A RNASE 3 GENE.

Hypothesis 2 asserts that the probability of erroneous concepts is higher in areas with large numbers of concepts and (almost) the same number of p-areas than in other areas. When interpreting large areas to contain at least 30 concepts, the percentage of erroneous concepts is 56%. If, on the one hand, the size of a large area is lowered to more than 15, in effect adding the two areas $\{0^*, 2^*\}$ and $\{0^*,1^+,2^+\}$ with a total of 4 erroneous concepts, the percentage is just slightly reduced to 52%. If, on the other hand, the size of a large area is raised to more than 100 concepts, and thereby lose the two areas $\{1^*, 2^*\}$ and $\{0^*,2^+,3^+\}$ and 31 erroneous concepts, the percentage is hardly changed at 55%. The consequence is that auditors should concentrate on such large areas with any reasonable interpretation of the term "large".

5.4.2 Improving the Modeling of the Gene Hierarchy

Lowering the Number of Role-Introducing Concepts

Table 5.5 shows that the majority (96%) of the p-areas have one concept. The majority of concepts (87%) are located at the leaves of the IS-A hierarchy. These concepts located at the leaves are introducing new roles. Many of these concepts have the same parent. If an intermediate "generalizing" concept were added as a parent of the role introducing concepts and as a child of the previous joint parent to centrally introduce the roles

only once, the number of p-areas with one concept would be reduced significantly. A better modeling could be gained for the price of adding a few new concepts. This would simplify the AT and PAT structure. Moreover, it would simplify the terminology itself. It suggests a new measure of "role-definition complexity" for terminologies as the ratio between the number of role introducing concepts to the total number of concepts. In a terminology with a lower role definition complexity, fewer roles need to be introduced explicitly and more are defined implicitly by inheritance. According to such a complexity measure, the transformation described above simplifies the terminology.

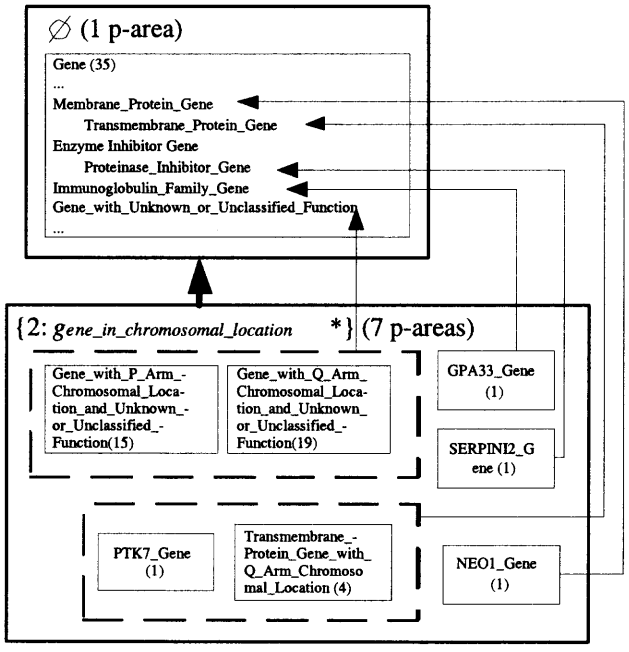


Figure 5.6 Example of new modeling of the p-area taxonomy.

The transformation is demonstrated. There are 42 concepts in {*gene_in_chromosomal_location**} (Figure 5.3). The role *gene_in_chromosomal_location* is introduced at each concept. Thus, they are divided into 42 p-areas with only one concept each. All of them are located in leaves of the Gene hierarchy. It is suggested that the targets of *gene_in_chromosomal_location* can be used to further group together many p-areas in this area. Thirty-four concepts in this area are children of GENE WITH UNKNOWN OR UNCLASSIFIED FUNCTION. Two generalizing concepts could be created as children of GENE WITH UNKNOWN

