Copyright Warning & Restrictions

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

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

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

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

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

ABSTRACT

APPLICATIONS OF BIG KNOWLEDGE SUMMARIZATION

by
Ling Zheng

Advanced technologies have resulted in the generation of large amounts of data ("Big Data"). The Big Knowledge derived from Big Data could be beyond humans' ability of comprehension, which will limit the effective and innovative use of Big Knowledge repository. Biomedical ontologies, which play important roles in biomedical information systems, constitute one kind of Big Knowledge repository. Biomedical ontologies typically consist of domain knowledge assertions expressed by the semantic connections between tens of thousands of concepts. Without some high-level visual representation of Big Knowledge in biomedical ontologies, humans cannot grasp the "big picture" of those ontologies. Such Big Knowledge orientation is required for the proper maintenance of ontologies and their effective use. This dissertation is addressing the Big Knowledge challenge – How to enable humans to use Big Knowledge correctly and effectively (referred to as the "Big Knowledge to Use" (BK2U) problem) – with a focus on biomedical ontologies.

In previous work, Abstraction Networks (AbNs) have been demonstrated successful for the summarization, visualization and quality assurance (QA) of biomedical ontologies. Based on the previous research, this dissertation introduces new AbNs of various granularities for Big Knowledge summarization and extends the applications of AbNs. This dissertation consists of three main parts. The first part introduces two advanced AbNs. One is the *weighted aggregate partial-area taxonomy* with a parameter

to flexibly control the summarization granularity. The second is the Ingredient Abstraction Network (IAbN) for the National Drug File – Reference Terminology (NDF-RT) *Chemical Ingredients* hierarchy, for which the previously developed AbNs for hierarchies with outgoing relationships, are not applicable. Since NDF-RT's *Chemical Ingredients* hierarchy has no outgoing relationships.

The second part describes applications of the two advanced AbNs. A study utilizing the weighted aggregate partial-area taxonomy for the identification of major topics in SNOMED CT's *Specimen* hierarchy is reported. A multi-layer interactive visualization system of required granularity for ontology comprehension, based on the weighted aggregate partial-area taxonomy, is demonstrated to comprehend the *Neoplasm* subhierarchy of National Cancer Institute thesaurus (NCIt). The IAbN is applied for drug-drug interaction (DDI) discovery.

The third part reports eight family-based QA studies on NCIt's *Neoplasm*, *Gene*, and *Biological Process* hierarchies, SNOMED CT's *Infectious disease* hierarchy, the Chemical Entities of Biological Interest ontology, and the *Chemical Ingredients* hierarchy in NDF-RT. There is no one-size-fits-all QA method and it is impossible to find a QA method for each individual ontology. Hence, family-based QA is an effective way, i.e., one QA technique could be applicable to a whole family of structurally similar ontologies. The results of these studies demonstrate that "complex concepts" and "uncommonly modeled concepts" are more likely to have errors. Furthermore, the three studies on overlapping concepts in partial-area taxonomies reported in this dissertation combined with previous three studies prove the success of "overlapping concepts" as a QA methodology for a whole family of 76 similar ontologies in BioPortal.

APPLICATIONS OF BIG KNOWLEDGE SUMMARIZATION

by
Ling Zheng

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

Department of Computer Science

August 2018

APPROVAL PAGE

APPLICATIONS OF BIG KNOWLEDGE SUMMARIZATION

Ling Zheng

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

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

Dr. James A. McHugh, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Michael Halper, Committee Member — Date
Professor of Informatics and IT Division Director, NJIT

Dr. Huanying (Helen) Gu, Committee Member — Date
Professor of Computer Science, NYIT

Dr. Mei Liu, Committee Member — Date
Assistant Professor of Internal Medicine, University of Kansas Medical Center

BIOGRAPHICAL SKETCH

Author: Ling Zheng

Degree: Doctor of Philosophy

Date: August 2018

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2018

- Master of Science in Biomedical Engineering,
  Zhejiang University, Hangzhou, P. R. China, 2012

- Bachelor of Science in Biomedical Engineering,
  Southern Medical University, Guangzhou, P. R. China, 2009

Major: Computer Science

Publications:

Published Journal Papers

Zheng L, Chen Y, Elhanan G, Perl Y, Geller J, Ochs C. Complex overlapping concepts: an effective auditing methodology for families of similarly structured BioPortal ontologies. Journal of biomedical informatics. 2018;83:135-149.

Halper M, Perl Y, Ochs C, Zheng L. Taxonomy-based approaches to quality assurance of ontologies. Journal of healthcare engineering. 2017;10.1155/2017/3495723.

Zheng L, Yumak H, Chen L, Ochs C, Geller J, Kapusnik-Uner J, et al. Quality assurance of chemical ingredient classification for the National Drug File - Reference Terminology. Journal of biomedical informatics. 2017;73:30-42.

Zheng L, Min H, Chen Y, Xu J, Geller J, Perl Y. Auditing National Cancer Institute thesaurus neoplasm concepts in groups of high error concentration. Applied Ontology. 2017;12(2):113-130.

Min H, Zheng L, Perl Y, Halper M, De Coronado S, Ochs C. Relating complexity and error rates of ontology concepts. More complex NCIt concepts have more errors. Methods of Information in Medicine. 2017;56(3):200-208.

Perl Y, Geller J, Halper M, Ochs C, Zheng L, Kapusnik-Uner J. Introducing the Big Knowledge to Use (BK2U) challenge. Annals of the New York Academy of Sciences. 2017;1387(1):12-24.

Ochs C, He Z, Zheng L, Geller J, Perl Y, Hripcsak G, et al. Utilizing a structural meta-ontology for family-based quality assurance of the BioPortal ontologies. Journal of biomedical informatics. 2016;61:63-76.

Journal Papers in Process

Chen Y, Zheng L, Perl Y, Halper M, De Coronado S. Quality assurance of concept roles in the National Cancer Institute thesaurus. Artificial intelligence in medicine. Submitted for review.

Yumak H, Chen L, Zheng L, Halper M, Perl Y. Quality assurance analysis of ChEBI concepts based on relationship types. Journal of biomedical informatics. Submitted for review.

Published Conference Papers

Zheng L, Min H, Perl Y, Geller J. Discovering additional complex NCIt gene concepts with high error rate. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017:653-657.

Zheng L, Perl Y, Elhanan G, Ochs C, Geller J, Halper M. Summarizing an ontology: a "Big Knowledge" coverage approach. Studies in health technology and informatics. 2017;245:978-982.

Zheng L, Ochs C, Geller J, Liu H, Perl Y, De Coronado S. Multi-layer Big Knowledge visualization scheme for comprehending neoplasm ontology content. 2017 IEEE International Conference on Big Knowledge (ICBK). 2017:127-134.

Yumak H, Chen L, Halper M, Zheng L, Perl Y, Elhanan G. A quality-assurance study of ChEBI. 2016 International Conference on Biological Ontology & BioCreative.

Ochs C, Zheng L, Gu H, Perl Y, Geller J, Kapusnik-Uner J, et al. Drug-drug interaction discovery using abstraction networks for "National Drug File - Reference Terminology" Chemical Ingredients. AMIA Annual Symposium Proceedings. 2015;2015:973-982.

Accepted Conference Papers

Zheng L, Liu H, Perl Y, Geller J, Ochs C, Case JT. Overlapping complex concepts have more commission errors, especially in intensive terminology auditing. 2018 AMIA Annual Symposium Proceedings.

Liu H, Chen L, Zheng L, Perl Y, Geller J. A quality assurance methodology for ChEBI Ontology focusing on uncommonly modeled concepts. 2018 International Conference on Biological Ontology. August 7-10, 2018, Corvallis, Oregon, USA.

Liu H, Zheng L, Perl Y, Geller J, Elhanan G. Can a Convolutional Neural Network support auditing of NCI thesaurus neoplasm concepts? 2018 International Conference on Biological Ontology. August 7-10, 2018, Corvallis, Oregon, USA.

Posters

Liu H, Zheng L, Perl Y, Chen Y, Elhanan G. Correcting ontology errors simplifies visual complexity. Studies in health technology and informatics. 2017;245:1330.

Zheng L, Perl Y, Geller J, Elhanan G. How to summarize Big Knowledge subjects. 2016 International Conference on Biological Ontology & BioCreative.

Presentations:

Discovering Additional Complex NCIt Gene Concepts with High Error Rate. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas City, MO, USA, November 14, 2017.

Summarizing an Ontology: A “Big Knowledge” Coverage Approach. MedInfo 2017, Hangzhou, China, August 25, 2017.

Multi-layer Big Knowledge Visualization Scheme for Comprehending Neoplasm Ontology Content. 2017 IEEE International Conference on Big Knowledge (ICBK), Anhui, China, August 9, 2017.

A Quality-Assurance Study of ChEBI. 2016 International Conference on Biological Ontology & BioCreative. Corvallis, OR, USA, August 4, 2016.

To my beloved parents, Yongman Zheng and Xiuyu Zhang

谨以此博士论文献给我挚爱的父母郑永满先生和张秀玉女士

ACKNOWLEDGMENT

I would like to express my deepest appreciation to my dissertation co-advisors Dr. Yehoshua Perl and Dr. James Geller for their continuous guidance and encouragement. Their expertise and professional dedication to scientific research has been having a profound effect upon me. The journey of being their Ph.D. student is amazing and precious during my whole life. I also would like to thank Dr. Michael Halper for his help on some of the research projects described in the dissertation. I am thankful to Dr. James McHugh, Dr. Huanying Gu and Dr. Mei Liu for serving on my dissertation committee.

The research projects in this dissertation were partially supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779.

Special thanks are given to Dr. Christopher Ochs, Dr. Gai Elhanan, Dr. Yan Chen, Dr. Hua Min, Dr. Ling Chen, Dr. Hasan Yumak, Dr. Julia Xu, and many other collaborators for all of their contributions to the work in the dissertation. I also would like to thank Prof. Janet Bodner for her continuous help with my English.

Lastly, I want to thank my parents, Yongman Zheng and Xiuyu Zhang; my brother, Long Zheng; my friends, Xiang Ji, Yanfei Liu, Hao Liu, Xiaowei Shang, Hongxiang Niu, Xiaohan Yang, and Vipina Kuttichi Keloth; Dr. Ning Deng and Dr. Li Wang, who are very important for my success in Zhejiang University; and my favorite musician, Jinyoung Jung for their support and encouragement.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(**Continued**)

Table **Page**

LIST OF TABLES
(**Continued**)

LIST OF TABLES
(**Continued**)

LIST OF FIGURES

LIST OF FIGURES
(**Continued**)

LIST OF FIGURES
(**Continued**)

Figure **Page**

LIST OF FIGURES
(**Continued**)

LIST OF FIGURES
(**Continued**)

LIST OF FIGURES
(**Continued**)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Motivation

The purpose of the "Big Data to Knowledge" (BD2K) initiative launched by the US National Institutes of Health in 2014 is to develop methodologies and techniques for extracting new knowledge hidden in large amounts of biomedical data [1]. However, if the resulting knowledge stored in a knowledge repository is too much for humans' comprehension, it is impossible for humans to make effective or innovative use of the knowledge. According to Perl et al. [2], knowledge that is so big that humans cannot easily comprehend it is defined as "Big Knowledge."

There are various kinds of knowledge. This dissertation concentrates on large biomedical ontologies, a special kind of knowledge repository typically consisting of many thousands of domain knowledge assertions. Concepts and relationships are two essential elements to represent knowledge in ontologies. A concept represents a unique entity in a domain. Concepts are linked by hierarchical IS-A relationships and lateral relationships ("relationships" for short). The hierarchical IS-A relationship between two concepts represents a concept that is a specification of the other concept. For example, *Neoplasm* IS-A *Disease or Disorder*, because the concept *Neoplasm* is more specific than the concept *Disease or Disorder*. The lateral relationships are used to define the semantics in the domain. For example, the concept *Breast Neoplasm* is connected to the concept *Breast* through the lateral relationship *Disease Has Associated Anatomic Site* to define the anatomic location of breast neoplasm. As an example of large biomedical

ontologies, National Cancer Institute Thesaurus (NCIt) [3], the most famous cancer-focused biomedical terminology, has more than 100,000 active concepts connected by more than 400,000 relationships. Hence, large biomedical ontologies are complex networks due to the large number of concepts and relationships respectively represented by nodes and links in the networks. Figure 1.1 demonstrates the complexity of a large biomedical ontology and the difficulty for humans to comprehend such Big Knowledge. Figure 1.1 only shows a small part of NCIt, i.e., 8,445 neoplasm-related concepts (7.8% of the complete NCIt).

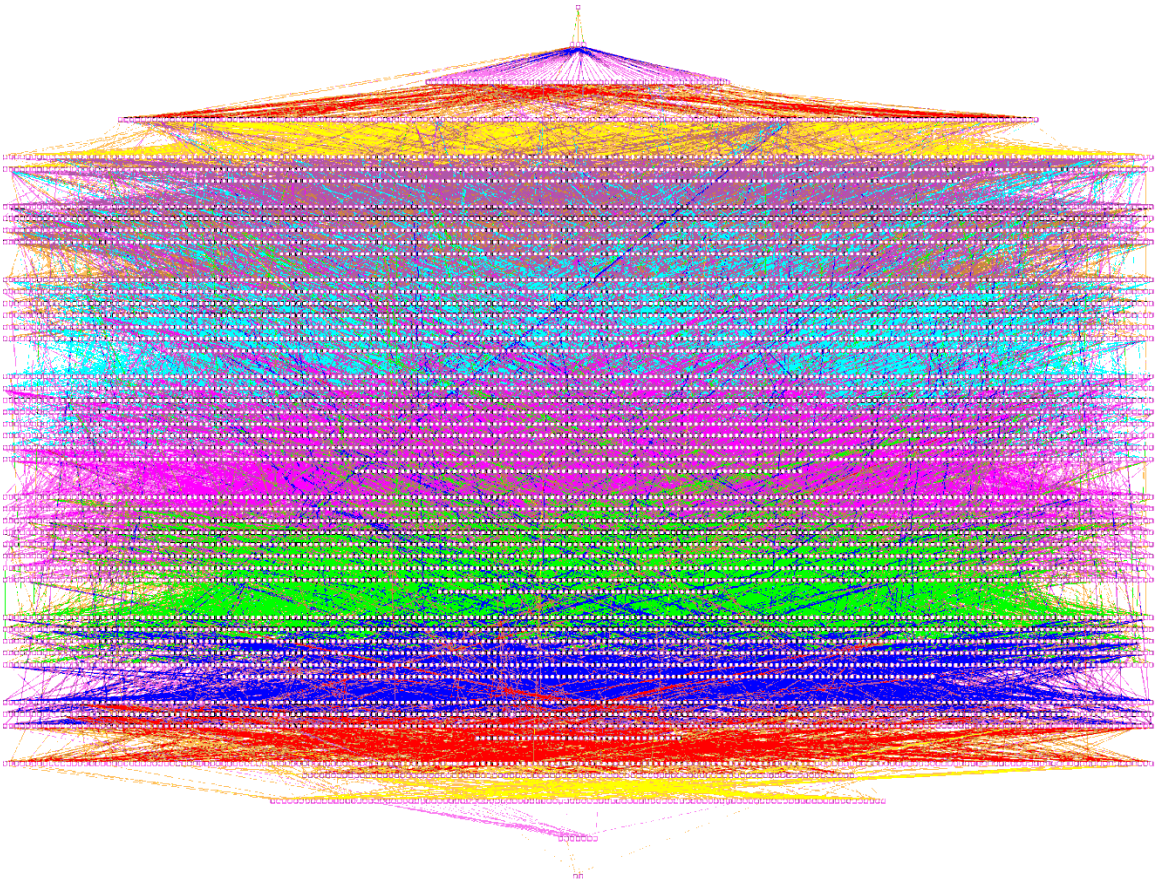

Figure 1.1 The 8,445 neoplasm concepts from NCIt. Concepts are drawn as white boxes organized into levels according to their longest-path distance from the root concept (i.e., *Neoplasm*). Only 14,420 hierarchical relationships are shown (as color-coded lines, based on the level of the child concept). At this scale, "white boxes" appear as white dots.

In general, humans typically comprehend complex knowledge by summaries and visual representations. Without some smart techniques, e.g., summarization and visualization tools to assist humans' high-level mental comprehension, even the curators of large biomedical ontologies cannot see the "big picture" of their ontologies. It would be even more difficult for external users who utilize large biomedical ontologies, to develop applications using such an ontology. Hence, this dissertation is trying to address the new challenge after BD2K: How to enable humans to use Big Knowledge correctly and effectively (referred to as the "Big Knowledge to Use" (BK2U) problem [2]).

The Structural Analysis of Biomedical Ontologies Center (SABOC) has developed various kinds of Abstraction Networks (AbNs) to support the summarization, visualization and quality assurance (QA) of biomedical ontologies [4]. An Abstraction Network derived from an ontology is itself a compact summary network consisting of "nodes," each representing a set of concepts that are similar in their structure and semantics. Thus, the Abstraction Network summarizes the structure and content of the ontology. Two basic kinds of AbNs are area taxonomy and partial-area taxonomy, which have been developed for various biomedical ontologies [5, 6] (e.g., NCIt [3] and SNOMED CT [7]).

Based on the previous work on AbNs conducted at SABOC, this dissertation presents advanced AbNs to summarize and visualize the content of Big Knowledge in large biomedical ontologies to support BK2U. The Big Knowledge summarization and visualization technique was demonstrated to be useful for the identification of major topics in a large ontology, as a multi-layer Big Knowledge visualization scheme for ontology comprehension, for the discovery of drug-drug interactions and also for the

quality assurance of large biomedical ontologies.

In a multi-year research program on quality assurance of biomedical ontologies, the SABOC team has observed that no one-size-fits-all QA method exists. However, Ochs et al. [8] have developed a family-based approach to ontology QA, where the same QA method can be applicable to most members of one family, while different families need different approaches. In order to demonstrate such scaling of a method to most ontologies of a family, the effectiveness of a method has to be demonstrated for "six out of six" members of the family to allow drawing a conclusion about the whole family. Thus, this dissertation also contains several quality assurance studies on different large biomedical ontologies utilizing different QA techniques to achieve the goal of showing the effectiveness of family-based QA for biomedical ontologies.

1.2 Dissertation Overview

Chapter 2 provides background information on biomedical ontologies used in this dissertation, i.e., SNOMED CT, National Drug File-Reference Terminology (NDF-RT), NCIt and Chemical Entities of Biological Interest (ChEBI). Chapter 2 also introduces the Abstraction Networks for biomedical ontologies developed by the SABOC team, the previous quality assurance studies of biomedical ontologies based on the Abstraction Networks, and a brief review of other QA studies for biomedical ontologies.

Chapter 3 describes two advanced Abstraction Networks, *weighted aggregate partial-area taxonomy* that provides a more compact summary of biomedical ontologies compared with a partial-area taxonomy, and *Ingredient Abstraction Network* (IAbN) to

summarize NDF-RT's *Chemical Ingredient* hierarchy, due to that terminology's unique modeling structure.

Chapter 4 presents two applications of the weighted aggregate partial-area taxonomy. One is the identification of major topics in an ontology, which was successfully demonstrated on SNOMED CT's *Specimen* hierarchy. The other is a multi-layer multi-granularity visualization scheme based on the weighted aggregate partial-area taxonomy, which was applied to comprehend the NCIt's *Neoplasm* subhierarchy. This chapter also includes one application of the Ingredient Abstraction Network, namely for Drug-Drug Interaction discovery.

Chapter 5 reports several family-based quality assurance studies in the framework of Abstraction Networks for NCIt's *Neoplasm* hierarchy, *Gene* hierarchy, and *Biological Process* hierarchy, for SNOMED CT's *Infectious Disease* hierarchy, for the ChEBI ontology, and for NDF-RT's *Chemical Ingredients* hierarchy. The results confirmed that two characterizations of concepts – complex concepts and uncommonly modeled concepts – which can be automatically identified by Abstraction Networks – are more likely to have errors than other concepts. The QA results of these studies in Chapter 5 pave the way to the family-based QA approach for biomedical ontologies. Chapter 6 concludes this dissertation.