OR UNCLASSIFIED FUNCTION. One could be called GENE WITH P-ARM CHROMOSOMAL LOCATION AND UNKNOWN OR UNCLASSIFIED FUNCTION. It would introduce the chromosomal location role with P-arm as its target. Fifteen out of these 34 concepts would be children of this generalizing concept. All these 15 concepts would inherit from the new concept the target P-arm for the chromosomal location role. The other generalizing concept could be called GENE WITH Q-ARM CHROMOSOMAL LOCATION AND UNKNOWN OR UNCLASSIFIED FUNCTION. It would introduce the chromosomal location role with the target Q-arm. It would be the parent of the 19 other concepts, which would inherit from it the target Q-arm for the role chromosomal location. The other five concepts in this area are children of TRANSMEMBRANE PROTEIN GENE. Similarly, a generalizing concept, say TRANSMEMBRANE PROTEIN GENE WITH Q-ARM CHROMOSOMAL LOCATION could be created as a child of TRANSMEMBRANE PROTEIN GENE. Four out of the five concepts would be the children of this new concept TRANSMEMBRANE PROTEIN GENE WITH Q-ARM CHROMOSOMAL LOCATION. Instead of 42 concepts introducing *gene_in_chromosomal_location*, it would be introduced at only three new concepts. All the children of these three concepts would inherit this role and the proper target. As shown in Figure 5.6, three new p-areas would be induced. The number of p-areas of $\{2^*\}$ would be reduced from 42 to seven. At the same time, this transformation would reduce the role-definition complexity of the terminology.

Lowering the Number of Multiple Subsumption Concepts

There are 42 gene concepts with two parents. The names of these concepts along with their parents and areas can be found in Table 5.9. No concept has more than two parents. Concepts with multiple parents are more complex as they are several things in one. Inheriting roles from multiple parents also adds to complexity. One can define a (subsumption) "multiplicity complexity" for a terminology as the ratio of the total number of extra parents (beyond the first one) to the total number of concepts. Note that in this definition, a concept with four parents contributes more to this complexity measure than a