The studies in this dissertation have been published in journals and proceedings of conferences on biomedical informatics. The weighted aggregate partial-area taxonomy in Section 3.1 and its application to major topic identification in Section 4.1 were published in the Proceedings of the 16th World Congress on Medical and Health Informatics (MedInfo 2017) [9]. The multi-layer visualization scheme in Section 4.2 was published in

the Proceedings of the 2017 IEEE International Conference on Big Knowledge (ICBK) [10]. The application of the Ingredient Abstraction Network to drug-drug interaction discovery was published in the 2015 AMIA Annual Symposium Proceedings [11] and in an extended form in the Annals of the New York Academy of Sciences [2]. Most of the QA studies in Chapter 5 have been published in the Journal of Biomedical Informatics [12, 13], the Methods of Information in Medicine [14], the Applied Ontology [15], and the Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) [16].

CHAPTER 2

BACKGROUND

2.1 Biomedical Ontologies

In recent years, ontologies have played an important role in the biomedical field to support the rapid increase of data processing in healthcare and basic research [17]. Biomedical ontologies have been used for data annotation, information integration, knowledge discovery and other applications [18-22]. This section will introduce several relevant and important biomedical ontologies.

2.1.1 SNOMED CT

SNOMED CT (SNOMED Clinical Terms) [7] is the most comprehensive clinical terminology, used in more than fifty countries in the world, providing multiple language versions. It is maintained and distributed by an international non-profit organization named SNOMED International which is the trading name of the International Health Terminology Standards Development Organisation (IHTSDO) [23]. SNOMED CT covers a wide range of clinical specialties, disciplines and requirements so that it enables consistent and processable representation of clinical content in electronic health records [24] and facilitates the semantic interoperability of health records.

Concepts are SNOMED CT's basic components to represent healthcare data [25]. SNOMED CT's concepts are organized into 19 top-level hierarchies (e.g., *Clinical finding* and *Specimen*) through IS-A relationships. A concept may have multiple parents in a hierarchy, i.e., a concept may have multiple IS-A relationships pointing to other concepts in the same hierarchy. The lateral relationships provide formal definitions for

concepts. There were about 316,840 active concepts connected by more than 574,000 IS-A hierarchical relationships and about 960,000 lateral relationships in SNOMED CT's July 2015 version. Figure 2.1 shows an excerpt of 15 concepts from the *Specimen* hierarchy with 1,620 concepts. The concept *Bile specimen* has three parents *Specimen from digestive system*, *Body substance sample* and *Fluid sample*. The lateral relationship *Specimen source topography* in the dashed green box defines the body structure where a specimen comes from. For example, the concept *Specimen from liver* has a lateral relationship *Specimen source topography* linking it to *Liver structure*.

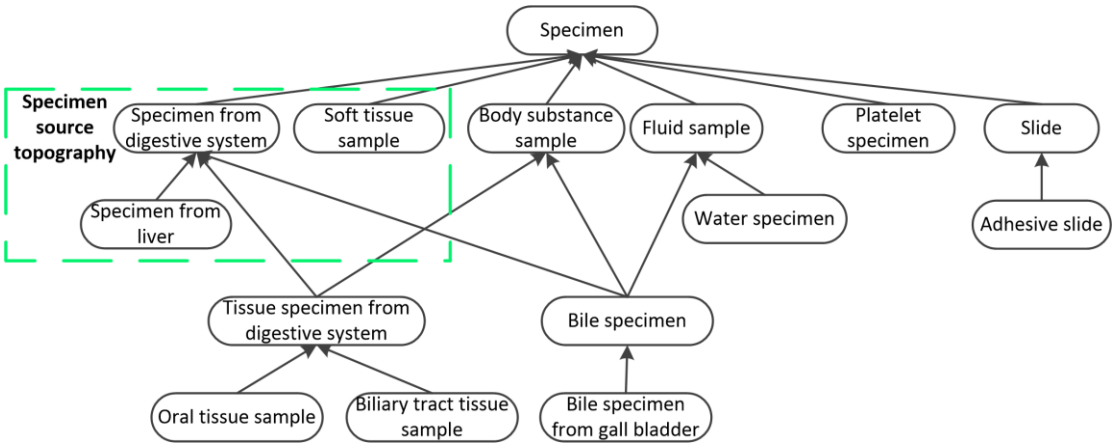

Figure 2.1 Excerpt of 15 concepts from the *Specimen* hierarchy of SNOMED CT. Concepts represented by boxes with rounded corners are connected by IS-A relationships shown as upward arrows. Each of the three concepts *Specimen from digestive system*, *Soft tissue sample*, and *Specimen from liver* enclosed in the dashed green box has a lateral relationship *Specimen source topography* with the corresponding values *Structure of digestive system*, *Soft tissues*, and *Liver structure*, respectively, in the *Body structure* hierarchy.

2.1.2 National Drug File – Reference Terminology (NDF-RT)

National Drug File – Reference Terminology (NDF-RT) is a drug terminology developed and maintained by the U.S. Department of Veterans Affairs (VA), Veterans Health Administration (VHA). NDF-RT is a formal representation of the VHA National Drug

File (NDF) [26], which is a drug classification hierarchy used to group orderable drug products into one of 579 drug classes. NDF-RT is used to support clinical applications at the VHA's clinical centers.

NDF-RT uses a description logic-based reference model to define drugs in the *Pharmaceutical Preparations* (*PP*) hierarchy according to multiple aspects (other hierarchies) [27]. These aspects include the *Chemical Ingredients* (*CI*) hierarchy, describing the chemical ingredients of drugs, the *Cellular or Molecular Interactions* (*MoA*) hierarchy, describing the drug effects at molecular, subcellular, or cellular levels, the *Physiological Effects* (*PE*) hierarchy, describing drug effects at tissue, organ, or system levels, the *Clinical Kinetics* (*PK – from Pharmacokinetics*) hierarchy, describing the absorption, distribution, and elimination of drugs, and the *Therapeutic Categories* (*TC*) hierarchy, which is an experimental hierarchy exclusively used to model FDA established pharmacologic class concepts to describe general therapeutic intents of drugs. Two more hierarchies are the *Diseases, Manifestations or Physiologic States* (*Disease*) hierarchy, describing the therapeutic, preventative, or diagnostic indications of drugs, and the *Dose Forms* hierarchy, describing the dose forms of drugs.

The *MoA*, *PE* and *CI* hierarchies were initially created by matching VHA drug ingredient names to terms from the National Library of Medicine's Medical Subject Headings (MeSH) [28]. Specifically, the *CI* hierarchy was derived from MeSH's *Chemicals and Drugs Category* and the *MoA* and *PE* hierarchies were created by extending and restructuring selected Pharmacologic Actions associated with ingredients in MeSH. Concepts in the *Disease* hierarchy were included from MeSH's *Diseases Category* [27, 29]. The purpose of developing the MeSH was to support the classification

of biomedical publications in the PubMed system [30] of the National Library of Medicine.

NDF-RT is available for download in Apelon DTS format at the National Cancer Institute's Enterprise Vocabulary Services (EVS) website [31]. NDF-RT is also released as part of the Unified Medical Language System (UMLS) [32] and it is available for download at the National Center for Biomedical Ontology (NCBO) BioPortal [33]. NDF-RT organizes concepts around the *PP* hierarchy (the triangle in Figure 2.2), which is the largest hierarchy in NDF-RT with 25,759 concepts (59.4% of the 43,397 NDF-RT concepts in the June 2015 version). The root concept of the *PP* hierarchy is *Pharmaceutical Preparations*. Besides IS-A relationships, concepts in the *PP* hierarchy can have role relationships (represented by the arrows in Figure 2.2) pointing to concepts in the other hierarchies (the seven rectangles in Figure 2.2). Role relationships (corresponding to lateral relationships) are used to define drugs according to their various aspects. Drug-disease relationships were mined from co-occurrence data in the UMLS [34] (see Figure 2.2). The *TC* hierarchy is exclusively used for concepts established by the FDA, so there are no NDF-RT asserted roles between the *PP* hierarchy and the *TC* hierarchy and the arrow in Figure 2.2 is not labeled with any NDF-RT asserted role.

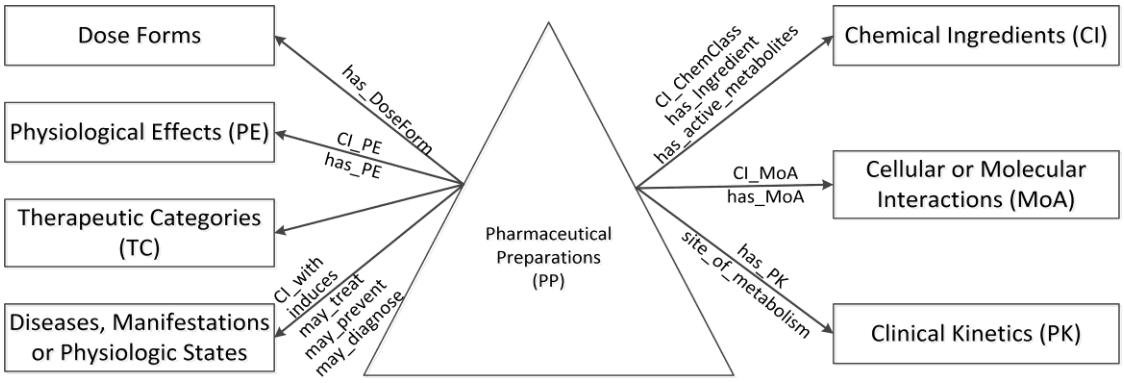

Figure 2.2 Content Model of NDF-RT [27] (The "CI" in role names means contraindicated, not Chemical Ingredient).

For example, in Figure 2.3, the drug preparation *ASPIRIN* in the *PP* hierarchy has the role relationship *has_Ingredient* pointing to the chemical ingredient *Aspirin* in the *CI* hierarchy, the second largest hierarchy in NDF-RT with 10,145 concepts. The role relationships of drug classes and drug preparations are inherited by orderable drug products, e.g., *ASPIRIN 300MG TAB* (a VA Product) inherits the role relationship *has_Ingredient* and its target concept *Aspirin* from its parent drug preparation *ASPIRIN*.

Concepts in each of the above hierarchies are organized as a generalization hierarchy; higher level concepts are more general than lower level concepts. Concepts may have multiple parents. For example, *ASPIRIN* in the *PP* hierarchy and *Salicylates* in the *CI* hierarchy each have two parents in Figure 2.3.

Concepts in the *PP* hierarchy may have different types of role relationships to concepts in the *CI* hierarchy. These role relationships are introduced at a drug class level or a drug preparation level. For example, the role relationship *has_Chemical_Structure*

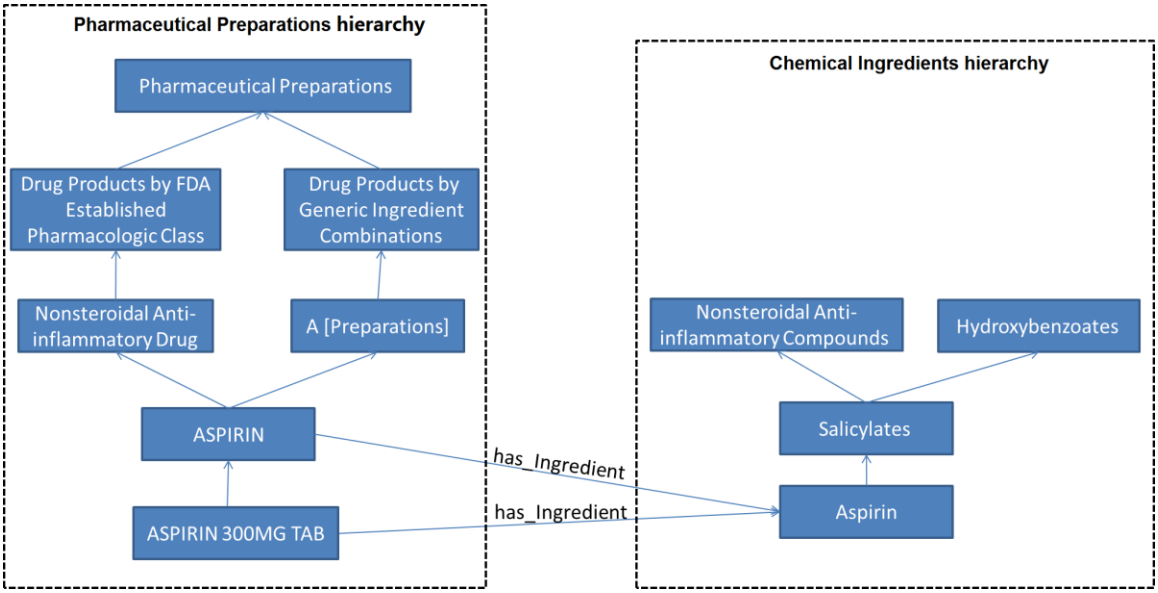

Figure 2.3 An excerpt from NDF-RT's *Pharmaceutical Preparations* and *Chemical Ingredients* hierarchies. Concepts are shown as blue boxes and hierarchical relationships are shown as upward directed blue arrows. The *has_Ingredient* roles linking the concepts in the two hierarchies are shown as labeled blue arrows.

that describes the chemical structure of an FDA-established pharmacologic class is introduced at a drug class level, while the roles *has_Ingredient*, *CI_ChemClass*, and *has_active_metabolites* are introduced at a drug preparation level.

Extensive research has been conducted on NDF-RT, e.g., on its content coverage, the adequacy of representation, drug normalization and classification, etc. Rosenbloom et al. [35] investigated the adequacy of representation in the *Physiologic Effect* hierarchy. Carter et al. [36] studied drug class names from three sources to understand how drugs were classified. They further evaluated NDF-RT's semantic coverage. Zhu et al. [37] normalized drug data in PharmGKB [38] by mapping extracted drugs and drug classes to NDF-RT. Pathak et al. [39] investigated drug-disease relationships in NDF-RT and PharmGKB to make both more robust and integratable. Pathak et al. [40] also evaluated the applicability of RxNorm [41] and NDF-RT to classification of medication data extracted from electronic health records.

2.1.3 National Cancer Institute Thesaurus (NCIt)

The National Cancer Institute Thesaurus (NCIt) [3] is a cancer-focused reference terminology developed and published by the National Cancer Institute (NCI) with the initial goal to facilitate interoperability and data sharing among various information systems at the NCI. It is released at the beginning of each month for free public access in OWL and flat file formats and has been used by an increasing number of information systems outside the NCI, both nationally and internationally [42].

The NCIt covers vocabulary in different domains important for cancer research, including clinical care, basic research, public information dissemination and administrative activities. The content of the NCIt is modeled based on description logic

[43, 44]. A "concept" is the basic unit in the NCIt, just as in many other ontologies/terminologies. The NCIt exists in two versions, the asserted and the inferred version. The asserted version contains assertions explicitly defined by the NCIt team, while the inferred version is obtained by running a reasoner on the asserted version. The studies reported in this dissertation were conducted on the inferred version of the NCIt. The 15.02d release of the NCIt had 108,376 active concepts organized into 19 disjoint IS-A hierarchies, e.g., *Disease Disorder or Finding*, *Gene*, *Biological Process*, *Molecular Abnormality*, and *Abnormal Cell*. Concepts in each hierarchy are connected by IS-A relationships to their parents, forming a directed acyclic graph (DAG), i.e., a concept may have multiple parents.

Roles are binary semantic relationships between pairs of concepts. Each role has a domain and a range, e.g., for the role *Disease Has Abnormal Cell*, relating a disease to the type of neoplastic cell present in the disease, the domain is the *Disease Disorder or Finding* hierarchy and the range is the *Abnormal Cell* hierarchy. Roles are inherited along the IS-A hierarchy. For example, as shown in Figure 2.4, the concept *Neoplasm* has the role *Disease Has Abnormal Cell* pointing to the concept *Neoplastic Cell*. Since *Neoplasm by Morphology* IS-A *Neoplasm*, it inherits the role *Disease Has Abnormal Cell* with the target *Neoplastic Cell* from its parent *Neoplasm*. In fact, all the concepts under *Neoplasm* in Figure 2.4 inherit the above role from *Neoplasm*, because those concepts are *Neoplasm*'s descendants.

Use cases determine, to a large extent, the modeling priorities of the NCIt. Hence, not every hierarchy is modeled with roles. Concepts in eight hierarchies, such as the *Organism* hierarchy and the *Biochemical Pathway* hierarchy, only serve as the ranges of

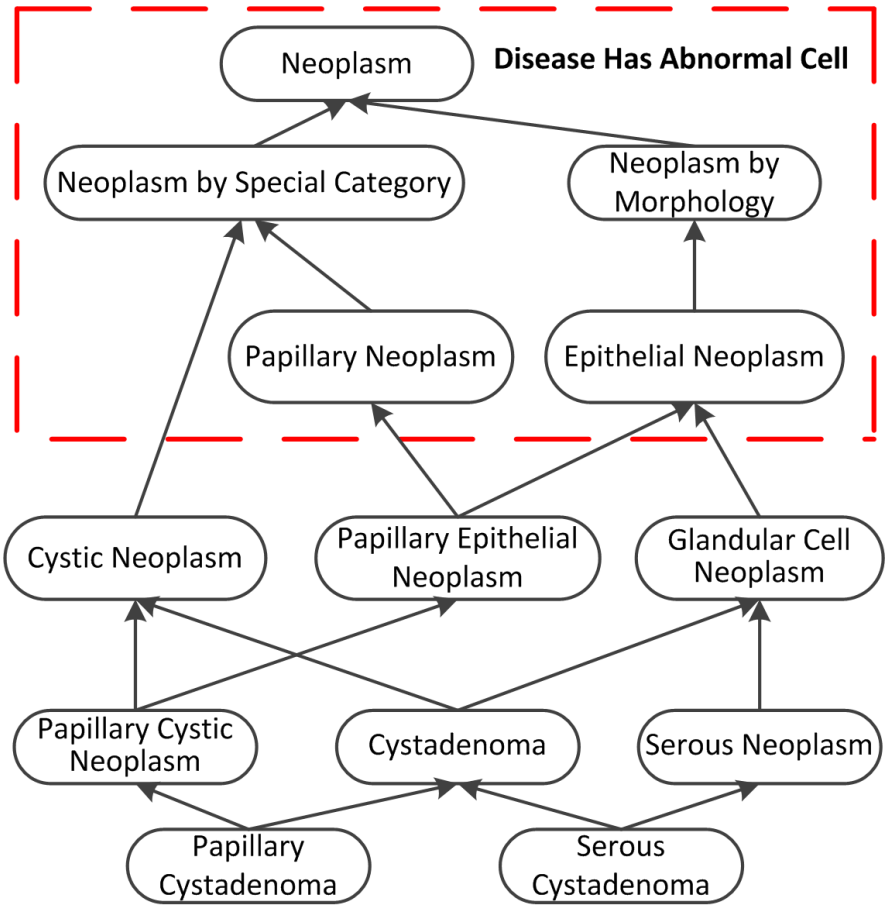

Figure 2.4 An excerpt of 13 neoplasm concepts in the *Disease, Disorder or Finding* hierarchy of NCIt. Concepts represented by boxes with rounded corners are connected by IS-A relationships shown as upward thin arrows.

roles and do not serve as the domains of roles. Each of the other 11 hierarchies has a list of associated defined role types. For example, the *Disease, Disorder or Finding* hierarchy has 29 role types, including *Disease Excludes Abnormal Cell* with the range *Abnormal Cell* hierarchy, *Disease Has Finding* with the range *Disease, Disorder or Finding* hierarchy, *Disease Mapped To Gene* with the range *Gene* hierarchy, and *Disease Has Normal Cell Origin* with the range *Anatomic Structure, System, or Substance* hierarchy.

Due to the mission of NCIt, cancer-related concepts are modeled with a higher priority and more detail than other concepts. In the February 2015 release, the *Disease, Disorder or Finding* hierarchy, which is the largest hierarchy in NCIt in all releases,

contained 25,360 concepts (23.4% = 25360/108376) with 7.79 roles on average. The *Neoplasm* subhierarchy in the *Disease, Disorder or Finding* hierarchy had 8,166 concepts (32.2% = 8166/25360) with an average of 23.02 roles. However, the corresponding average was 0.55 for non-neoplasm concepts, because only 2,858 non-neoplasm concepts (16.6% = 2858/17194) had roles. These 2,858 non-neoplasm concepts had an average number of 3.33 roles. The average number of parents for concepts in the *Neoplasm* subhierarchy was 1.73, while it was 1.10 for the remaining concepts in the *Disease, Disorder or Finding* hierarchy.

The *Gene* hierarchy which had 9,540 concepts in the September 2016 release is another important component of NCIt, because it contains cancer-related knowledge about genes, which are organized according to biological functions [45]. It has a list of 16 defined role types. Some of these role types are *Gene Plays Role In Process*, specifying a biological process in which the gene participates, *Gene Associated With Disease*, indicating a disease associated with molecular abnormalities in a gene, and *Gene In Chromosomal Location*, describing the general location of a gene by chromosomal band position. Table 2.1 shows the number of concepts in the *Gene* hierarchy with each of the five most frequent role types. The other roles are less frequent.

Table 2.1 Distribution of Concepts in the *Gene* Hierarchy of NCIt

Five Most Frequent Role Types	# of Concepts	Percentage (%)
Gene Plays Role In Process	8775	91.98
Gene In Chromosomal Location	3548	37.19
Gene Found In Organism	3258	34.15
Gene Is Element In Pathway	2234	23.42
Gene Associated With Disease	1377	14.43

The *Biological Process* (*BP*) hierarchy is also a core hierarchy in NCIt, because of its relevance for cancer research and treatment, containing 1,145 concepts in the February 2015 release, with seven defined role types (whose full names and abbreviated names are given in Table 2.2). The *Gene* hierarchy and the *Biological Process* hierarchy are closely related to one another. This relation is manifested, for example, by the role *Gene Plays Role In Process* that exists for 92% of the concepts of the *Gene* hierarchy shown in Table 2.1. Among the *BP* hierarchy's 1,145 concepts, 513 (44.8%) have no roles at all. The levels of these concepts without roles (i.e., the maximum distance from the *BP* hierarchy root to each concept) varied from 0 to 9.

Table 2.2 Roles in the *Biological Process* Hierarchy and their Abbreviations

Role	Abbreviated Name
Biological Process Has Associated Location	*Location*
Biological Process Has Initiator Chemical Or Drug	*Initiator Chemical or Drug*
Biological Process Has Initiator Process	*Initiator BP*
Biological Process Has Result Anatomy	*Resulting Anatomy*
Biological Process Has Result Biological Process	*Resulting BP*
Biological Process Has Result Chemical Or Drug	*Resulting Chemical or Drug*
Biological Process Is Part Of Process	*Part of Process*

2.1.4 Chemical Entities of Biological Interest (ChEBI)

The Chemical Entities of Biological Interest (ChEBI) ontology [46] is a structure that houses terminological knowledge concerning chemicals in biological contexts. It serves as an important electronic reference for software systems needing such knowledge. For example, ChEBI has been used in many annotation, text-mining, and chemical-analysis applications. Also, ChEBI's hierarchy has been integrated into the Gene Ontology (GO) [47-50] to support the integration of data across the biology and chemistry domains.

ChEBI's August 2016 release comprised 103,478 concepts, 161,256 IS-A relationships, and 68,395 lateral relationships.

The ChEBI ontology is maintained by the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) and is updated monthly in OBO and OWL format. The version used in this dissertation is the February 2016 inferred version in OWL format. It contained 61,896 concepts, including 47,752 fully annotated chemical entities. Each concept in ChEBI has a unique id, written in the general form "CHEBI: *num*." For example, the concept *Neticonazole hydrochloride* has the id "CHEBI: 31900."

ChEBI divides its classification of molecular entities into three hierarchies. The *chemical entity* hierarchy categorizes molecular entities based on their chemical structure. The *subatomic particle* hierarchy classifies particles that are smaller than atoms. The *role* hierarchy, with its three subhierarchies, defines the roles of compounds in three different settings: (1) their intended use by humans (e.g., fuel, anti-inflammatory agent), (2) their biological context (e.g., growth regulator, inhibitor), and (3) their chemical role (e.g., acid, base).

ChEBI employs three primary relationships in the modeling of its concepts. The hierarchical IS-A relationship denotes the standard subsumption relationship between concepts in the hierarchies. The relationship *has part* indicates the whole/part association between compounds. The relationship *has role* serves to link concepts in the chemical entity hierarchy with those in the role hierarchy. Seven other chemistry-specific, non-hierarchical ("lateral") relationships are also used in modeling concepts. These are *is conjugate base of, is conjugate acid of, is tautomer of, is enantiomer of, has functional*

parent, *has parent hydride*, and *is substituent group from*. Certain pairs of these relationships form converses. For example, the two relationships *is conjugate base of* and *conjugate acid of* are converses of each other, as are *is tautomer of* and *is enantiomer of* [51].

ChEBI is highly user-driven. Users can make requests (e.g., to add a new concept) to the ChEBI curatorial team using the ChEBI submission tool [52]. In the case of a new concept, users must provide minimal unique information, including classifications. Issues and bugs of ChEBI's concepts can be reported using ChEBI's GitHub issue tracking system [53]. As of August 2016, there were 2,951 closed and 243 open issues in the ChEBI GitHub. After ChEBI's curators have validated requests, changes are made available in subsequent releases. For example, a user reported on July 31, 2016 that the *has role* relationship between *protein polypeptide chain* (CHEBI: 16541) and *mouse metabolite* (CHEBI: 75771) is questionable. Two days later, a ChEBI curator responded that the problem was fixed.

2.2 Abstraction Networks for Biomedical Ontologies

As demonstrated by Figure 1.1 in Chapter 1 and the introduction to biomedical ontologies in the previous section, Big Knowledge in large biomedical ontologies is beyond humans' comprehension ability. In order to facilitate the comprehension of the complex content in biomedical ontologies, in a long range research program, the SABOC team [54] has developed an Abstraction Network-based framework to support the summarization and visualization of biomedical ontologies. An Abstraction Network (AbN) of an ontology is a compact summary network consisting of "nodes," each representing a set of concepts

that are *similar* in their structure and semantics. Nodes are connected by hierarchical *child-of* links that are derived from the IS-A relationships in the ontology.

The definition of "similar" depends on an ontology's structural characteristics and is not the same for all ontologies, hence there are various types of Abstraction Networks. For example, the SABOC team has developed the *area taxonomies* and *partial-area taxonomies* [5, 6, 55] for the National Cancer Institute thesaurus (NCIt) [3], SNOMED CT [7], and the Gene Ontology [56]. Furthermore, the *disjoint partial-area taxonomies* [57] and the *tribal abstraction networks* [58] have been designed for SNOMED CT. Besides, they have introduced the *domain-defined partial-area taxonomy* [59, 60] for the Ontology of Clinical Research (OCRe) [61] and the Cancer Chemoprevention Ontology (CanCo) [62], the *restriction-defined partial-area taxonomy* [63] for the Sleep Domain Ontology (SDO) [64], and the *domain-defined and restriction-defined partial-area taxonomies* [65] for the Drug Discovery Investigations Ontology [66]. An extensive review of Abstraction Networks has been presented by Halper et al. [4]. The Ontology Abstraction Framework (OAF) created by Ochs et al. [67] is an open source software system and tool for deriving Abstraction Networks, which is available at http://saboc.njit.edu/. The following sections will describe the Abstraction Networks associated with this dissertation using example neoplasm concepts from NCIt.

2.2.1 Area Taxonomy and Partial-area Taxonomy

The first discussed Abstraction Network is the *area taxonomy*. It is a network composed of *area* nodes and links denoted *child-of*. Another basic Abstraction Network is called *partial-area taxonomy* [6], which is derived from an *area taxonomy*.

Figure 2.5 shows the derivation of the *area taxonomy* and *partial-area taxonomy*

for an excerpt of 13 neoplasm concepts. Concepts with the exact same set of roles are enclosed in dashed, colored boxes in Figure 2.5(a). An *area* represents such a group of concepts with the exact same set of roles and is named by the set of roles. For example, in Figure 2.5(a) the five concepts in the dashed gray box *Neoplasm*, *Neoplasm by Special Category*, *Neoplasm by Morphology*, *Papillary Neoplasm*, and *Epithelial Neoplasm* have only one role *Disease Has Abnormal Cell*. Therefore, there is a corresponding area node named {*Disease Has Abnormal Cell*} summarizing these five concepts in Figure 2.5(b). An *area taxonomy* is an Abstraction Network composed of area nodes connected by *child-of* links, which are derived from the underlying IS-A relationships in the terminology.

A *root* of an area is a concept such that its parent concept(s) are not in this same area. An area may have multiple root concepts. For example, the dashed blue box has the two roots *Papillary Epithelial Neoplasm* and *Glandular Cell Neoplasm.* An area **A** is *child-of* another area **B** if a root in **A** has a parent in **B**. Figure 2.5(b) is the area taxonomy for the 13 concepts in Figure 2.5(a). Area nodes in Figure 2.5(b) are color coded by the number of roles, i.e., areas with the same number of roles have the same color. *Child-of* links are displayed as bold upward arrows. For example, the single red area node at the bottom of Figure 2.5(b) is *child-of* both the green area node and the blue area node.

The area taxonomy summarizes groups of concepts with similar structure. A root concept and all its descendant concepts in an area share a similar semantics, as they are all specializations of the same root concept. Since an area may have multiple roots representing various semantics, an area is further divided into partial-area(s) to get groups of concepts sharing similar structure and similar semantics. A *partial-area* is composed

Figure 2.5 (a) An excerpt of 13 neoplasm concepts in the *Disease, Disorder or Finding* hierarchy of NCIt. Concepts represented by boxes with rounded corners are connected by IS-A relationships shown as upward thin arrows. **(b)** The area taxonomy for the excerpt in (a). **(c)** The partial-area taxonomy for the excerpt in (a).

of a root concept in an area and all its descendant concepts (i.e., children, grand-children, etc.) in the same area. Partial-area nodes represent partial-areas of the terminology in the derived partial-area taxonomy. Partial-area nodes are connected by *child-of* links to form the partial-area taxonomy. Figure 2.5(c) is the partial-area taxonomy for Figure 2.5(a), in which partial-area nodes are represented as white boxes within area nodes. A partial-area node is labeled by its root concept, with the number of concepts that the node summarizes in () parentheses. *Child-of* links connecting partial-area nodes are also represented as bold

arrows. For example, *Papillary Cystic Neoplasm (2)* is *child-of* both *Cystic Neoplasm (1)* and *Papillary Epithelial Neoplasm (1)*.

2.2.2 Disjoint Partial-area Taxonomy

Note that the red (bottom) area in Figure 2.5(b) has five concepts, while the sum of the numbers of concepts in the three partial-areas in Figure 2.5(c) is 7 (= 2+3+2). That is the case, because both *Papillary Cystadenoma* and *Serous Cystadenoma* have two parents which are roots of the red area. Therefore, both concepts are simultaneously summarized by two partial-areas. Concepts that are summarized by more than one partial-area are called "overlapping concepts." Note that overlapping concepts cause some ambiguity in the summarization due to their multiple summarizations.

In order to eliminate the phenomenon of summarization ambiguity of overlaps among partial-areas, the *disjoint partial-area taxonomy* [57] was developed. The basic idea is to extract overlapping concepts from their original partial-areas and place them into their own partial-areas. As a result all resulting partial-areas become disjoint. Figure 2.6 illustrates the derivation of the disjoint partial-area taxonomy for an excerpt of 15 neoplasm concepts in the area {*Disease Excludes Abnormal Cell, Disease Excludes Finding, Disease Has Abnormal Cell, Disease Has Finding, Disease Has Normal Cell Origin, Disease Has Normal Tissue Origin*}. Figure 2.6(a) shows that the 15 concepts of the excerpt are organized into four partial-areas *Serous Neoplasm (5)* enclosed by dashed orange lines, *Cystadenoma (12)* enclosed by dashed green lines, *Papillary Cystic Neoplasm (8)* enclosed by dashed red lines and *Mucinous Neoplasm (4)* enclosed by dashed blue lines. The upward thin arrows represent IS-A relationships between concepts. These four partial-areas have 10 overlapping concepts. For example, the concept

Papillary Serous Cystadenoma appears in three partial-areas (red, yellow, and green).

In Figure 2.6(b), the four concepts in the first row (=Level 1) in different solid colors are the roots of the area ("area roots"). The concept *Borderline Cystadenoma* (without color at Level 2) and the four root concepts are non-overlapping concepts. The concept *Serous Cystadenoma* (Level 2) is a child of two root concepts. Previously, in the partial-area taxonomy, this concept would appear twice, namely once in the partial-area defined by each of the two roots. In the *disjoint partial-area taxonomy*, however, this concept is promoted to becoming a root of its own partial area. Such a concept is called an *overlapping root.* The same process is repeated for other concepts that have two or

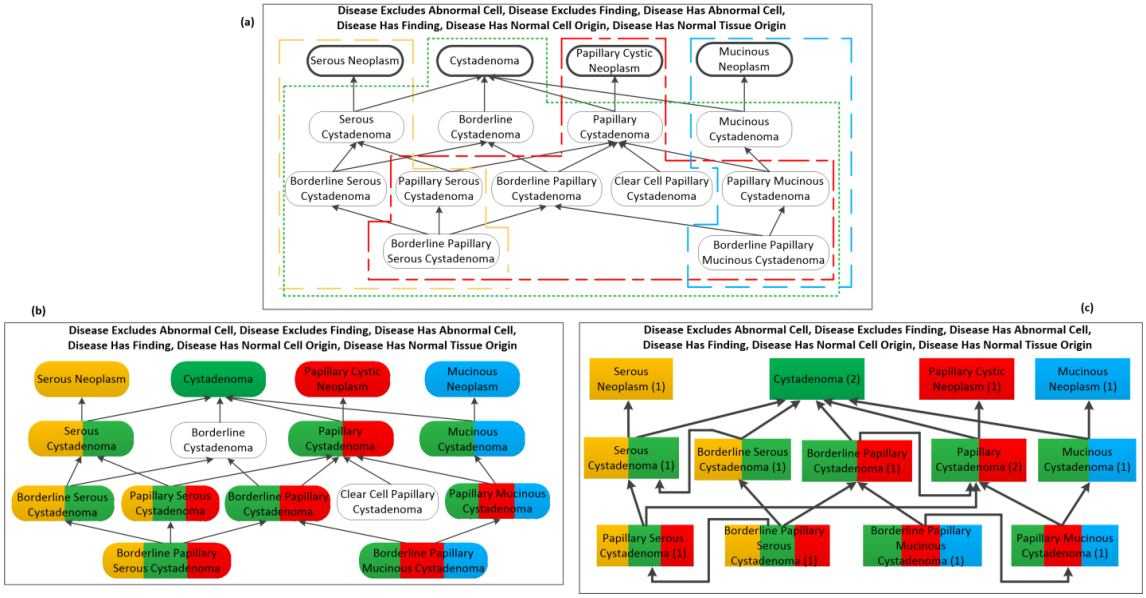

Figure 2.6 (a) An excerpt of 15 neoplasm concepts from the area {*Disease Excludes Abnormal Cell, Disease Excludes Finding, Disease Has Abnormal Cell, Disease Has Finding, Disease Has Normal Cell Origin, Disease Has Normal Tissue Origin*} distributed in four partial-areas enclosed by four different colored dashed boxes. **(b)** The roots of disjoint partial-areas are colored. Area roots have a single color and overlapping roots have multiple colors according to the colors of their multiple ancestor area roots. **(c)** The disjoint partial-area taxonomy for the excerpt in (a). Disjoint partial-areas are color coded according to the colors of their roots. Disjoint partial-areas with the same number of colors are placed at the same level, e.g., the five disjoint partial-areas with two colors are at the second level. There may be child-of relationships between disjoint partial-areas at the same level.

more parents that are area roots. As a side remark, the concept *Clear Cell Papillary Cystadenoma* (Level 3) is without color, because it is not a root; however, it is an overlapping concept, as it inherits from *Cystadenoma* and from *Papillary Cystic Neoplasm.*

There is a complication that requires performing the above operation recursively at every level of the taxonomy. *Papillary Serous Cystadenoma* (at Level 3) is a child of two concepts that have now become overlapping roots. Thus, it would have to appear in the partial-areas defined by both these two overlapping roots in the partial-area taxonomy (Figure 2.6(a)). To avoid this, the same method is applied again one level down, and *Papillary Serous Cystadenoma* is itself promoted to overlapping root.

Overlapping roots are multicolored according to the colors of their multiple ancestor area roots. For example, the overlapping root *Serous Cystadenoma* has two colors, orange and green, because its two parents *Serous Neoplasm* (orange) and *Cystadenoma* (green) are area roots. Another example is the overlapping root *Papillary Serous Cystadenoma* with three colors, orange, green and red, because it has the combined semantics inherited (over two levels) from the three area roots *Serous Neoplasm*, *Cystadenoma* and *Papillary Cystic Neoplasm*.

Figure 2.6(b) still displays "concept space," like the original terminology. To arrive at the final *disjoint partial-area taxonomy*, three more steps have to be taken. All non-root concepts have to be deleted, because they are represented by their disjoint partial-area root nodes. Any IS-A link pointing to a deleted concept is redirected to the disjoint partial-area root node that will represent the deleted concept. Thus, the two uncolored concepts of Figure 2.6(b) are eliminated. Secondly, it is more intuitive to

organize overlapping roots (and thus disjoint partial-area nodes) by the number of colors in one node. Thus, all nodes with two colors are shown next to each other at Level 2 in Figure 2.6(c). Finally, for every root concept, whether overlapping root or area root, the number of concepts that this disjoint partial-area node represents is shown in () parentheses. (By this step, a root concept is effectively converted into a disjoint partial-area node.) Thus, *Cystadenoma* appears in Figure 2.6(c) as *Cystadenoma (2)*, because this disjoint partial-area node represents the deleted concept *Borderline Cystadenoma*. The arrows represent *child-of* links between disjoint partial-area nodes, with a similar interpretation as the previously described *child-of* links in a partial-area taxonomy and an area taxonomy.

2.3 Quality Assurance of Biomedical Ontologies

Building an ontology is a burdensome task, requiring a thorough understanding of the application domain as well as authoring skills following ontological rules. Many important biomedical ontologies (e.g., NCIt) have a large, complex network structure that poses significant maintenance challenges. It is not reasonable to expect that ontologies are completely free of modeling errors and inconsistencies. Errors and inconsistencies in biomedical ontologies impede their applications. Hence, Quality Assurance (QA) is a fundamental part of the life cycle of an ontology [5]. However, QA is a challenging and resource-intensive task. Without the help of automatic or semi-automatic techniques and tools, it is impossible to maintain ontologies with a high quality.

Automated support for terminology and ontology QA has been a focus of much research, especially for large complex structures playing important roles in the

biomedicine field like the UMLS Metathesaurus [68-72], SNOMED CT [73-77], and Gene Ontology (GO) [78-81]. For NCIt, both internal and external QA reviews of NCIt have been conducted. Different internal QA techniques, including various automated and manual methods, have been employed during the whole life-cycle of NCIt [42]. For example, during the editing phase, concept definitions are reviewed by editors following the NCIt Editor Guide. Externally, a qualitative analysis of NCIt determining its adherence to relevant ISO terminology standards and ontological principles was performed [82]. In that study, it was concluded that the particular version of the NCIt suffered from the same broad range of problems (e.g., missing or inappropriately assigned verbal and formal definitions) as other biomedical ontologies.

Structural characterizations based on Abstraction Networks have been applied to the *Biological Process* hierarchy of NCIt to identify sets of concepts with different kinds of errors in the hierarchy [5]. The main observation was that small partial-areas, which are units comprising few concepts that are all similar to each other in their structure and semantics, tend to exhibit higher error rates than large partial-areas. A comparative QA methodology focusing on the biological processes of different genes was carried out on NCIt's *Gene* hierarchy with the use of the National Center for Biotechnology Information's (NCBI's) Entrez Gene database [83]. A multiphase QA methodology based on Abstraction Networks was also used on the *Gene* hierarchy to detect different kinds of role errors [84]. Another QA methodology based on semantic web technologies and using relationships defined in the UMLS Semantic Network identified inconsistencies in the hierarchical relationships and roles in the NCIt [85].