Table 5.9 Concepts with Multiple Parents

Concept	Concept Area	First Parent	First Parent Area	Second Parent	Second Parent Area
GRB7 Gene	{1+,2+,3+}	Adaptor Signaling Protein Gene	{3*}	Link-GEFII plus H1171Gene	{1+,2+,3+}
MADD Gene	{1+,2+,3+}	Apoptosis Promoter Gene	{3*}	Adaptor Signaling Protein Gene	{3*}
MAGED1 Gene	{1+,2+,3+}	Apoptosis Promoter Gene	{3*}	Adaptor Signaling Protein Gene	{3*}
HTATIP2 Gene	{1+,2+,3+}	Apoptosis Promoter Gene	{3*}	Transcription Coactivator Gene	{1*,3+}
BCL2 Gene	{0*,1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Inhibitor Gene	{3*}
BCL2L2 Gene	{1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Inhibitor Gene	{3*}
BAX Gene	{0*,1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
BAK1 Gene	{1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
BCL2L11 Gene	{1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
BIK Gene	{1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
BID Gene	{1+,2+,3+}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
PMAIP1 Gene	{2*,3+}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
IL8 Gene	{1+,2+,3+}	BTC Gene	{1+,2+,3+}	Interleukin Gene	{1*,3+}
TBP Gene	{1+,2+,3+}	B-TFIID Subunit Gene	{1*,3+}	TFIID Subunit Gene	{1*,3+}
ETV6 Gene	{0*,1+,2+,3+}	CREBL2 Gene	{1+,2+,3+}	ETS Family Gene	{1*,3+}
PHEMX Gene	{2*,3+}	Gene with Unknown or Unclassified Function	∅	Transmembrane Protein Gene	∅
GRAP Gene	{1+,2+,3+}	GIT1 Gene	{1+,2+,3+}	Adaptor Signaling Protein Gene	{3*}
RASD1 Gene	{1+,2+,3+}	GIT1 Gene	{1+,2+,3+}	RAS Family Gene	{1*,3+}
HLA-A Gene	{0*,1+,2+,3+}	Immunoprotein Gene	{3*}	MHC Class-I Gene	{3*}
KLRD1 Gene	{1+,2+,3+}	Immunoprotein Gene	{3*}	Receptor Gene	{3*}
Oncogene ERB-A	{1+,2+,3+}	Oncogene Transcription Factor	{1*,3+}	ERB Oncogene Family	{1*,3+}
SMARCC1 Gene	{1+,2+,3+}	Protein Complex Subunit Gene	∅	SWI-SNF-Related Gene	{3*}
SMARCC2 Gene	{1+,2+,3+}	Protein Complex Subunit Gene	∅	SWI-SNF-Related Gene	{3*}
SMARCE1 Gene	{1+,2+,3+}	Protein Complex Subunit Gene	∅	SWI-SNF-Related Gene	{3*}
IFITM1 Gene	{1+,2+,3+}	Protein Complex Subunit Gene	∅	Transmembrane Protein Gene	∅
TOP1 Gene	{1+,2+,3+}	Regulatory Gene	∅	Isomerase Gene	∅
TOP2A Gene	{1+,2+,3+}	Regulatory Gene	∅	Isomerase Gene	∅
TOP2B Gene	{1+,2+,3+}	Regulatory Gene	∅	Isomerase Gene	∅
LATS1 Gene	{1+,2+,3+}	Regulatory Gene	∅	Serine-Threonine Protein Kinase Gene	{3*}
LATS2 Gene	{1+,2+,3+}	Regulatory Gene	∅	Serine-Threonine Protein Kinase Gene	{3*}
LAMR1 Gene	{1+,2+,3+}	Ribosome Subunit Gene	{3*}	Cell Adhesion Receptor Gene	{3*}
MAP2K6 Gene	{1+,2+,3+}	Signaling Pathway Gene	{3*}	Apoptosis Promoter Gene	{3*}
GPI Gene	{1+,2+,3+}	Signaling Pathway Gene	{3*}	Isomerase Gene	∅
PTGS1 Gene	{1+,2+,3+}	Signaling Pathway Gene	{3*}	Oxidoreductase Gene	{3*}
PTGS2 Gene	{1+,2+,3+}	Signaling Pathway Gene	{3*}	Oxidoreductase Gene	{3*}
RPS6KA5 Gene	{2*,3+}	Signaling Pathway Gene	{3*}	Serine-Threonine Protein Kinase Gene	{3*}
DAXX Gene	{1+,2+,3+}	Signaling Pathway Gene	{3*}	Transcription Corepressor Gene	{3*}
MAP3K1 Gene	{1+,2+,3+}	Signaling Pathway Gene	{3*}	Ubiquitin Ligase Gene	{3*}
NR0B2 Gene	{1+,2+,3+}	Transcription Corepressor Gene	{3*}	Orphan Nuclear Receptor Gene	{3*}
MADH2 Gene	{0*,1+,2+,3+}	Transcription Factor Gene	{1*,3+}	Signaling Pathway Gene	{3*}
GPR54 Gene	{0*,1+,2+,3+}	TRIP10 Gene	{2*,3+}	G Protein-Coupled Receptor Gene	{3*}
LDLR Gene	{1+,2+,3+}	TRIP10 Gene	{2*,3+}	Receptor Gene	{3*}

concept with two parents. According to this definition, the multiplicity complexity of the Gene hierarchy is 42/1786 = 0.023.

In Table 5.9, there are six different gene concepts, e.g., BAX GENE and BAK1 GENE, which have the same two parents, BCL-2 GENE FAMILY and APOPTOSIS PROMOTER GENE (see Figure 5.7(a)). These six contribue 14% of the multiplicity complexity of the hierarchy.

Consider two category concepts that have the same set of children. One can then define a new category concept which will be a child of both given category concepts while being a parent of their original children. Such a new node is called as an intersection category concept since it conveys the category expressed by each of its children. For example, one could define an new intersection category concept as a child of the above two parents. It could be named BCL-2 APOPTOSIS PROMOTER GENE FAMILY. Then the six above mentioned gene concepts would be made children of this new intersection category concept rather than being children of their original two concepts (Figure 5.7(b)). This new configuration would stress the similarity of the six concepts and reduce the multiplicity complexity of the hierarchy. There are altogether 20 concepts, out of 42 in Table 5.9, that share both parents with another concept. By creating seven such new intersection category joint parents, the number of multiple subsumption gene concepts would be reduced to 22. At the same time, seven new subsumption category concepts would have been created for a net gain of $20 - 7 = 13$. See Table 5.10 for details of these 7 new concepts. As a result, the multiplicity complexity will be reduced by a third to 29/(1786+7) = 0.016.

5.4.3 Transfer of Concepts between Areas

It notes that out of the three kinds of role errors, only "missing role" affects the area of a concept. A concept with new roles is effectively removed from its previous area and inserted into another area with the inclusion of the proper set of roles. The other two types

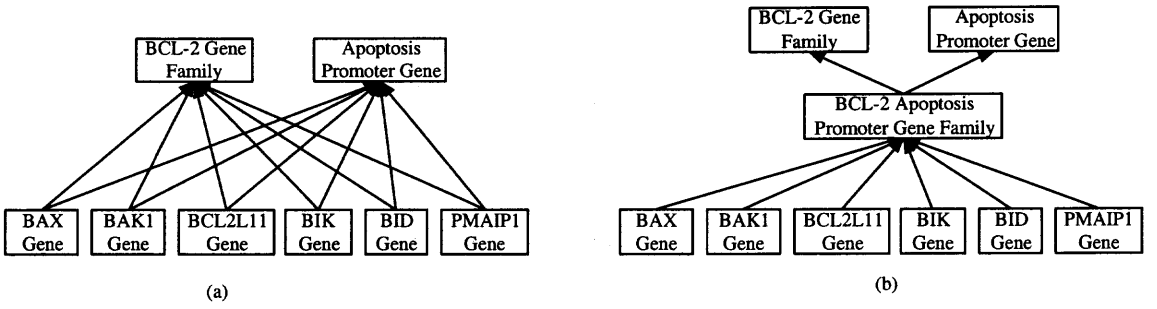


Figure 5.7 Example of a transformation of the hierarchy of concepts with multiple parents.

of errors, "missing target" and "incorrect target," do not impact the area of a concept, since the role existed already and the changes in the target do not affect the areas.

There are 362 concepts with missing roles. Their movement to new areas is shown in Table 5.11. An interesting case is the area {2*}. There are 42 concepts in this area and all of them are moved to new areas according to the addition of previously missing roles (Table 5.11).

5.5 Conclusion

A structural auditing methodology has been applied to audit role errors in the Gene hierarchy of the NCIT. The Gene hierarchy is divided into areas and p-areas, and two abstraction taxonomies (the AT and PAT) are derived. These taxonomies provide guidance for auditing priority by pointing to groups of concepts with a high likelihood of errors. The auditing conducted according to this methodology has found that about half of the concepts have role errors of three kinds: missing roles, missing targets, and incorrect targets. Error distributions have been investigated. The error percentage in small p-areas (having one or two concepts) is much higher (54%) than for larger p-areas (6%), confirming a proposed hypothesis. The error percentage for the large areas with the number of p-areas close to the number of concepts is high (above 50% of the concepts), confirming another hypothesis.

Table 5.10 List of New Intersection Category Concepts

New Concept	Area	First Parent	First Parent Area	Second Parent	Second Parent Area
Apoptosis Promoter Adaptor Signaling Protein Gene	{3*}	Apoptosis Promoter Gene	{3*}	Adaptor Signaling Protein Gene	{3*}
BCL-2 Apoptosis Inhibitor Gene Family	{3*}	BCL-2 Gene Family	{3*}	Apoptosis Inhibitor Gene	{3*}
BCL-2 Apoptosis Promoter Gene Family	{3*}	BCL-2 Gene Family	{3*}	Apoptosis Promoter Gene	{3*}
Protein Complex Subunit SWI-SNF-Related Gene	{3*}	Protein Complex Subunit Gene	∅	SWI-SNF-Related Gene	{3*}
Regulatory Isomerase Gene	∅	Regulatory Gene	∅	Isomerase Gene	∅
Regulatory Serine-Threonine Protein Kinase Gene	{3*}	Regulatory Gene	∅	Serine-Threonine Protein Kinase Gene	{3*}
Signaling Pathway Oxidoreductase Gene	{3*}	Signaling Pathway Gene	{3*}	Oxidoreductase Gene	{3*}

After correcting the role errors, when the concepts will have a more accurate structure than originally, one can apply structural auditing for other errors as well.