More and more biomedical ontologies, including large ones such as SNOMED CT [7], National Drug File – Reference Terminology (NDF-RT) [27] and NCIt [42] are modeled using description logic (DL) to ensure logical consistency of these ontologies. DL reasoners can automatically identify logical inconsistencies but cannot detect semantic errors (e.g., missing or incorrect relationships) that do not cause logical conflicts. Wei et al. [86] demonstrated that other methods (e.g., Abstraction Network-based methodologies) are needed to complement DL classifiers to identify semantic errors.

Below is a discussion of two general overviews of quality assurance of biomedical ontologies. Rogers [87] reviewed literature on the quality assurance of logic-based medical ontologies from scholar.google.com in 2006 and proposed a framework to evaluate ontologies according to four aspects, namely philosophical rigor, ontological commitment, content correctness and fitness for purpose. Zhu et al. [88] performed an extensive review of auditing methods applied to various biomedical terminologies in 130 studies, which appeared in the first journal special issue on auditing of terminologies [89]. They extended the review target from ontologies to all forms of controlled biomedical terminologies and presented a framework to characterize various auditing methods, applied to different terminologies, with appropriate examples.

The SABOC research team has shown that Abstraction Networks are useful in support of ontology QA [4]. In particular, the alternative view of an ontology offered by an Abstraction Network supports the identification of sets of concepts with high likelihood of errors.

Two main characterizations of concepts with high likelihood of errors identified by the SABOC team are “uncommonly modeled” concepts and “complex” concepts. For different ontologies, the definitions of “uncommon” and “complex” may be different. For example, Halper et al. [90] identified partial-areas of up to seven concepts in the *Specimen* hierarchy of SNOMED CT as sets of “uncommon” concepts. Min et al. [5] also obtained a similar result for the *Biological Process* hierarchy of NCIt, but only for partial-areas of up to three concepts. Several studies have demonstrated that overlapping concepts in partial-area taxonomies are “complex” concepts [4, 91-93] such as for the *Specimen* hierarchy and the *Clinical finding* hierarchy of SNOMED CT, and for the Uber Anatomy Ontology (UBERON) [94].

CHAPTER 3

ADVANCED ABSTRACTION NETWORKS

The Big Data to Knowledge (BD2K) initiative is expected to produce many knowledge items that can be expressed as assertions or as rules. However, orientation into large knowledge bases is a challenge by itself, the "Big Knowledge" challenge. Without some high-level mental representation of the kinds of content in a large knowledge base, effective use of the knowledge may be limited [95]. When an ontology surpasses many thousands of assertions, even its curators are confronted with the problem of seeing the "big picture" of its content, which would impede the ontology's maintenance and applications. Hence, in order to facilitate the users' "big picture" comprehension, it is important to provide automated tools for summarization of the content in a large ontology. This is one manifestation of the "Big Knowledge" challenge [4].

In order to address the "Big Knowledge" challenge, this chapter introduces two advanced Abstraction Networks, the *weighted aggregate partial-area taxonomy* [9, 10] that provides a more compact and flexible summary of biomedical ontologies compared with the original partial-area taxonomy, and the *Ingredient Abstraction Network* (IAbN) [11, 12], summarizing and visualizing NDF-RT's *Chemical Ingredient* hierarchy. The latter had to be derived due to the terminology's unique modeling structure, for which the previously developed kinds of Abstraction Networks cannot be derived.

3.1 Weighted Aggregate Partial-area Taxonomy

In previous studies by the SABOC team, various types of Abstraction Networks have been developed to summarize, visualize, and support the quality assurance of biomedical ontologies. However, biomedical ontologies are complex knowledge systems in terms of their large numbers of concepts and tens or even hundreds of thousands of relationships. Although the previously derived Abstraction Networks, e.g., the partial-area taxonomy, are compact compared to the ontologies themselves, they are usually still too overwhelming to comprehend, since there are many nodes, some of which are summarizing only a few concepts. In order to address the "Big Knowledge" challenge, this section introduces a more compact Abstraction Network for ontologies, named *weighted aggregate partial-area taxonomy*.

A *weighted aggregate partial-area taxonomy* [9, 10] ("*aggregate taxonomy*" for short) is a variation of a partial-area taxonomy with an adjustable parameter *b* that is used to control the granularity of summarization. First, a complete partial-area taxonomy is created for the ontology (the *original partial-area taxonomy)*. Next, based on the given parameter *b*, "small" partial-areas are aggregated into their closest "large" ancestor partial-area(s) [96]. In this way, the small partial-areas are not removed but are instead hidden to obtain a more compact summary. The parameter *b* is used to distinguish between small and large partial-areas. The *size* of a partial-area is defined as the number of concepts it summarizes. The *aggregated weight* of a partial-area in the original partial-area taxonomy is defined as the sum of its size and the sizes of all its descendant partial-areas.

The parameter *b* specifies the minimum aggregated weight of a partial-area in the

original partial-area taxonomy that will appear explicitly in the aggregate taxonomy. More precisely, all partial-areas with an aggregated weight greater than or equal to *b* will be included in the resulting aggregate taxonomy. Using a *topological sort*, an aggregate taxonomy is generated by aggregating any partial-areas with an aggregated weight less than *b* into their closest parent/ancestor partial-area(s) (which have an aggregated weight $\geq b$). The root partial-area will be included regardless of size, to ensure that there is always a root in the aggregate taxonomy. The nodes in the aggregate taxonomy, called *aggregate partial-areas*, summarize their descendant partial-areas from the original partial-area taxonomy. A partial-area may appear unchanged after performing the aggregation process.

Figure 3.1 illustrates the derivation of an aggregate taxonomy with b=20 for a subhierarchy of eight partial-areas rooted at *Colorectal Carcinoma (19)*. The partial-areas in the blue areas in Figure 3.1(a) are child partial-areas of *Colorectal Carcinoma (19)*. The aggregated weight of *Colorectal Carcinoma (19)* is thus 56 (=19 + 24 + 4 + 3 + 1 + 3 + 1 + 1) and the aggregated weights of its seven child partial-areas are their sizes, as they have no descendant partial-areas. With b=20, the six child partial-areas with an aggregated weight less than 20 are aggregated into the partial-area node at the first level. The only child partial-area is *Colorectal Adenocarcinoma (24)* with an aggregated weight greater than 20. It stays unchanged and is represented by a white rectangle with "sharp corners" in Figure 3.1(b). A rectangle with rounded corners summarizes small descendant partial-areas.

Figure 3.3 shows the weighted aggregate taxonomy with the parameter *b* as 200 for the NCIt *Neoplasm* subhierarchy with 8,445 concepts, consisting of only 25 aggregate

partial-areas. Compared to the *Neoplasm* subhierarchy itself in Figure 1.1 on Page 2 and its partial-area taxonomy with 4,177 partial-areas shown in Figure 3.2, Figure 3.3 captures the "big picture" of the subhierarchy. For example, there are 1199 concepts related to *Reproductive System Neoplasm* and 909 *Connective and Soft Tissue Neoplasm* concepts.

Note that, given the original partial-area taxonomy and the aggregated weight of each of its partial-areas, the parameter *b* can be automatically adjusted such that the aggregate taxonomy will consist of no more than some fixed number of aggregate partial-areas. In this way, the resulting aggregate taxonomy becomes smaller and more comprehensible.

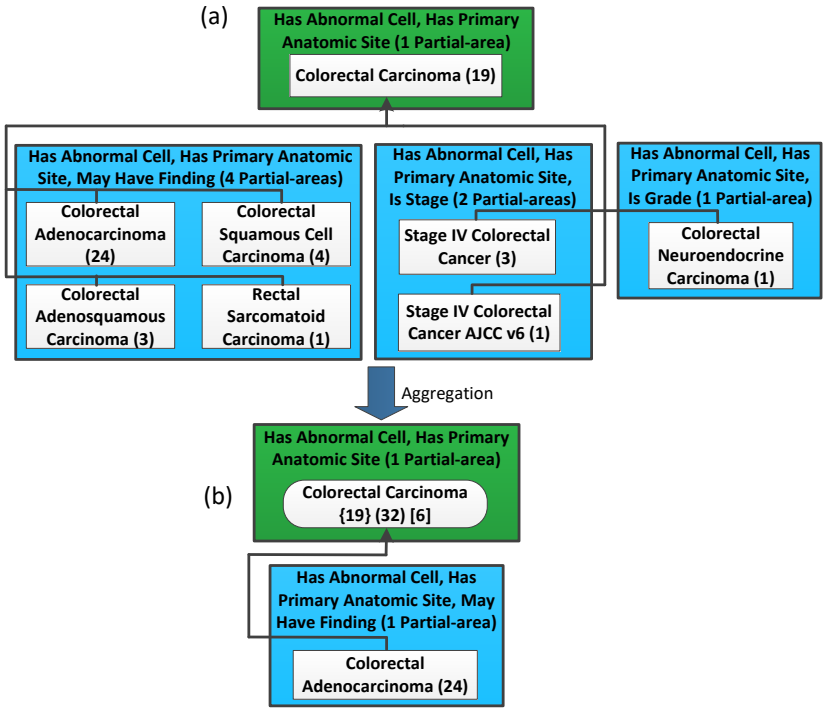

Figure 3.1 (a) An excerpt of eight partial-areas in the NCIt *Neoplasm* partial-area taxonomy. **(b)** Weighted aggregate partial-area taxonomy for (a) with *b*=20. A "rounded" white rectangle represents an aggregate partial-area with its number of concepts in the original partial-area taxonomy in {}, its number of concepts in the aggregated taxonomy in () including all concepts from aggregated partial-areas, and the number of its aggregated partial-areas in []. A white rectangle with "corners" represents an aggregate partial-area that does not summarize any descendant partial-areas.

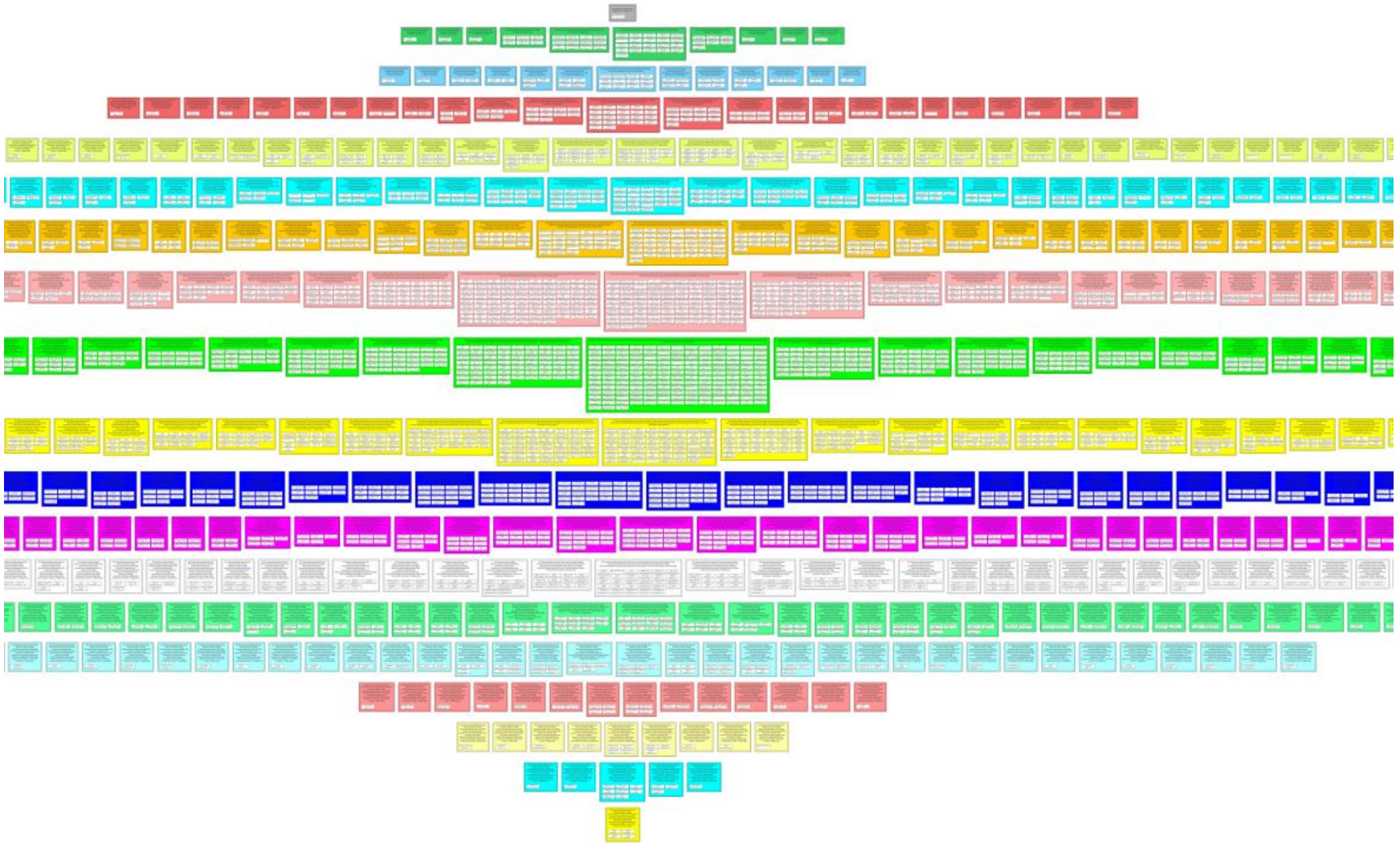

Figure 3.2 The partial-area taxonomy for the NCIt *Neoplasm* subhierarchy with 8,445 concepts shown in Figure 1.1.

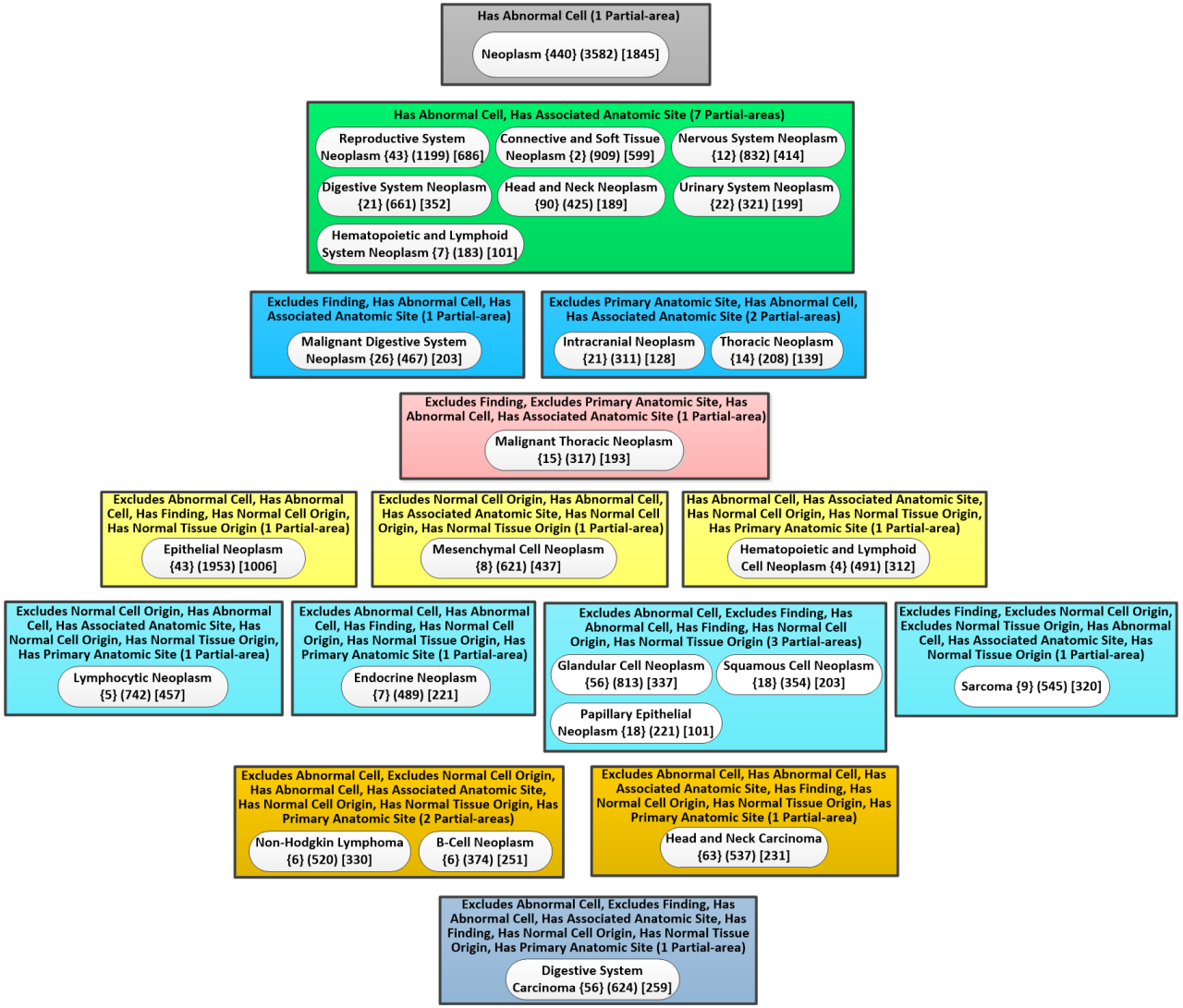

Figure 3.3 The weighted aggregate partial-area taxonomy with 25 aggregate partial-areas for the *Neoplasm* subhierarchy (b=200).

3.2 Ingredient Abstraction Network (IAbN)

The effects of drugs depend mostly on their chemical ingredients and each drug in NDF-RT is linked to its chemical ingredients via *has_Ingredient* roles (see Figure 2.2). Improving the modeling of NDF-RT's chemical ingredient concepts would improve the modeling of NDF-RT's drug concepts. The *Chemical Ingredients* (*CI*) hierarchy of NDF-RT is relevant to Drug-Drug Interactions (DDIs), since in many cases drugs that are chemically similar tend to have similar interactions [97]. Hence, it is necessary to make sure the modeling of the *Chemical Ingredients* hierarchy is of high quality.

In a long range research program, the SABOC team [54] has developed a framework that combines summarization, visualization and quality assurance (SVQA) into a sequence of well-ordered steps. The overall aim of the SVQA paradigm is to identify sets of concepts in a terminology that are expected to have a higher error rate than other concepts. The identification of theses sets is based on *summaries* derived from the terminology's structure and semantics. Limited QA resources can be applied to the concepts in such a set, improving the rate of error detection and correction.

However, it is impossible to derive *partial-area taxonomies* for large portions of NDF-RT; the structures of several of its concept hierarchies do not contain enough information to perform such a derivation. Seven of the NDF-RT hierarchies have no roles emanating from their concepts (i.e., their concepts only have hierarchical relationships). All of the roles in NDF-RT emanate from concepts in the *Pharmaceutical Preparations* (*PP*) hierarchy and point to concepts in the other hierarchies (see Figure 2.2). Hence, the only hierarchy of the NDF-RT that lends itself to deriving the partial-area taxonomy (see Section 2.2.1) is the *PP* hierarchy.

Thus, to apply the SVQA process to NDF-RT's *Chemical Ingredients* (*CI*) hierarchy, a new summarization process is needed. The *Ingredient Abstraction Network* (IAbN) is the new Abstraction Network that summarizes NDF-RT's chemical ingredients and their associated drug concepts.

An *Ingredient Abstraction Network* (IAbN) is an Abstraction Network where the nodes summarize (1) the ingredients in the *Chemical Ingredients* hierarchy **and** (2) those drug concepts in the *Pharmaceutical Preparations* hierarchy that have no dosage information but that do have at least one *has_Ingredient* role to a drug ingredient in the *Chemical Ingredients* hierarchy.

Drug ingredients are chemical ingredients that are used in prescription drugs. Five categories of concepts in the *Chemical Ingredients* (*CI*) hierarchy were defined. The right side of Figure 3.4(a) illustrates the following categories of drug concepts for an excerpt of 14 *CI* concepts.