Table 5.11 Movement of Concepts with Missing Roles

Original Area	Another Area	# Concepts Moved
$\emptyset$	$\{3^*\}$	14
$\emptyset$	$\{2^*,3^+\}$	1
$\{1^*\}$	$\{1^*,3^+\}$	1
$\{2^*\}$	$\{0^*,2^*\}$	1
$\{2^*\}$	$\{1^*,2^*\}$	12
$\{2^*\}$	$\{1^+,2^+,3^+\}$	14
$\{2^*\}$	$\{0^*,1^+,2^+\}$	9
$\{2^*\}$	$\{0^*,1^+,2^+,3^+\}$	6
$\{2^*,3^+\}$	$\{1^+,2^+,3^+\}$	305
$\{1^*,2^*\}$	$\{1^+,2^+,3^+\}$	8
$\{0^*,2^*\}$	$\{0^*,2^+,3^+\}$	1
$\{1^*,3^+\}$	$\{1^+,2^+,3^+\}$	1
$\{2^*,5^*\}$	$\{2^*,3^*,5^*\}$	1
$\{0^*,1^+,2^+\}$	$\{0^*,1^+,2^+,3^+\}$	2
$\{1^+,2^+,3^+\}$	$\{0^*,1^+,2^+,3^+\}$	1

CHAPTER 6

SUMMARY

As a part of the development life cycle, it is necessary to audit controlled terminologies for quality assurance. The size of terminologies is typically very large and their complexity is high. It is a major challenge for the medical informatics community to carry out such auditing. Several auditing methodologies are developed based on various structural characteristics of controlled terminologies.

A methodology is designed to identify the inconsistencies in hierarchical realtionships of the UMLS. It is based on comparing the parent-child (IS-A) relationship between concepts in META to the ancestor-descendant relationship between their corresponding semantic types. The result detected that a large portion of parent-child relationships are in need of correction.

The metaschema is built up from the SN of the UMLS. It provides a higher-level abstract view of the SN. A divide and conquer approach is designed to audit the concepts of intersections of meta-semantic types in the metaschema. Concepts located in such intersections have a high likelihood of errors. This methodology has been applied successfully to the UMLS, confirming a hypothesis of higher percentage of errors for concepts in intersections of meta-semantic types.

Concepts in the NCIT are grouped into areas and p-areas based on the structural and semantic similarity. Two abstraction taxonomies, the Area Taxonomy and P-area Taxonomy, are derived from these divisions. These taxonomies can be used to guide the auditing of the terminologies as they highlight groups of concepts with potential errors. An auditing methodology is designed to identify different kinds of errors in the NCIT. The auditing results supports the two hypotheses: (1) The probability of erroneous concepts is

higher for small p-areas (2) The likelihood of errors in concepts of small p-area is higher in an area with low number of p-areas.

All of these methodologies provide some computational support for auditing. They help to focus human review on problematic groups of concepts. It is especially important for controlled terminology quality assurance due to the limitation of typically available resources. Each methodology was tailored for specific structural characteristics of some controlled terminologies. It can be only applied to a terminology satisfying these characteristics. This phenomenon limits the generality and reuse of each methodology.

Future research should follow this direction of identifying characteristics of terminologies, identify families of terminologies satisfying such characteristics and design auditing methodologies tailored to utilize these characteristics. In the future, designers of new terminologies may take into account the support offered by specific characteristics for auditing, when they choose which characteristics, their terminology should satisfy. Such approach can limit the cost for resources required for auditing to assure the quality of the terminology. The more structural auditing methodologies will be designed, the better value will be returned for the human auditing review, due to the focus on groups with high likelihood of errors.

REFERENCES

[1] Apelon. URL: http://www.apelon.com (accessed Feb,2006).

[2] F. Baader. Restricted role-value-maps in a description logic with existential restrictions and terminological cycles. In D. Calvanese, G. De Giacomo, and E. Franconi, editors, *2003 International Workshop on Description Logics DL2003*, volume 81 of *CEUR Workshop Proceedings*, pages 31–38, August 2003.