Definition 1: A *drug ingredient concept* is a concept in the *Chemical Ingredients (CI)* hierarchy that is the target of *has_Ingredient* role(s) from concepts in the *Pharmaceutical Preparation* hierarchy.

Definition 2: A *classification ingredient concept* is a concept in *CI* that "organizes" other drug ingredient concepts below it. In other words, it has drug ingredient concepts as children. It may or may not be itself a target of a *has_Ingredient* role.

Definition 3: A *dual ingredient concept* is both a drug ingredient concept and a classification ingredient concept in *CI*. Such a concept is a target of a *has_Ingredient* role and has children that are drug ingredient concepts.

Definition 4: A *strict classification ingredient concept* is a classification ingredient concept that is **not** also a drug ingredient concept. That is, it is not a target of a *has_Ingredient* role.

In other words, a *classification ingredient concept* is either a dual ingredient concept or a strict classification ingredient concept.

Definition 5: An *uncategorized ingredient concept* is a concept in the *CI* hierarchy that is neither a drug ingredient concept nor a classification ingredient concept. Such concepts are not summarized in the IAbN.

The design of an Abstraction Network for the *CI* hierarchy poses a challenge for several reasons: (1) A lack of roles emanating from *CI* concepts prevents the derivation of a partial-area taxonomy (see Section 2.2.1) that can be derived for many other description logic-based terminologies. (2) The need to distinguish between *drug ingredient concepts* and *classification ingredient concepts* is further complicated by the existence of *dual ingredient concepts*. (3) To obtain a "big picture" of the *Chemical Ingredients* hierarchy there is a need to summarize the drug concepts, which in NDF-RT are parts of the *PP* hierarchy, according to their ingredient concepts in *CI*, as was illustrated by Ochs et al. [11].

The derivation algorithm for an IAbN begins with identifying all of the drug concepts in the *PP* hierarchy that have a *has_Ingredient* role but no *has_DoseForm* role. *PP* concepts with dosage information are ignored, since an ancestor concept, typically a parent (a *PP* generic drug ingredient), introduces the *has_Ingredient* role, which is inherited to such concepts. Hence, there is no need for direct summarization of *PP* concepts with dosage information, since such a summary is offered indirectly through the

summarization of the ancestor *PP* concepts without dosage information. All the *PP* concepts in Figure 3.4(a), except *Pharmaceutical Preparations*, have one *has_Ingredient* role to a concept in the *CI* hierarchy. Different drug concepts in the *PP* hierarchy can have a *has_Ingredient* role to the same *CI* concept, e.g., both *Aspirin* and *Acetylsalicylate Sodium* have the ingredient *Aspirin*. *PP* concepts may also have multiple *has_Ingredient* roles, e.g., *Aspirin/Caffeine* has distinct *has_Ingredient* roles to both *Aspirin* and *Caffeine*.

In the next step, drug ingredient concepts (see Definition 1 above) are identified by collecting the target concepts of all the *has_Ingredient* roles. Classification ingredient concepts (see Definition 2 above) are identified by analyzing the parent concept(s) of each drug ingredient concept. Next, for each drug ingredient concept, the lowest ancestor(s) that are a strict classification ingredient concept(s) (see Definition 4 above) are identified, with the intention of finding groups of drug ingredient concepts. (Common ancestors will be used in the next step to define groups.)

For example, for the *Aspirin CI* concept, the lowest ancestor that is a strict classification ingredient concept is *Salicylates*. *Salicylates* is the lowest common ancestor for *Aspirin*, *Magnesium Salicylate*, and *Diflunisal*. *Warfarin Sodium*'s parent concept *Warfarin* is a classification ingredient concept, but it is also a drug ingredient concept (i.e., it is a dual ingredient concept; see Definition 3 above). Thus, the lowest ancestor of *Warfarin Sodium* that is a strict classification ingredient concept is *Warfarin*'s parent, *4-Hydroxycoumarins*. Many *CI* hierarchy concepts have multiple parents, thus, a given drug ingredient concept may have more than one lowest ancestor that is a strict classification ingredient concept.

In the next step of deriving the Abstraction Network, the drug ingredient concepts

are grouped together according to their common ancestor(s) that are strict classification ingredient concepts. For example, *Aspirin, Magnesium Salicylate, and Diflunisal* share *Salicylates* as a lowest common ancestor. Similarly, *Warfarin, Warfarin Sodium, and Phenprocoumon* share *4-Hydroxycoumarins* as a lowest common ancestor. Figure 3.4(b) models the right side of Figure 3.4(a) and shows the "drug ingredient groups" induced by the lowest common ancestors. Color coding in Figure 3.4(b) helps to keep the groups apart: Every group has its own color.

In the following step, each strict classification ingredient concept is recast as a *root* for its ingredient group. Roots of a group are shown with solid fill in Figure 3.4(b). Thus *Salicylates* becomes the root of the group with *Aspirin, Magnesium Salicylate, and Diflunisal* in it. The *CI* root concept, *Chemical Ingredients*, is also a root. Roots represent groups of *CI* concepts in the IAbN (Figure 3.4(c)). The text line "3 Ingredients" under *Salicylates* in Figure 3.4(c) indicates how much information is summarized by this box. Ingredient groups are not disjoint; drug ingredient concepts with multiple parents may be summarized by multiple ingredient groups. With this step, a summary (Figure 3.4(c)) of the "right side" (the *Chemical Ingredients* hierarchy) of Figure 3.4(a) has been created. In the next step, information from the left (*PP*) side of Figure 3.4(a) will be included into Figure 3.4(c).

For each ingredient group, the *PP* drug concepts that have a *has_Ingredient* role to a drug ingredient concept in the ingredient group are identified. For example, the *Aspirin* and *Acetylsalicylate Sodium* drug concepts in *PP* both have *Aspirin* in *CI* as the target of their *has_Ingredient* roles. The *Aspirin* drug ingredient concept belongs to the *Salicylates* ingredient group, thus, the *Aspirin* and *Acetylsalicylate Sodium* drug concepts

from *PP* are also summarized by the *Salicylates* ingredient group. This is expressed by the text line "4 Drugs" under *Salicylates* in Figure 3.4(c). (The other two drug concepts are *Magnesium salicylate* and *Diflunisal*). Since ingredients may belong to multiple ingredient groups, a given *PP* drug concept may be represented by multiple ingredient groups.

Within the IAbN, ingredient groups are organized into a hierarchy according to *child-of* links derived from the underlying IS-A hierarchy. An ingredient group **A** is a *child-of* another ingredient group **B** if **A**'s root has **B**'s root as an ancestor in the *CI* hierarchy and there are no other roots of the IAbN on any path from **A**'s root to **B**'s root in the *CI* hierarchy. An ingredient group may be a *child-of* multiple ingredient groups. In summary, Figure 3.4(c) shows the IAbN derived from NDF-RT excerpt in Figure 3.4(a).

In the visualization of an IAbN, it is necessary to organize the ingredient groups in a way that helps the summary reflect the "big picture." Thus, ingredient groups *may be* organized into color coded levels according to the length of the longest *child-of* path to the root ingredient group (*Chemical Ingredients*). This will be shown later in Figure 3.5. Figure 3.4(c) does **not** use this color level encoding.

Note that *Ethyl Biscoumacetate* is an uncategorized ingredient concept (see Definition 5 above), as shown in Figure 3.4(a). This occurs when an ingredient is modeled in *CI* but no *PP* drug concept has a *has_Ingredient* role to this ingredient. For the current research, such concepts are not summarized by any ingredient group and are not considered part of the IAbN.

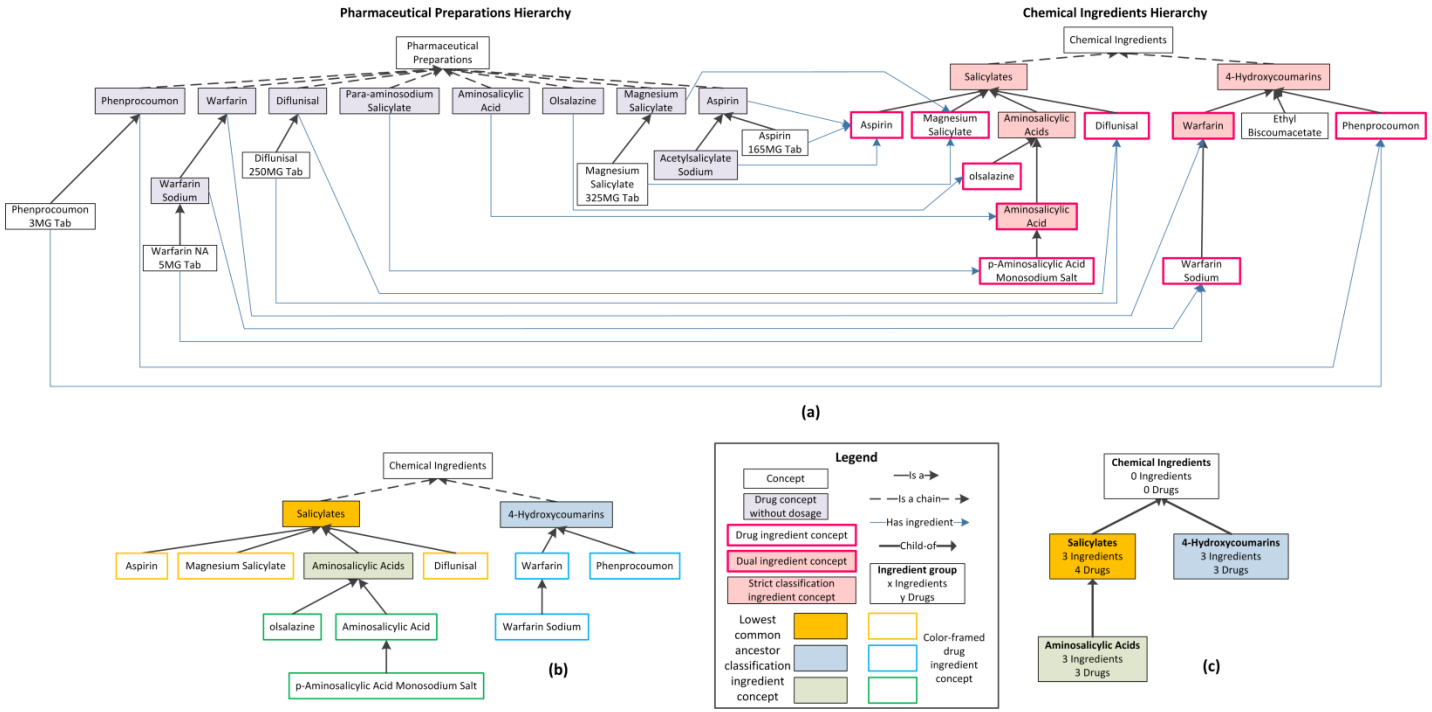

Figure 3.4 (a) An excerpt of concepts from NDF-RT's *Pharmaceutical Preparations* (*PP*) and *Chemical Ingredients* (*CI*) hierarchies. On the left, drug concepts in the *PP* hierarchy with no dosage information have a shaded background. On the right, nine drug ingredient concepts have red borders and five classification ingredient concepts have a pink background. Two concepts, *Aminosalicylic Acid* and *Warfarin*, are both drug ingredient concepts and classification ingredient concepts, i.e., they are dual ingredient concepts. *Ethyl Biscoumacetate* is neither a drug ingredient concept nor a classification ingredient concept, i.e., it is an uncategorized ingredient concept. **(b)** *CI* grouped. Drug ingredient concepts are not shaded and their lowest common ancestor classification ingredient concepts are shaded. Each drug ingredient concept is color-framed according to its lowest common ancestor classification ingredient concept. **(c)** The IAbN for Figure 3.4(a). Ingredient groups are shown as boxes that are labeled with the name of the lowest common ancestor from Figure 3.4(b). In each box are the total number of ingredient concepts summarized by the group, and the total number of drug concepts (without dosage information!) with *has_Ingredient* roles pointing to the summarized concepts in the *CI* hierarchy. *Child-of* links between ingredient groups are shown as upward directed bold arrows.

An IAbN for the June 2015 release of NDF-RT's *Chemical Ingredients* (*CI*) hierarchy, consisting of 10,145 concepts, was derived. This IAbN consists of 860 ingredient groups, which summarize 2,664 drug ingredients and 6,872 *Pharmaceutical Preparation* hierarchy drug concepts. The *abstraction ratio* of the IAbN was defined to be the average number of drug ingredients per ingredient group. The abstraction ratio of the June 2015 IAbN is 3.07 (=2,664/860). There are 813 drug ingredient concepts summarized by more than one ingredient group (with a total of 535 such ingredient groups), and each such drug ingredient is summarized by an average of 1.52 (=813/535) ingredient groups. The average number of *PP* drug concepts summarized by each ingredient group is 7.99 (=6,872/860).

Figure 3.5 shows an excerpt of 128 of the IAbN's ingredient groups, as the IAbN is too large to fit on a single page. By reviewing the ingredient groups of the IAbN, one can see the major types of drug ingredients used in NDF-RT's drugs. For example, the *Polymers* group (Level 2: green) summarizes 26 ingredients and 81 drugs, *Piperidines* (Level 3: blue) summarizes 47 ingredients and 76 drugs, *Tetracyclines* summarizes 17 ingredients and 35 drugs, *Ethanolamines* summarizes 45 ingredients and 231 drugs, and *Penicillins* summarizes 34 ingredients and 64 drugs.

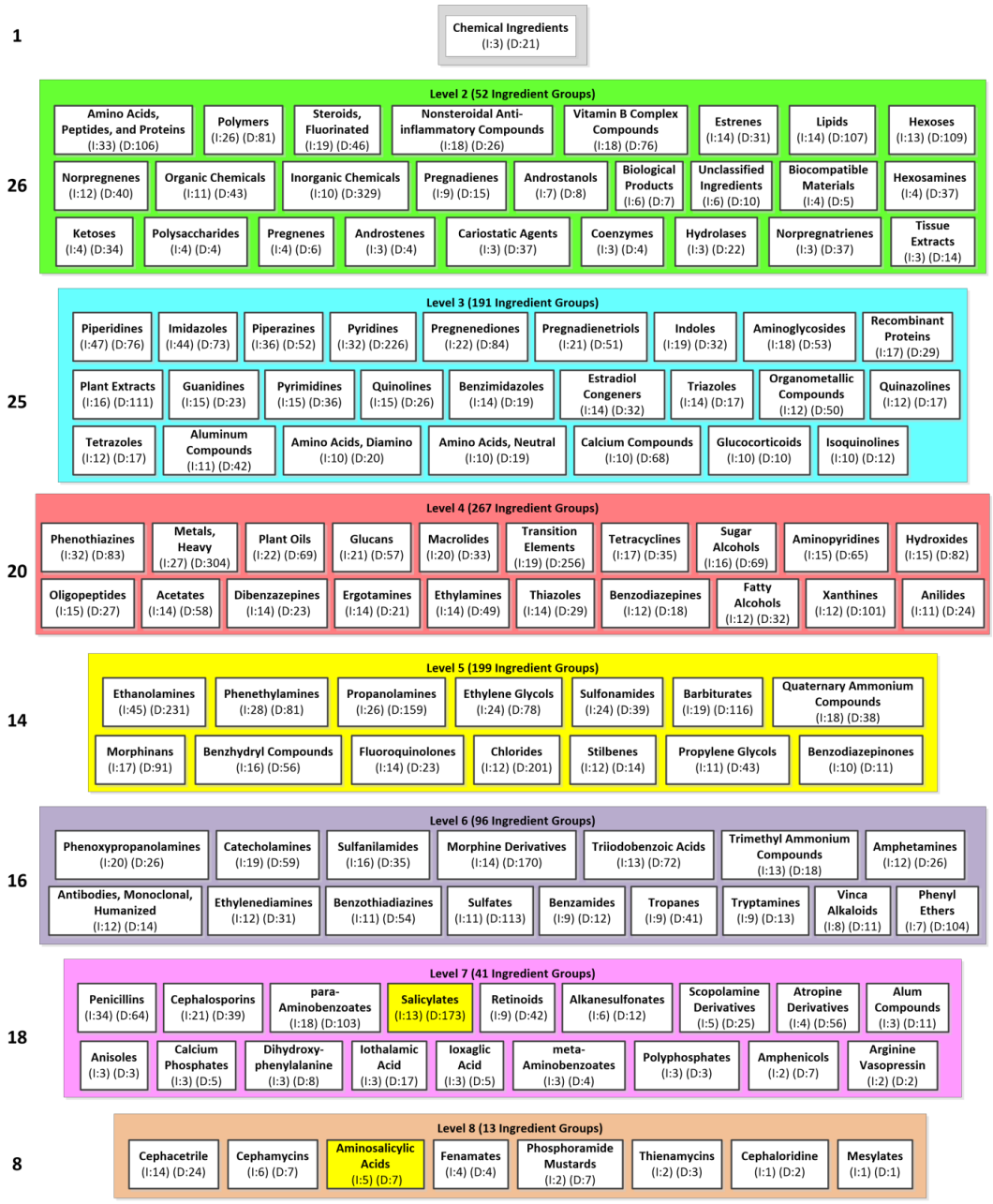

Figure 3.5 An excerpt of 128 (15%) ingredient groups from the IAbN for the June 2015 version of the *CI* hierarchy. The smaller ingredient groups have been hidden as follows. Each level shows as many groups as possible, in decreasing order by the number of ingredients in each group, while keeping the group names readable. *Child-of* links are hidden for readability. The numbers of ingredients and drugs summarized by each ingredient group are shown in parentheses and prepended with I: and D:, respectively. *Salicylates* and *Aminosalicylic Acids*, from Figure 3.4(c), are highlighted in yellow.

CHAPTER 4

BIG KNOWLEDGE COMPREHENSION

This chapter presents three applications of the two advanced Abstraction Networks introduced in Chapter 3 to demonstrate their effectiveness for the Big Knowledge comprehension. The first application is a summarization approach for the automatic identification and display of major topics covered by an ontology's content. This approach is based on the weighted aggregate partial-area taxonomy. SNOMED CT's *Specimen* hierarchy was the test-bed for evaluating the effectiveness of this approach. Another application of the weighted aggregate partial-area taxonomy is a multi-layer, multi-granularity Big Knowledge visualization scheme. The visualization scheme is demonstrated on the National Cancer Institute thesaurus's *Neoplasm* subhierarchy to support its comprehension. The innovative application of the Ingredient Abstraction Network is Drug-Drug Interaction discovery.

4.1 Major Topic Identification

This section presents a study on the summarization of the "big picture" of an ontology by automatically deriving concept groups that represent major topics in a specific domain. It is a parameterized methodology to identify major topics in an ontology based on the weighted aggregate partial-area taxonomy, followed by manual enhancement. Because of SNOMED CT's importance in clinical applications and its large size, an experiment on its *Specimen* hierarchy is presented to test the effectiveness of such summarization

measured by “Big Knowledge” coverage of a given list of major topics related to the corresponding domain.

4.1.1 Partial-area Taxonomies for Major Topic Identification

A topic of an ontology, represented by a concept *c*, is considered a major topic if *c* has a large number of descendants. The above definition of major topic is based on the following two assumptions. First, it was assumed that concepts belonging to a given topic are all hierarchically related (i.e., they share a common ancestor concept *c* that represents and names the topic). That is, all the descendant concepts of a topic *c* belong to that topic, since they are specializations of *c*. Second, it was assumed that if there are relatively many concepts for a topic then it is “more important.” For example, there are 262 concepts related to digestive system specimens, but only 12 related to bone marrow specimens. Thus the topic “digestive system specimens” was considered as more important in SNOMED CT. Note that it is not necessarily clinically more important, since this depends on the clinical context.

The approach for evaluating the quality of the automatically identified major topics was based on a gold standard list provided by a domain expert. The domain expert, Dr. Gai Elhanan (GE), was asked to select a list of major topics for the specimen domain. (GE) is an MD with long experience in ontologies. A gold standard may also be derived from a published ontology of an authoritative organization. No other ontology for specimens was found, e.g., in the UMLS Metathesaurus. For the sake of normalization and to simplify the eventual matching task, each chosen topic was semi-automatically mapped to a SNOMED CT concept in the *Specimen* hierarchy, utilizing UMLS

synonyms. For example, the topic “Bone specimen” was mapped to the concept *Specimen from bone*.

One straightforward heuristic for identifying major topics in an ontology is to review the ontology root’s children, which are typically general and cover high-level topics. For example, *Specimen* has 59 children (e.g., *Biopsy sample* and *Blood specimen*). However, among the 59 children, many would not be considered major topics (based on the second assumption above), since they have few descendants. For example, 13 of *Specimen*’s children do not themselves have children (e.g., *Muscle specimen*). Nine have few children and no grandchildren (e.g., *Fibroblast specimen* has one child). Of the remaining 37 children, only 13 were in the major topic list of the domain expert, while another eight on that list were not children of *Specimen* (e.g., *Stool specimen* is a grandchild of *Specimen*). Hence, a better methodology for identifying major topics is required.

In previous studies, the SABOC team has derived a *partial-area taxonomy* [6] for each of the seven SNOMED CT hierarchies, including *Specimen*, that have outgoing lateral relationships. The partial-area taxonomies were not designed for the purpose of major topic identification, but for structure and content summarization. Indeed, the roots of partial-areas are not necessarily intuitive topics. A root of a partial-area is distinguished by the introduction of a new relationship type into the ontology, but it may or may not be a major topic. Moreover, a partial-area may be small, and thus, may not define a broad topic. A partial-area taxonomy typically has many small partial-areas [96]. As a result, the partial-area taxonomy for a large ontology, although smaller by an order of magnitude than the ontology, can still fail to identify major topics. Metaphorically, the

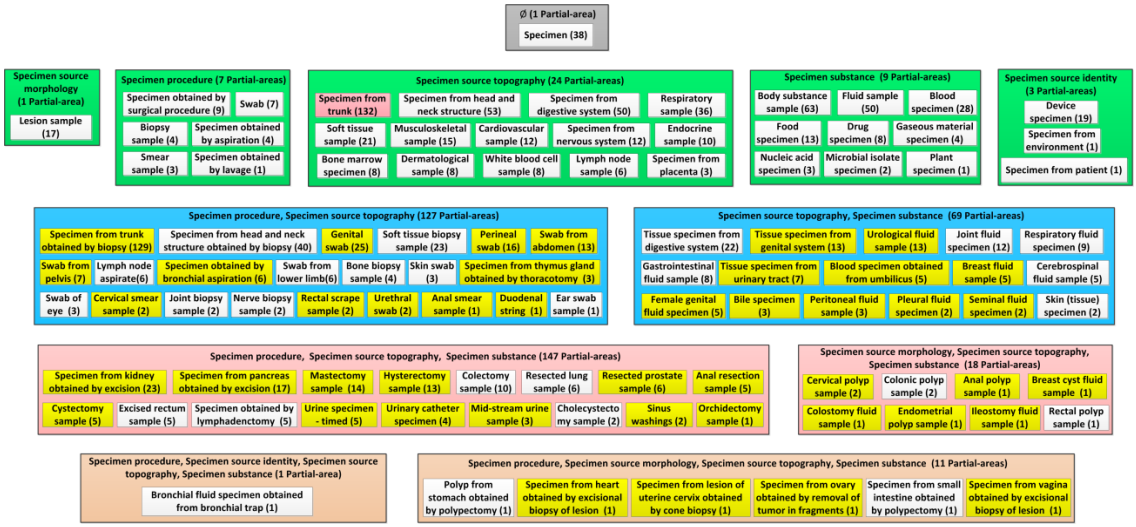

Figure 4.1 An excerpt of the partial-area taxonomy for the *Specimen* hierarchy. Partial-areas are sorted (left to right and top to bottom) according to their numbers of concepts. The yellow partial-areas are the descendant partial-areas of the pink partial-area *Specimen from trunk.* That is, there is a path of child-of relationships from any yellow partial-area to *Specimen from trunk.*

"forest" summary of the topics is not seen for the many small "trees" (see Figure 4.1). Figure 4.1 shows an excerpt of the partial-area taxonomy (with *child-of* links omitted) for the entire *Specimen* hierarchy.

Hence, a better solution for identifying major topics is to pick *only the large partial-areas* (with, e.g., dozens or more concepts). To illustrate these points, consider Figure 4.1, which shows an excerpt of *Specimen*'s partial-area taxonomy. Some concepts appear as (labels of) relatively large partial-areas. For example, *Specimen from trunk* (132), *Specimen from head and neck structure* (53), and *Specimen from digestive system* (50) from the area {*Specimen source topography*}, are partial-areas with 50 or more concepts. However, the seven large partial-areas account for only 536 *Specimen* concepts (33.1%). One may wonder about the topics of the other 66.9% of concepts.

Moving to medium-sized partial-areas with 20–49 concepts, eight partial-areas cover 218 (13.5%) concepts, e.g., *Blood specimen* (28) and *Soft tissue biopsy sample*

(23). Together, the large and medium partial-areas cover only 754 specimen concepts (46.5%). There are other problems with the summarization provided by the large/medium partial-areas. For example, all descendant partial-areas (yellow) of *Specimen from trunk* (pink) in Figure 4.1 contain concepts that are refinements of this topic. They are in separate partial-areas because they have (an) extra relationship(s) and appear in another area. For example, *Swab from abdomen* (13) has an additional *Specimen procedure* relationship. Overall, there are 201 partial-area descendants *of Specimen from trunk*, covering 551 concepts.

In summary, by only focusing on large and medium partial-areas, useful knowledge that is distributed among the many small partial-areas is ignored. Frequently, a large partial-area has many descendant small partial-areas. The concepts in these descendant partial-areas cover the same topic as the large parent/ancestor partial-area, but in more detail. Hence, they belong to the topic of the parent/ancestor partial-area.

4.1.2 Weighted Aggregate Partial-area Taxonomies for Major Topic Identification

During the aggregation process to derive the weighted aggregate partial-area taxonomy, small partial-areas are allowed to contribute to the identification of major topics. Since small partial-areas are folded into their larger ancestor partial-area(s), the lost knowledge in small partial-areas is accounted for. As an example, Figure 4.2 shows the aggregation process. The aggregated weight of the partial-area *Endocrine sample (10)* is 26, because it has nine descendant partial-areas summarizing 16 descendant concepts. Since the nine descendant partial-areas have no child partial-areas, their aggregated weights are the same as their original sizes, namely less than 11. Hence, in the weighted aggregate partial-area taxonomy with b=11 (Figure 4.2(b)), all the descendants partial-areas are

aggregated into and are represented by the partial-area *Endocrine sample*.

In the topic identification experiment, the threshold *b* was iterated over the range 1…30 and the weighted aggregate taxonomy was derived for each *b*. Each such weighted aggregate taxonomy was inspected to determine its effectiveness in capturing major topics. Precision, recall, and F measure [98] were calculated for each weighted aggregate taxonomy, with the expert's topic list serving as a gold standard. Recall is the ratio of the number of correctly identified topics and the number of total topics. Precision is the ratio of the number of correctly identified topics and the number of partial-areas.

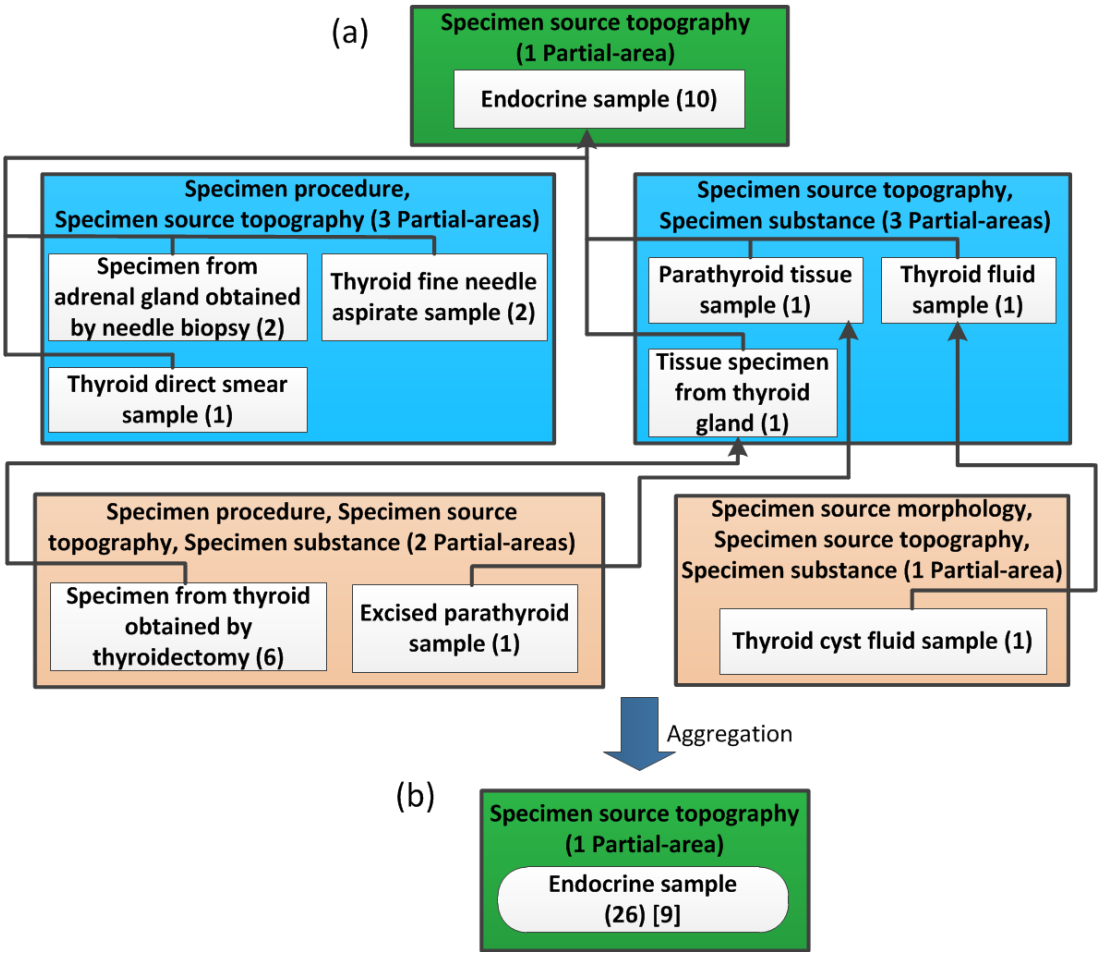

Figure 4.2 (a) An excerpt of 10 partial-areas **(b)** Weighted aggregate partial-area with b=11 for (a), shown as a rounded white rectangle with its number of concepts in ( ) including all concepts from aggregated partial-areas and the number of aggregated partial-areas in [ ].

As a preliminary experiment, it was determined how many of the gold standard topics appeared as partial-areas in the original partial-area taxonomy (not the weighted aggregate taxonomies). Out of the 21 topics chosen by the expert, 13 appear as partial-areas. This yields a recall of 0.62 (13/21) and, with 503 partial-areas in the taxonomy, very low precision of 0.03 (13/503). Note that many partial-areas are very small. In contrast, the weighted aggregate taxonomy, which eliminates the small partial-areas, is more effective. To balance recall and precision, the weighted aggregate taxonomy with the *b* value that maximizes the F measure was considered as optimal.

If the root concept *r* of a partial-area appears in the weighted aggregate taxonomy of threshold *b*, then *r* is considered a major topic identified by that weighted aggregate taxonomy; a corresponding checkmark "✓" is placed in Table 4.1. Otherwise, a dash "–" is written. For example, the topic *Bone marrow specimen* is captured by a partial-area *Bone marrow specimen* (8) with an aggregated weight 13 (Table 4.1). Therefore, it is identified by all weighted aggregate taxonomies with $b <= 13$ (b=1, 5, 10). However, for $b > 13$, *Bone marrow specimen* (8) is folded into a larger ancestor partial-area and disappears. No weighted aggregate taxonomy with $b > 13$ identifies the topic *Bone marrow specimen*. As another example, *Bone specimen* was not identified by the weighted aggregate taxonomy with any *b* value as major topic (Row 5 of Table 4.1), since its mapped SNOMED CT concept *Specimen from bone* (Row 5, Column 2 of Table 4.1) is not a root of a partial-area.

The bottom of Table 4.1 shows the totals of the identified topics for the respective aggregate taxonomies. For example, for b=5, the total is 13. Table 4.2 shows each weighted aggregate taxonomy's number of partial-areas (*A*), recall, precision, and F.

Recall is the ratio of identified topics and total topics ($R=C/S$, where S=21). Precision is the ratio of the identified topics and the number of partial-areas ($P=C/A$). For example, for b=25, the number of partial-areas is 29, the number of identified topics is 12, R=0.57, P=0.41 and F=0.48. Table 4.2 shows that b=25 yields the aggregate taxonomy where F is maximized. In this case, the weighted aggregate taxonomy captures 12 of the 21 topics. Figure 4.3 shows this weighted aggregate taxonomy with the 12 partial-areas identifying topics highlighted in yellow. The total number of concepts in these 12 aggregate partial-areas is 988, accounting for 61.0% (988/1620) of the concepts in the *Specimen* hierarchy.

Table 4.1 Identification Results for 21 Chosen Topics in Weighted Aggregate Taxonomies with Different Thresholds *b*

Topic	**Concept**	**Partial-area**	**Weight**	*b*=1	5	10	15	20	**25**	30
Blood specimen	*Blood specimen*	*Blood specimen* (28)	43	✓	✓	✓	✓	✓	✓	✓
Body substance sample	*Body substance sample*	*Body substance sample* (63)	498	✓	✓	✓	✓	✓	✓	✓
Fluid sample	*Fluid sample*	*Fluid sample* (50)	257	✓	✓	✓	✓	✓	✓	✓
Bone marrow specimen	*Bone marrow specimen*	*Bone marrow specimen* (8)	13	✓	✓	✓	–	–	–	–
Bone specimen	*Specimen from bone*	*Musculoskeletal sample* (15)	44	–	–	–	–	–	–	–
Specimen from nervous system	*Specimen from nervous system*	*Specimen from nervous system* (12)	42	✓	✓	✓	✓	✓	✓	✓
Dermatological specimen	*Dermatological sample*	*Dermatological sample* (8)	30	✓	✓	✓	✓	✓	✓	✓
Device specimen	*Device specimen*	*Device specimen* (19)	40	✓	✓	✓	✓	✓	✓	✓
Digestive system specimen	*Specimen from digestive system*	*Specimen from digestive system* (50)	126	✓	✓	✓	✓	✓	✓	✓
Endocrine system specimen	*Endocrine sample*	*Endocrine sample* (10)	26	✓	✓	✓	✓	✓	✓	–
Genital system specimen, male	*Male genital sample*	*Specimen from trunk* (132)	489	–	–	–	–	–	–	–
Genitourinary specimen	*Genitourinary sample*	*Specimen from trunk* (132)	489	–	–	–	–	–	–	–
Hair specimen, scalp	*Hair specimen*	*Dermatological sample* (8)	30	–	–	–	–	–	–	–
Musculoskeletal specimen	*Musculoskeletal sample*	*Musculoskeletal sample* (15)	56	✓	✓	✓	✓	✓	✓	✓
Skin specimen	*Specimen from skin*	*Dermatological sample* (8)	30	–	–	–	–	–	–	–
Soft tissue specimen	*Soft tissue sample*	*Soft tissue sample* (21)	92	✓	✓	✓	✓	✓	✓	✓
Cardiovascular sample	*Cardiovascular sample*	*Cardiovascular sample* (12)	28	✓	✓	✓	✓	✓	✓	–
Specimen from eye	*Specimen from eye*	*Specimen from head and neck structure* (53)	196	–	–	–	–	–	–	–
Specimen from joint	*Joint sample*	*Musculoskeletal sample* (15)	56	–	–	–	–	–	–	–
Lesion sample	*Lesion sample*	*Lesion sample* (17)	118	✓	✓	✓	✓	✓	✓	✓
Stool specimen	*Stool specimen*	*Body substance sample* (63)	498	–	–	–	–	–	–	–
# Identified topics (*C*)				13	13	13	12	12	12	10

Table 4.2 Performance of Weighted Aggregate Taxonomies for Various Thresholds

***b* =**	1	5	10	15	20	**25**	30
# Identified topics	13	13	13	12	12	12	10
# Partial-areas (*A*)	503	89	54	40	35	29	26
Recall (*R* = *C*/*S*)	0.62	0.62	0.62	0.57	0.57	0.57	0.48
Precision (*P* = *C*/*A*)	0.03	0.15	0.24	0.30	0.34	0.41	0.38
F = 2·*P*·*R*/(*P*+*R*)	0.05	0.24	0.35	0.39	0.43	**0.48**	0.43

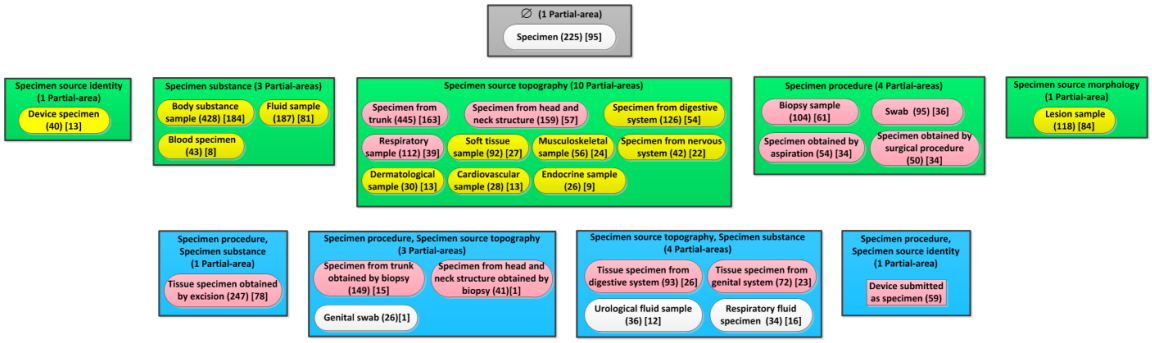

Figure 4.3 Weighted aggregate taxonomy for the Specimen hierarchy with ***b***=25. The 12 partial-areas corresponding to the original given topics are highlighted in yellow. The 13 topics added during the enhancement step are highlighted in pink.

An ancillary experiment was carried out as a feedback step with the domain expert (GE). When inspecting the weighted aggregate taxonomy for threshold *b*, one can assess whether its other partial-areas beyond those in the gold standard list are worthy of the designation "major topic," for example, those aggregate partial-areas in Figure 4.3 categorizing over 25 concepts that are not in the given list. Some important topics may have been overlooked originally, due to various reasons, e.g., *Specimen from head and neck structure*, a compound topic name with two body parts, and *Tissue specimen obtained by excision*, corresponding to two relationships *Specimen procedure* and *Specimen source topography*. Figure 4.3 was shown to (GE). He manually determined that 13 more partial-areas, highlighted in pink, warranted inclusion in the list of major specimen topics, while the other three (in white) are deemed as non-major topics. Reevaluating the experiment (with 21+13=34 major topics), it was obtained that R=0.74

(=25/34), P=0.86 (=25/29) and F=**0.79** for *b*=25. The number of concepts in these 25 aggregate partial-areas is 1,524 (94.1% of the concepts in the *Specimen* hierarchy.)

In summary, summarizing a large ontology is a challenge as there is a lack of an objective, universally accepted criterion for what constitutes a "good summarization" of an ontology. Various applications require different summaries of various granularities. Nevertheless, the management of ontologies requires "big picture" comprehension that can be enabled by compact summarization networks such as the w*eighted aggregate partial-area taxonomies* introduced in this dissertation.

This study utilized a knowledge-oriented approach, where the importance of a topic is based on the number of concepts related to that topic in an ontology. To measure the quality of the summarization, the number of identified major topics was compared with a gold standard list of topics selected by a domain expert, who selected topics from a clinical perspective. The results showed that the weighted aggregate partial-area taxonomy is viable as a method for capturing the major topics of a domain.

4.2 Multi-layer Big Knowledge Visualization Scheme for Comprehending Neoplasm Ontology Content

Visualizing Big Knowledge is challenging, due to the inherent complexity of knowledge. Consider, for example, an ontology as a structured knowledge repository. An ontology consists of a network of nodes (called concepts) interconnected by hierarchical relationships and semantic lateral relationships. Large ontologies contain several hundred thousand concepts and at least as many relationships.