[3] A. D. Baxevanis. The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Res*, 31(1):1–12, 2003.

[4] O. Bodenreider. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. In S. Bakken, editor, *Proc 2001 AMIA Annual Symposium*, pages 57–61, Washington, DC, November 2001.

[5] O. Bodenreider. An object-oriented model for representing semantic locality in the UMLS. In *Proc Medinfo2001*, pages 161–165, London, UK, September 2001.

[6] O. Bodenreider, A. Burgun, and T. C. Rindflesch. Assessing the consistency of a biomedical terminology through lexical knowledge. *Int J Med Inf*, pages 85–95, 2002.

[7] O. Bodenreider and A. T. McCray. Exploring semantic groups through visual approaches. *Journal of Biomedical Informatics*, 36(6):414–432, December 2003.

[8] M. J. Brescia, J. E. Cimino, K. Appel, and B. J. Hurwich. Chronic hemodialysis using venipuncture and a surgically created arteriovenous fistula. *N Engl J Med*, 275(20):1089–92, November 1966.

[9] caCORE technical guide. URL: http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/caBIO/guide (accessed Feb,2006).

[10] K. E. Campbell, D. E. Oliver, and E. H. Shortliffe. The Unified Medical Language System: Toward a collaborative approach for solving terminologic problems. *JAMIA*, 5(1):12–16, 1998.

[11] W. Ceusters, B. Smith, and L. Goldberg. A terminological and ontological analysis of the NCI Thesaurus. In *Methods of Information in Medicine*, volume 44, pages 498–507, 2005.

[12] W. Ceusters, B. Smith, C. Kumar, and C. Dhaen. Mistakes in medical ontologies: Where do they come from and how can they be detected? In D. M. Pisanelli, editor, *Ontologies in Medicine: Proc. Workshop on Medical Ontologies*, pages 145–164, Rome, Italy, October 2003.

[13] W. Ceusters, B. Smith, C. Kumar, and C. Dhaen. Ontology-based error detection in SNOMED-CT. In *Proc Medinfo 2004*, pages 482–6, San Francisco, CA, September 2004.

[14] Z. Chen, Y. Perl, M. Halper, J. Geller, and H. Gu. Partitioning the UMLS Semantic Network. *IEEE Trans. Information Technology in Biomedicine*, 6(2):102–108, June 2002.

[15] J. J. Cimino. Auditing the Unified Medical Language System with semantic methods. *JAMIA*, 5(1):41–51, 1998.

[16] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37:394–403, November 1998.

[17] J. J. Cimino. Battling Scylla and Charybdis: The search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. In Bakken S, editor, *Proc 2001 AMIA Annual Symposium*, pages 120–124, Washington, DC, November 2001.

[18] J. J. Cimino, P. D. Clayton, G. Hripcsak, and S. B. Johnson. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *JAMIA*, 1(1):35–50, 1994.

[19] J. J. Cimino, H. Min, and Y. Perl. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *Journal of Biomedical Informatics*, 36(6):450–461, December 2003.

[20] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, L. Walters, et al. New goals for the U.S. human genome project: 1998–2003. *Science*, 282:682–689, 1998.

[21] R. Cornet and A. Abu-Hanna. Using non-primitive concept definitions for improving dl-based knowledge base. In V. Haarslev and R. Möller, editors, *2004 International Workshop on Description Logics - DL2004*, pages 138–147, Whistler, Canada, June 2004.

[22] N. F. de Keizer, A. Abu-Hanna, R. Cornet, J. H. Zwersloot-Schonk, and C. P. Stoutenbeek. Analysis and design of an ontology for intensive care diagnoses. *Methods of Information in Medicine*, 38(2):102–12, June 1999.

[23] B. G. Druss and S. C. Marcus. Growth and decentralization of the medical literature: implications for evidence-based medicine. *J Med Libr Assoc.*, 93(4):499–501, 2005.

[24] G.O. Consortium. Creating the Gene Ontology resource: Design and Implementation. Genome Res. volume 11, pages 1425–1433, 2001.