Figures are an effective way of presenting knowledge in a comprehensible format. Indeed, the knowledge in an ontology is often presented as a node-link diagram [99]. In

this type of visualization, concepts are displayed as nodes labeled with their names. Node-link diagrams for small portions of an ontology can be displayed on a single computer screen or on a single printed page. However, visualizing larger portions of an ontology requires larger and more complex figures, which pose significant problems related to the comprehension capacity of humans (which is limited, independent of the screen size).

Heuristically, node-link diagrams become overwhelming if more than about 20 to 30 nodes, and their associated links, are displayed [100, 101]. A figure with more nodes (and thus, more links between nodes) most likely will pose a challenge to the mental comprehension of most humans. Assuming that there is a capacity limit on comprehension, how can humans cope with comprehending large ontologies with hundreds of thousands of concepts?

Large ontologies are often divided into disjoint subhierarchies, each dedicated to a specific topic. However, even the individual subhierarchies are typically far beyond a human's comprehension ability. For example, the *Disease, Disorder or Finding* hierarchy of the National Cancer Institute (NCI) thesaurus (NCIt) [3] contains 27,045 concepts. The *Neoplasm* subhierarchy, dedicated to neoplastic diseases like *Cancer*, contains 8,445 concepts.

The overwhelming complexity of Figure 1.1 illustrated the challenges when trying to comprehend the contents of Big Knowledge repositories such as ontologies, in addition to the technical limitations associated with generating such a view on a computer screen (e.g., a lack of screen space and limited human visual acuity).

Without some level of mental comprehension of the contents of a large body of knowledge, humans will experience difficulty using the knowledge for innovative and sophisticated applications [95]. Furthermore, the curator(s) in charge of maintaining large knowledge repositories require such comprehension to achieve correct and exact modeling of the knowledge.

How can a human achieve comprehension of large hierarchies of concepts? As can be seen in Figure 1.1, the amount of knowledge expressed in a node-link diagram can be overwhelming to the point of being unusable.

The approach outlined in this section is a vision for coping with the "Big Knowledge challenge" using two techniques. First, Big Knowledge is automatically summarized into a compact visualization that captures the "big picture" of the knowledge by hiding less important details. This summary view allows a user to concentrate on the "major subjects" in a knowledge base. An interactive mechanism for recovering details that were lost in the summarization process allows a user to obtain more information on-demand. Hence, in this approach, details are not lost, they are only hidden until they are exposed upon request. The second part of this approach is based on the heuristic that humans struggle to comprehend a node-link diagram with more than 30 labeled nodes. Thus, when visualizing Big Knowledge summaries, a limit on the number of named nodes will be strictly enforced.

The contribution of this section consists of a multi-layer interactive system, based on the theory of Abstraction Networks, where the initial "big picture" summary presented to a user is of low granularity. From this view, a user "drills down" into more details as desired. Such a dynamic system enables a user to navigate through several layers of

summarization, where at each point in time the user views only one network with a fixed number of nodes (specifically, an example limit of 25 is enforced in this section).

4.2.1 Hypothesis for Limited Human Comprehension Capacity

In humans, working memory is a cognitive system with a limited capacity. It is responsible for temporarily holding information available for processing [102]. Working memory is important for reasoning and decision making. Although there are various hypotheses about the quantitative measure of working memory capacity, the general consensus is that there is a limit to that capacity.

One of the most famous papers in cognitive psychology is "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" by Miller [103]. According to Miller, the number of objects an average human can hold in working memory is 7 ± 2. A more recent paper by Cowan [104] identifies a smaller number (four items).

The partial-area taxonomy created from the 8,445 neoplasm concepts in the September 2016 NCIt release consists of 4,177 partial-areas in 1,301 areas. Although the partial-area taxonomy is smaller than the underlying ontology, following the theory of limited working memory, it is still far too large for a human to comprehend its contents or the contents of the partial-area taxonomy.

Based on the above heuristic, the hypothesis that an Abstraction Network (such as a partial-area taxonomy) with no more than 30 nodes, when displayed as a figure, will be comprehensible to a human was formulated. Therefore, a more compact summary than is provided by a partial-area taxonomy is required. As noted before, the weighted aggregate partial-area taxonomy offers a more compact summary of a size that can be controlled. It

will now be used to develop a multi-layer display where each layer is limited in its complexity.

4.2.2 Multi-layer Visualization Scheme for Big Knowledge

The technique defined in this section provides a compact summary of an ontology in the form of an aggregate taxonomy, created in a software system by using an automatically determined parameter *b* at each layer. Following the vision outlined above, the system should be able to expand a given aggregate partial-area to show its details (i.e., its constituent "small" partial-areas). The expansion process is the inverse procedure of aggregation. For example, if a user wants to see which partial-areas are summarized by the aggregate partial-area *Colorectal Carcinoma {19}(32)[6]* in Figure 3.1, then he/she could obtain the details shown in Figure 3.1(a) by "re-expanding" that aggregate partial-area in the interactive system. This corresponds to a drill down operation.

Expanding an aggregated partial-area into its constituent small partial-areas can result in a view that is overwhelming when the aggregate partial-area summarizes many small partial-areas. Thus, it may be required to apply the aggregate taxonomy process recursively, with a different automatically selected bound *b*, on the subhierarchy of small partial-areas that is to be displayed.

For example, if the aggregate partial-area *Malignant Digestive System Neoplasm {26} (467) [203]* is expanded with 203 "small" partial-areas in the *Neoplasm* aggregate taxonomy shown in Figure 4.4, the resulting partial-area taxonomy will have 204 partial-areas (far more than the recommended limit of 25).

Thus, it is necessary to apply the aggregation process to the resulting partial-area taxonomy to obtain an aggregate taxonomy with no more than 25 aggregate partial-areas.

The automatically identified parameter *b*=8 makes the resulting aggregate taxonomy for *Malignant Digestive System Neoplasm* have 24 nodes, in the range of human comprehension ability. This bound is significantly lower than the parameter *b*=200, which limits the number of aggregate partial-areas in the *Neoplasm* aggregate taxonomy to 25.

With repeated applications of this process, a multi-layer visualization scheme is obtained, where each summarizing view has at most 25 aggregate partial-areas. The first summary, of the least granularity, summarizes the entire ontology (or a selected subhierarchy of an ontology, e.g., *Neoplasm*). The second layer of summarization summarizes a chosen subject (e.g., *Malignant Digestive System Neoplasm*). Further summaries, as obtained by expanding additional aggregate partial-areas, will display more details for more specific subjects (e.g., *Colorectal Carcinoma* selected from within the *Malignant Digestive System Neoplasm* aggregate taxonomy). The final summary will be a partial-area taxonomy, with at most 25 partial-areas. From this summary, a user could "drill down" to individual concepts in individual partial-areas. This dynamic multi-layer visualization system enables a user to view details in a desired part of the "big picture" through recursive application of the same aggregate-taxonomy-based summarization methodology.

The multi-layer visualization scheme is demonstrated below using the *Neoplasm* subhierarchy in NCIt as a test bed. There are 27 role types defined for the *Disease, Disorder or Finding* hierarchy. To simplify the following figures, the 27 roles as coded as numbers in the figures. Table 4.3 shows the code numbers (index numbers) representing the 19 roles that appear in the following figures.

Table 4.3 The Index Numbers for the Roles in NCIt Appearing in this section

Role Type	Index Number
Disease Excludes Abnormal Cell	1
Disease Excludes Finding	3
Disease Excludes Normal Cell Origin	5
Disease Excludes Normal Tissue Origin	6
Disease Excludes Primary Anatomic Site	7
Disease Has Abnormal Cell	8
Disease Has Associated Anatomic Site	9
Disease Has Associated Disease	10
Disease Has Finding	12
Disease Has Normal Cell Origin	15
Disease Has Normal Tissue Origin	16
Disease Has Primary Anatomic Site	17
Disease Is Grade	18
Disease Is Stage	19
Disease Mapped To Gene	21
Disease May Have Associated Disease	23
Disease May Have Cytogenetic Abnormality	24
Disease May Have Finding	25
Disease May Have Molecular Abnormality	26

Layer 1: *Neoplasm* Aggregate Taxonomy

Figure 4.4 shows the aggregate taxonomy for the *Neoplasm* subhierarchy with *b*=200 (the smallest *b* resulting in an aggregate taxonomy with [at most] 25 nodes). There are 25 aggregate partial-areas, shown in white. This would be the first layer of summarization for the *Neoplasm* subhierarchy. Note that the aggregate partial-areas shown in Figure 4.4 have root concepts that are general in nature and each summarizes a large number of concepts. This view provides the "big picture" of the contents of the *Neoplasm* subhierarchy. For example, the aggregate partial-area node *Malignant Digestive System Neoplasm {26} (467) [203]* summarizes 467 neoplasm concepts, about 20% of the *Neoplasm* subhierarchy.

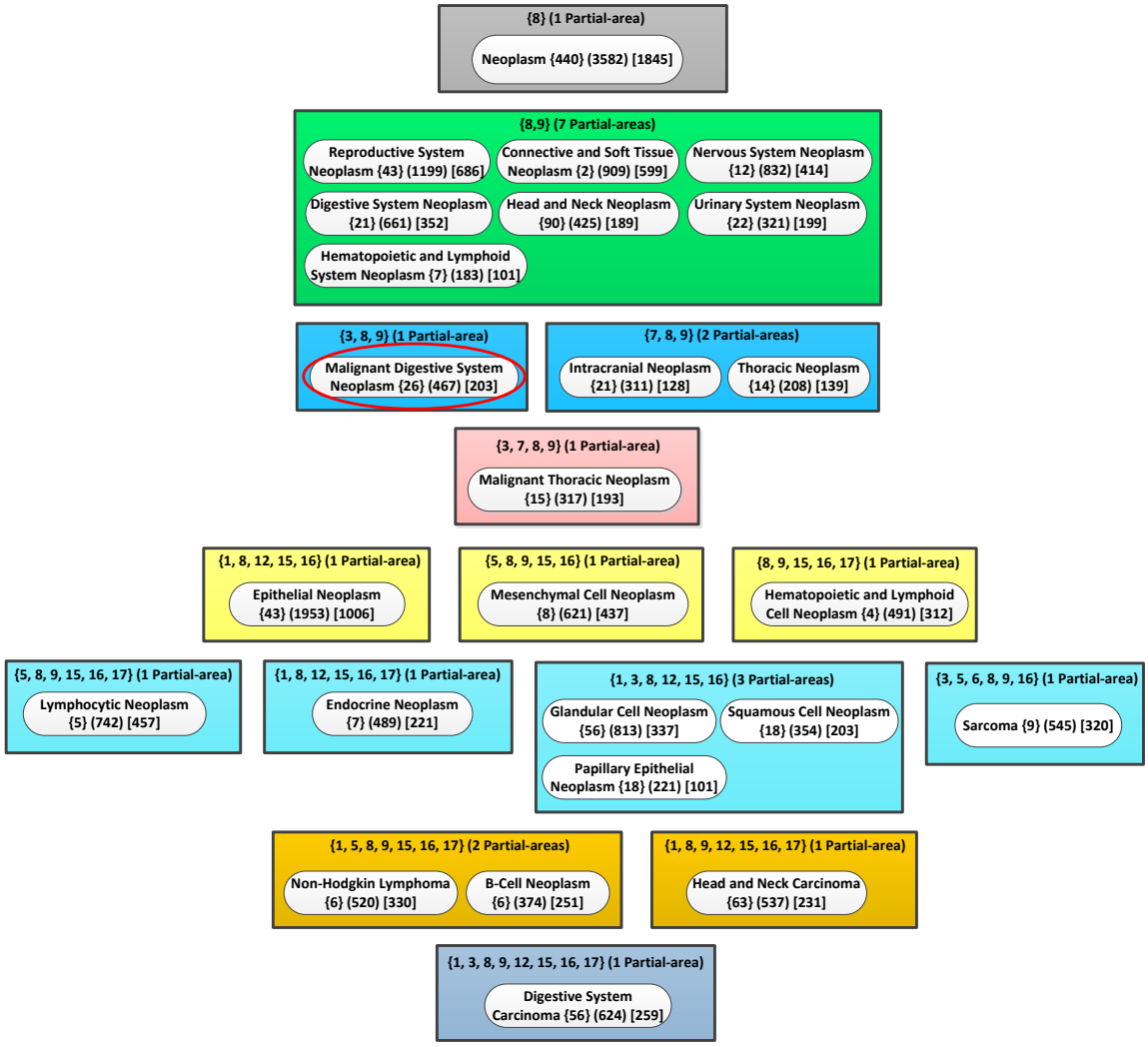

Figure 4.4 *Neoplasm* Aggregate Taxonomy with 25 aggregate partial-areas (*b*=200).

Layer 2: *Malignant Digestive System Neoplasm* Aggregate Taxonomy

A user may be interested in the concepts summarized by *Malignant Digestive System Neoplasm* (surrounded by a red ellipse in Figure 4.4). In this case, the user can expand this aggregate partial-area. Since the partial-area taxonomy obtained from expanding *Malignant Digestive System Neoplasm* has more than 25 partial-areas (namely 204), the aggregation process is recursively applied to obtain an aggregate taxonomy with at most 25 aggregate partial-areas (this time using *b*=8).

Figure 4.5 shows the aggregate taxonomy for the *Malignant Digestive System Neoplasm* subhierarchy with b=8, which is composed of 24 aggregate partial-area nodes. Comparing Figure 4.4 with Figure 4.5, the number of concepts summarized by each aggregate partial-area is much smaller in Figure 4.5. This is due to the aggregate partial-areas in Figure 4.5 capturing more specific subjects. This view captures a relatively small part of the "big picture" in Figure 4.4.

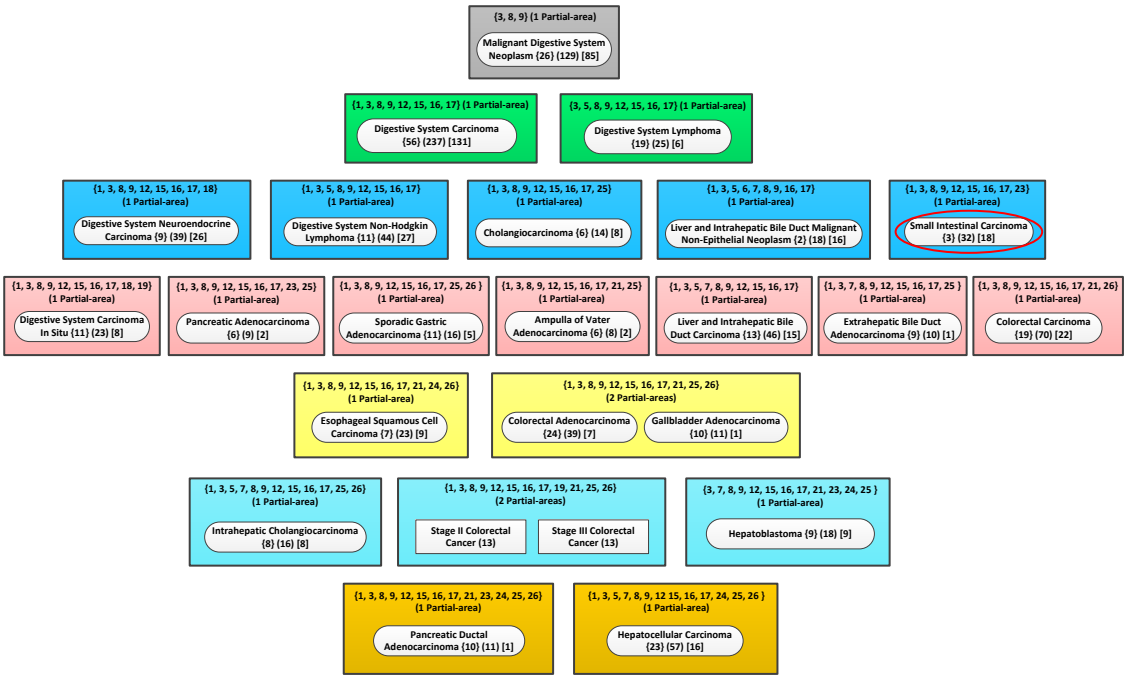

Figure 4.5 *Malignant Digestive System* Neoplasm (from Figure 4.4) Aggregate Taxonomy with 24 aggregate partial-areas (b=8).

Layer 3: *Small Intestinal Carcinoma* Partial-area Taxonomy

Note that all nodes in Figure 4.5 are aggregate partial-areas (shown as rounded corner white rectangles) that summarize at least one descendant partial-area, except for *Stage II Colorectal Cancer (13)* and *Stage III Colorectal Cancer (13)* (shown as white rectangles with sharp corners). In this view, the expansion and aggregation process can be applied again to get a more detailed picture with at most 25 partial-areas for any

aggregate partial-area in the figure. For example, if *Small Intestinal Carcinoma* in Figure 4.5 (again, marked by a red ellipse) is expanded, the resulting partial-area taxonomy has only 19 partial-areas. Thus, there is no need to apply aggregation after the expansion. Figure 4.6 shows the partial-area taxonomy for the *Small Intestinal Carcinoma* subhierarchy.

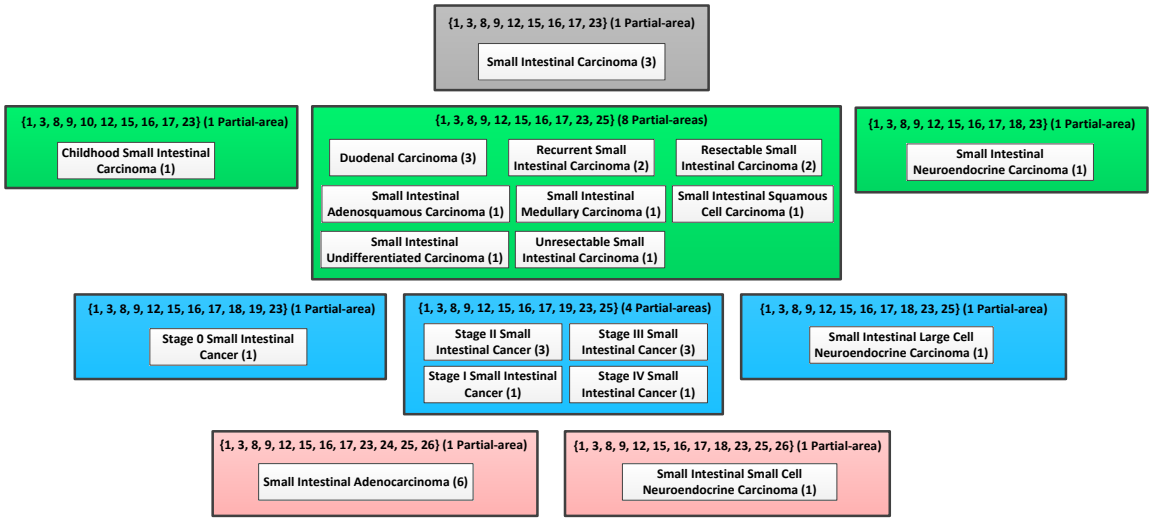

Figure 4.6 *Small Intestinal Carcinoma* (from Figure 4.5) Partial-area Taxonomy with 19 partial-areas.

The multi-layer visualization scheme for ontologies has been integrated into the Ontology Abstraction Framework (OAF) software tool, described in detail by Ochs et al. [67]. The OAF provided all of the necessary modules for this visualization scheme, namely, the aggregate partial-area taxonomy module for the aggregation process, and the expanded sub-taxonomy module for the expansion process.

In this study, the *Neoplasm* subhierarchy of NCIt was utilized to demonstrate the methodology. However, the OAF tool supports ontologies in various formats. Since the multi-layer visualization scheme is fully integrated into the OAF, the technique described

in this section will be applicable to many ontologies, thus enabling users to obtain a better understanding of the Big Knowledge in the displayed ontologies.