[25] L. Goldstein, D. Schneider, and M. Siegel. *Finite mathematics and its applications.* Prentice-Hall, 2003.

[26] H. Gu, M. Halper, J. Geller, and Y. Perl. Benefits of an object-oriented database representation for controlled medical terminologies. *JAMIA*, 6(4):283–303, July/August 1999.

[27] H. Gu, Y. Li, and S. Haque. Discovering inconsistencies in UWDA – a UMLS source vocabulary. In *Proceedings of the Int'l Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS'03)*, pages 300–304, Las Vegas, June 2003.

[28] H. Gu, Y. Perl, G. Elhanan, H. Min, L. Zhang, and Y. Peng. Auditing concept categorizations in the UMLS. *Artificial Intelligence in Medicine*, 31(1):29–44, 2004.

[29] H. Gu, Y. Perl, J. Geller, M. Halper, L. Liu, and J. J. Cimino. Representing the UMLS as an OODB: Modeling issues and advantages. *JAMIA*, 7(1):66–80, January/February 2000. Selected for reprint in R. Haux, C. Kulikowski, eds.: Yearbook of Medical Informatics, International Medical Informatics Association, Rotterdam, 2001: 271–285.

[30] H. Gu, Y. Perl, M. Halper, J. Geller, F. Kuo, and J. J. Cimino. Partitioning an object-oriented terminology schema. *Methods of Information in Medicine*, 40(3):204–212, July 2001.

[31] M. Halper, Z. Chen, J. Geller, and Y. Perl. A metaschema of the UMLS based on a partition of its Semantic Network. In S. Bakken, editor, *Proc. 2001 AMIA Annual Symposium*, pages 234–238, Washington, DC, November 2001.

[32] W. T. Hole and S. Srinivasin. Discovering missed synonymy in a large concept-oriented metathesaurus. *JAMIA*, 7:354–358, 2000.

[33] B. L. Humphreys and D. A. Lindberg. Building the Unified Medical Language System. In *Proc. Thirteenth Annual Symposium on Computer Applications in Medical Care*, pages 475–480, Washington, DC, November 1989.

[34] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett. The Unified Medical Language System: An informatics research collaboration. *JAMIA*, 5(1):1–11, 1998.

[35] Z. E. Karanjawala and F. S. Collins. Genetics in the context of medical practice. *JAMA*, 280(17):1533–1534, 1998.

[36] A. Kumar and B. Smith. The Unified Medical Language System and the Gene Ontology: Some critical reflections. In A. Günter, R. Kruse, and B. Neumann, editors, *KI 2003, Advances in Artificial Intelligence (Lecture Notes in Artificial Intelligence 2821)*, pages 135–148, Berlin: Springer, 2003.

[37] A. Kumar and B. Smith. Oncology ontology in the nci thesaurus. In S. Miksch, J. Hunter, and E. T. Keravnou, editors, *AIME*, pages 213–220, 2005.

[38] A. Kumar, B. Smith, and C. Borgelt. Dependence relationships between Gene Ontology terms based on TIGR gene product annotations. In S. Ananadiou and P. Zweigenbaum, editors, *CompuTerm 2004, 3rd International Workshop on Computational Terminology*, pages 31–38, Coling, Geneva, August 2004.

[39] J. H. Lin. Divining and altering the future: Implications from the human genome project. *Science*, 282:1532, 1998.

[40] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The Unified Medical Language System. *Methods of Information in Medicine*, 32:281–291, 1993.

[41] J. Lomax and A. T. McCray. Mapping the Gene Ontology into the Unified Medical Language System. *Comparative and Functional Genomics*, 5(5):354–361, 2004.

[42] A. T. McCray. Representing biomedical knowledge in the UMLS Semantic Network. In N. C. Broering, editor, *High-Performance Medical Libraries: Advances in Information Management for the Virtual Era*, pages 45–55, Mekler, Westport, CT, 1993.

[43] A. T. McCray. An upper-level ontology for the biomedical domain. *Comp Funct Genom*, 4:80–84, 2003.

[44] A. T. McCray, A. Burgun, and O. Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. In R. Rogers, R. Haux, and V. Patel, editors, *Proc. Medinfo 2001*, pages 171–175, London, UK, September 2001. Amsterdam: IOS Press.

[45] A. T. McCray and W. T. Hole. The scope and structure of the first version of the UMLS Semantic Network. In *Proc. Fourteenth Annual SCAMC*, pages 126–130, Los Alamitos, CA, November 1990.

[46] A. T. McCray and S. J. Nelson. The representation of meaning in the UMLS. *Methods of Information in Medicine*, 34:193–201, 1995.

[47] C. J. McDonald, S. M. Huff, and J. G. Suico et al. LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry*, 49(4):624–633, 2003.

[48] H. Min, Y. Perl, Y. Chen, M. Halper, J. Geller, and Y. Wang. Auditing as part of the terminology design life cycle. Submitted for journal publication.

[49] M. Minsky. A framework for representing knowledge. *MIT-AI Laboratory Memo*, 306, June 1974.

[50] D. Nardi and R. J. Brachman. An introduction to description logics. *The Description Logic Handbook*, pages 5–44, 2002.

[51] URL: http://nciterms.nci.nih.gov/NCIBrowser/Startup.do (accessed Feb,2006).

[52] Y. Peng, M. Halper, Y. Perl, and J. Geller. Auditing the UMLS for redundant classifications. In *Proc. 2002 AMIA Annual Symposium*, pages 612–616, San Antonio, TX, November 2002.

[53] Y. Perl, C. Chen, M. Halper, J. Geller, L. Zhang, and Y. Peng. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. *Journal of Biomedical Informatics*, 35(3):194–212, June 2002.

[54] Y. Perl, Z. Chen, M. Halper, J. Geller, L. Zhang, and Y. Peng. The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network. *Journal of Biomedical Informatics*, 35(3):194–212, June 2003.

[55] Protégé. URL: http://protege.stanford.edu (accessed Feb,2006).

[56] C. Rosse and J. L. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, 2003.

[57] S. Schulze-Kremer, B. Smith, and A. Kumar. Revising the UMLS Semantic Network. In M. Fieschi, E. Coiera, and Y.-C. Li, editors, *Proc Medinfo 2004*, San Francisco, CA, September 2004. 1700.

[58] B. Smith, J. Köhler, and A. Kumar. On the application of formal principles to life science data: A case study in the Gene Ontology. In *DILS 2004 (Data Integration in the Life Sciences), (Lecture Notes in Bioinformatics 2994)*, pages 79–94, Berlin: Springer, 2004.

[59] B. Smith and A. Kumar. On controlled vocabularies in bioinformatics: A case study in the Gene Ontology. In *BIOSILICO: Information Technology in Drug Discovery*, volume 2, pages 246–252, 2004.

[60] B. Smith, J. Williams, and S. Schulze-Kremer. The Ontology of the Gene Ontology. In M. Musen, editor, *Proc 2003 AMIA Annual Symposium*, pages 609–613, Washington, DC, November 2003.

[61] SNOMED International: The Systematized Nomenclature of Medicine. URL: http://www.snomed.org (accessed Feb,2006).

[62] Software quality assurance. URL: http://satc.gsfc.nasa.gov/assure/assurepage.html (accessed Feb,2006).

[63] M. S. Tuttle, W. G. Cole, D. D. Sherertz, and S. J. Nelson. Navigating to knowledge. *Methods of Information in Medicine*, 34:214–31, 1995.

[64] U. S. Dept. of Health and Human Services, National Institutes of Health, National Library of Medicine. Unified Medical Language System (UMLS), 2002.

[65] National Library of Medicine. UMLS knowledge sources experimental edition. Bethesda, MD: The Library, updated annually.

[66] P. R. Watkins and L. B. Eliot, editors. *Expert systems in business and finance: issues and applications*. John Wiley & Sons, Inc., New York, NY, USA, 1993.

[67] L. Zhang, Y. Perl, M. Halper, J. Geller, and J. J. Cimino. Enriching the structure of the umls semantic network. In *Proc AMIA Symp*, pages 939–943, 2002.