The multi-layer visualization scheme is based on the heuristic that a human has a limited comprehension capacity that was assumed to be about 25 nodes in a node-link diagram. In future work, evaluation studies will be conducted to test this heuristic. This section proposed a general process for supporting the comprehension of Big Knowledge through summarization. In the future, usability studies will be performed to evaluate the effectiveness of this technique.

To conclude, comprehension of Big Knowledge is a significant challenge. In this study, a multi-layer visualization scheme was described for Big Knowledge repositories, which are ontologies in this research. The approach was based on Abstraction Networks, which, using a process of aggregation, can be tuned to automatically limit the amount of information presented to a user. This technique was illustrated using the *Neoplasm* subhierarchy of the NCIt ontology.

4.3 Application of the IAbN to Drug-Drug Interaction Discovery

A Drug-Drug Interaction (DDI) is a particularly important type of Adverse Drug Reaction (ADR) [105-108] that can cause excessive responses or altered toxicity [109]. The risk of adverse DDIs increases exponentially for each additional medication [110-114]. One application of the IAbN is the discovery of candidate drug-drug interactions missing from existing DDI knowledge bases.

The rationale is that drugs with similar chemical ingredients tend to have similar DDIs [97]. Given DDIs of the form ($DrugConcept_1$, $DrugConcept_2$,

ClinicalConsequence), each $DrugConcept_1$ and $DrugConcept_2$ element is coded as a concept in the NDF-RT's *Chemical Ingredients* hierarchy [115]. In this study, the DDI knowledge base from First Databank (FDB) [116] was used as the test-bed to demonstrate the approach. By reviewing the known DDIs in FDB associated with the chemical ingredients in an IAbN ingredient group, one may discover candidate DDIs missing from First Databank's DDI knowledge base [11].

Figure 4.7 illustrates the approach. There are 18 drug ingredients summarized by the IAbN's *Salicylates* ingredient group, including its child ingredient group *Aminosalicylic Acid* (the two yellow highlights in Figure 3.5). Out of the 18 NDF-RT *Salicylates* ingredients, 13 ingredients appear in FDB's DDI knowledge base. The DDI interactions between ten of these salicylates and seven anticoagulant drugs are "Avoid concurrent use when possible" (AVD) and "Increases the effect of latter drug" (INL), for a total of 70 DDIs between these two groups. However, three extra *Salicylates* (balsalazide, mesalamine, and salsalate) have no DDIs with any anticoagulant in the FDB

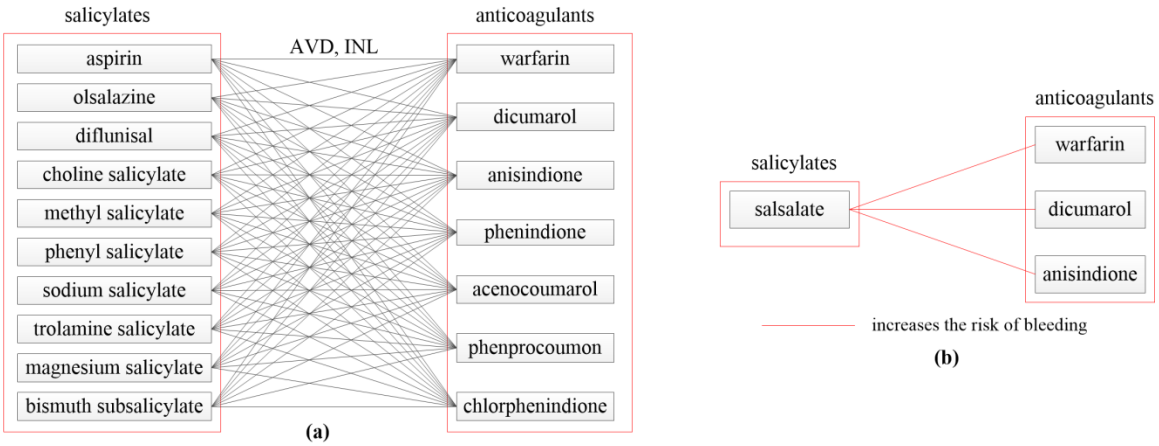

Figure 4.7 (a) Illustration of 70 DDIs. There are 10x7=70 DDIs between the ten salicylates on the left and the seven anticoagulants on the right in FDB's DDI knowledge base. AVD = "Avoid concurrent use when possible" and INL = "Increases the effect of latter drug." **(b)** Three new candidate DDIs not appearing in FDB's DDI knowledge base, between the Salicylate *Salsalate* on the left and the three anticoagulants on the right.

DDI knowledge base. This raised doubts regarding the existence of DDIs between the seven anticoagulants and these three salicylates. Indeed, upon investigating the DDIs between the three extra salicylates and these seven anticoagulants in another public source, Drugs.com (https://www.drugs.com/), DDIs between one salicylate *Salsalate* and three of the anticoagulants shown in Figures 4.7(b) were discovered. The reason FDB did not include these candidate DDIs in their knowledge base is that in these cases the drug formulation has a low potential for interactions. Nevertheless, FDB staff (Joan Kapusnik-Uner) confirmed that this example demonstrates the fact that summaries of NDF-RT have the potential for supporting the discovery of new candidate DDIs. Of course, pharmacological investigation is required for each potential DDI.

One has to realize that, as a leading Pharmacological knowledge company, FDB's DDI knowledge is widely used by pharmacies, doctors and hospitals for decisions about preventing patients from taking drugs prescribed due to their conditions. These are critical clinical decisions that sometimes involve issues of life or death. Other DDI sources, like drugs.com, which are not used in this way in the healthcare industry, are thus more lenient in including questionable pairs of drugs as DDIs.

Similar to the above study on one drug pair, another study was performed to examine several pairs of families of drugs, known to have DDIs, in search of drug pairs with potential DDIs that are not listed in the FDB knowledge base. For each such pair, one family is a chemical classification while the other is a pharmaceutical classification. Typically, the NDF-RT contains more drugs under the chemical classification than the FDB DDI knowledge base. Drugs.com was explored for DDIs for those additional drugs from the NDF-RT, looking for interactions with the drugs that are classified by the

corresponding pharmaceutical classification.

Table 4.4 reports the details of this study for seven pairs of families. Column 2 lists the DDI family pairs (A, B) as given in the FDB DDI knowledge base. For all seven pairs, A is a *Chemical Ingredient* family and B is a *Pharmaceutical* family. Column 3 represents the number of ingredients in Family A (e.g., *Sulfonamides*) in FDB. Column 4 shows the number of ingredients in Family B (e.g., *Antidiabetics, Oral*) in FDB. Column 5 gives the number of drug DDI (A, B) pairs in FDB, which is the product of the number in Column 3 and the number in Column 4. Column 6 represents the number of ingredients in Family A in the NDF-RT. Column 7 shows the number of ingredients in Family A in both NDF-RT and FDB. Column 8 shows the total number of potential DDIs found in other sources than FDB's knowledge base for the specific (A, B) drug pairs.

For example, the first pair is (*Sulfonamides*; *Antidiabetics, Oral*). In FDB the DDIs between six sulfonamides (Family A) and eight antidiabetics (Family B) have the clinical effect "INL" (Increased effect of the latter drug). However, in the NDF-RT there are 52 drugs classified under *Sulfonamides.* When pairing the additional (i.e., not in FDB's knowledge base) 48 A drugs with those eight B drugs, 93 pairs of drugs were found in the other sources as DDI pairs between Family A and Family B. An analysis of these 93 pairs revealed that for (A, B) pairs the interaction between the two drugs consists of a "protein binding displacement mechanism" that applies only for sulfonamide *antibiotics,* which describes the six sulfonamides listed in the FDB DDI knowledge base. This mechanism does not apply to the remaining sulfonamides found in the NDF-RT. One outcome from this study is that FDB will change, in its DDI knowledge base, the name of this Family A from "Sulfonamides" to "Sulfonamide Antibiotics," which

describes it more accurately.

Out of the 73 potential DDIs (Table 4.4, line 3) for the family pair (NSAID, ACE Inhibitor or ARBS) found in other sources, 16 potential DDIs for the *Nonsteroidal Anti-inflammatory Compound* (NSAID) *Diflunisal* with 16 different ACE Inhibitors should be added to the FDB knowledge base. According to line 5 in the table, eight potential DDIs for the NSAID *Diclofenac* with various Beta Blockers were found. These should be considered for addition to FDB's knowledge base. Similarly, for its salt form *Diclofenac Potassium*, nine DDIs were found to be missing from FDB's knowledge base. However, the clinical studies reported in drugs.com for these pairs did not prove the DDIs to be at the stricter level required for inclusion by FDB. For the two NSAIDs *Meclofenamate* and *Mefenamic Acid*, which are currently not on the US market, eight and 11 DDIs, respectively, were found with various Beta Blockers. They should be added to FDB's knowledge base for the case that these drugs will be made available for sale in the US.

The DDI family pair in line 6 of Table 4.4 is known to cause many DDI alerts with a low severity level, i.e., it is a prime example of a combination that causes alert fatigue. Thus, line 6 interactions will be recommended for removal from FDB's DDI knowledge base. A detailed pharmacological analysis of these various families of drugs will appear in a future publication.

To summarize the results of this study, out of 394 potential DDI drug pairs found in other sources, 80 (20.3%) were approved for inclusion in FDB's DDI knowledge base. Another 19 (4.8%) drug pairs will be added if their drugs are made available in the US. The following additional outcomes of this study do not relate to the potential DDI pairs of Column 8, but to the actual FDB DDI pairs of Column 5 in Table 4.4. One such

important outcome was that 66 DDI drug pairs for the family pair (Phenotiazines, Narcotics) were removed from the FDB DDI knowledge base, since they almost always cause false alerts.

Even more interestingly, a deeper analysis of the FDB DDI drug pairs from three family pairs (rows 3, 4, and 5 of Table 4.4) showed that the issue was not the interaction of one drug with another drug, but the interaction between one drug and the *disease* that is present when the other drug is used. These interactions will be removed from the FDB *DDI* knowledge base, since the adverse drug reaction (ADR) is between a drug and a disease, rather than an interaction between two drugs. The relevant ADR knowledge will be placed in the proper FDB ADR knowledge base. The total number of DDI pairs removed is 1632+490+324=2446. Hence, beyond the additional DDI pairs, the study led to important changes in the storage of DDI and ADR knowledge in the FDB knowledge base.

Table 4.4 Potential DDI Findings for Seven Pairs of Drug Families

	2	3	4	5	6	7	8
	DDI Family Pair (A, B)	**# of Ingredients in Family A in FDB**	**# of Ingredients in Family B in FDB**	**# of Actual FDB DDIs**	**# of Ingredients in Family A in NDF-RT**	**# of Ingredients in Family A in Both FDB and NDF-RT**	**# of Potential DDIs Found in Other Sources**
1	(Sulfonamides, Antidiabetics, Oral)	6	8	48	52	4	93
2	(Sulfonamides, Anticoagulants)	22	6	132	52	15	21
3	(NSAIDs, ACE Inhibitors or ARBs)	68	24	1632	43	21	73
4	(NSAIDs, Loop Diuretics)	70	7	490	43	21	16
5	(NSAIDs, Beta Blockers)	12	27	324	43	7	81
6	(Phenothiazines, Narcotics)	6	11	66	32	5	73
7	(Benzodiazepines, Macrolide Antibiotics)	4	5	20	22	3	37
Total:		n/a	n/a	2712	n/a	n/a	394

CHAPTER 5

FAMILY-BASED QUALITY ASSURANCE OF BIOMEDICAL ONTOLOGIES

As described in Chapter 2 and Chapter 3, the SABOC team has developed different Abstraction Network-based quality assurance (QA) techniques for individual biomedical ontologies. To improve the efficiency of the Abstraction Network-based QA methodology, Ochs et al. [8] have classified BioPortal [117] ontologies into families according to ontologies' structural features and have introduced a family-based QA approach such that one QA methodology could be applicable to a whole family of structurally similar ontologies. Statistically, in order to correctly draw the conclusion that a QA technique is likely to work for at least half of the ontologies in a family, the QA methodology has to be demonstrated successfully for six out of six sample ontologies in the family.

It has been demonstrated that two main characterizations of concepts – complex concepts and uncommonly modeled concepts – are more likely to have errors for individual ontologies (Section 2.3). To demonstrate the effectiveness of these two characterizations for six ontologies in the same family, this chapter presents several Abstraction Network-based quality assurance studies on some ontologies in the same family. These ontologies are the NCIt's *Neoplasm* subhierarchy of the *Disease, Disorder or Finding* hierarchy, the *Gene* hierarchy and the *Biological Process* hierarchy, SNOMED CT's *Infectious Disease* subhierarchy of the *Clinical finding* hierarchy, the ChEBI ontology and NDF-RT's *Chemical Ingredients* hierarchy.

5.1 Quality Assurance of Complex Concepts

In the long-range research of the SABOC team [54], a repeated theme in QA of ontologies has been that "complex" concepts tend to have a significantly higher error rate than "simple" concepts. There are various interpretations of "complex concept" for different methodologies and different ontologies. A likely explanation is that the human activity of modeling complex concepts is more challenging and thus there is more room for errors in the modeling of a complex concept. This section mainly involves two types of complex concepts, overlapping concepts and concepts with more lateral relationship types. The latter concepts, laterally complex concepts, intuitively can be deemed to be more complex than a concept with fewer lateral relationship types. Two studies on this type of concepts in the NCIt's *Biological Process* hierarchy and the ChEBI ontology are presented in the following sections.

Overlapping concepts are hierarchically complex, because they derive semantics from two or more source concepts that are roots of partial-areas in the partial-area taxonomy. For example, the concept *Papillary Serous Cystadenoma* in Figure 2.6(a) inherits semantics from three partial-area roots *Serous Neoplasm*, *Cystadenoma*, and *Papillary Cystic Neoplasm*. Its two parents *Serous Cystadenoma* and *Papillary Cystadenoma* are themselves overlapping concepts. Hence, from the view point of hierarchy complexity, *Papillary Serous Cystadenoma* is more complex than its two parents, which in turn are more complex than the area roots.

The SABOC team has demonstrated that overlapping concepts are more likely to have errors than non-overlapping concepts for three ontologies [16, 91-93]: the *Specimen* hierarchy of SNOMED CT, the *Bleeding* subhierarchy in the *Clinical finding* hierarchy of

SNOMED CT [7], and the Uber Anatomy Ontology (Uberon) [94]. In order to confidently make a statement that concentrating on overlapping concepts constitutes a successful methodology for a whole family of ontologies, the effectiveness of the methodology needs to be shown for six out of six sample ontologies. To achieve six out of six, studies on overlapping concepts in three more ontologies that belong to the same family as the above mentioned three ontologies will be presented in the following sections. These six ontologies belong to the family of ontologies in BioPortal with (a) object properties used only in restrictions and (b) with multiple parents allowed.

5.1.1 Quality Assurance of Complex Neoplasm Concepts in NCIt

The QA study on overlapping neoplasm concepts in NCIt presented in this section was conducted with the goal to add a fourth ontology to the set of three ontologies (the *Specimen* hierarchy and the *Bleeding* subhierarchy of SNOMED CT and Uberon) for which the QA methodology of overlapping concepts was previously shown as effective.

5.1.1.1 Methods. As mentioned in Section 2.1.3, the concepts in the *Neoplasm* subhierarchy of the NCIt's *Disease, Disorder or Finding* hierarchy are modeled with more details, compared to the other concepts in the *Disease, Disorder or Finding* hierarchy. Furthermore, according to Ochs et al. [8], the *Neoplasm* subhierarchy has the same structural features as the above three ontologies for which the "overlapping concept" QA methodology was demonstrated as effective. Hence, this QA study concentrated on the *Neoplasm* subhierarchy. Although the number of concepts (8,166) in the *Neoplasm* subhierarchy is much smaller than that of the complete *Disease, Disorder or Finding* hierarchy (25,360), it was still impossible to review all concepts in the *Neoplasm* subhierarchy, considering the reality of limited QA resources.

The disjoint partial-area taxonomy for the *Neoplasm* subhierarchy clearly distinguishes between overlapping concepts and non-overlapping concepts. In this study overlapping concepts were considered as complex concepts and non-overlapping concepts were considered as simpler concepts which serve as control concepts. The following hypothesis was investigated.

Hypothesis 5.1: Overlapping concepts are more likely to have errors than non-overlapping concepts in the disjoint partial-area taxonomy for the *Neoplasm* subhierarchy of the *Disease, Disorder or Finding* hierarchy of NCIt.

Hypothesis 5.1 is of practical importance. If Hypothesis 5.1 is confirmed with statistical significance, then the disjoint partial-area taxonomy can be viewed as a fully automatic screening test that identifies sets of concepts with a likely higher error yield than other neoplasm concepts, defined by the ratio of the number of discovered errors to the number of reviewed concepts. Thus, it is justified to invest QA resources, such as the time of domain experts, into a careful review of overlapping concepts.

A randomized controlled trial was conducted on a sample of neoplasm concepts to evaluate Hypothesis 5.1. The *Neoplasm* disjoint partial-area taxonomy contains exactly 225 overlapping concepts, which were used as the study concepts. A sample of 350 non-overlapping concepts from the same areas that the study concepts came from was randomly picked as a control group. Since concepts in small partial-areas are prone to have more errors, as mentioned in Section 2.3, the control population excluded such concepts. The study concepts and control group concepts were combined into a list. The order of the concepts in the list was randomized and the resulting list was presented to two domain experts for review.

The two domain experts, Dr. Gai Elhanan and Dr. Yan Chen, were trained in medicine and have extensive terminology QA experience. The QA study consisted of three steps. First, the two experts reviewed all 575 concepts independently. Each of the reviewers generated a report of errors with reasons, error severities (moderate or severe) and suggested corrections. Non-critical errors were not reported. In the second step, a combined list of errors reported by the two experts in the first step was created and presented to the same two reviewers. They had to express agreement or disagreement with each error in the list. The information of who had marked a concept as erroneous in the combined list was not included to avoid biased results.

In the third step, all concepts that were considered erroneous by only one reviewer in the second step were eliminated. Concepts on the list were then divided according to whether they came from the study group (overlapping concepts) or from the control group (non-overlapping concepts) and the numbers of errors were counted. The two tailed *p*-value of Fisher's exact test [118] was calculated to evaluate the statistical significance of the different error rates for overlapping concepts and for non-overlapping concepts.

5.1.1.2 Results. The partial-area taxonomy for the *Neoplasm* subhierarchy of the February 2015 release of the NCIt was first derived. The 8,166 neoplasm concepts are summarized by 920 areas and 4,824 partial-areas in this partial-area taxonomy. The partial-area taxonomy for the complete *Disease, Disorder or Finding* hierarchy of 25,360 concepts contains 986 areas and 5,080 partial-areas.

Comparing the numbers of areas and partial-areas for the *Neoplasm* subhierarchy versus the whole *Disease, Disorder or Finding* hierarchy, 95% (4,824/5,080) of the

Disease, Disorder or Finding partial-area taxonomy summarize all the neoplasm concepts, which account for only 32% (8,166/25,360) of the complete *Disease, Disorder or Finding* hierarchy. The remaining 68% of the hierarchy are covered by only 5% of the partial-areas. In order to perform a direct quantitative comparison, the *abstraction ratio* of a partial-area taxonomy is defined as the average number of concepts summarized per partial-area. The abstraction ratio for the *Neoplasm* subhierarchy is 1.69 (=8,166/4,824) and the standard derivation is 6.49, while the abstraction ratio is 4.99 (=25,360/5,080) and the standard derivation is 201.55 for the whole *Disease, Disorder or Finding* hierarchy. A lower number is indicative of more structural and semantic diversity, which is the result of detailed modeling efforts. The structural diversity is due to the large average number (23) of roles per neoplasm concept, since every combination of roles defines a different area. Thus, the structural diversity is reflected in the large number of areas. The semantic diversity is borne out by the many partial-areas.

The partial-area taxonomy for the complete *Disease, Disorder or Finding* hierarchy has 396 overlapping concepts. Among those, 225 overlapping concepts are in the *Neoplasm* partial-area taxonomy, and they appear in 45 areas. Most overlapping concepts are summarized by two partial-areas each. Only six overlapping concepts appear in three partial-areas simultaneously.

There are six areas with more than 10 overlapping neoplasm concepts in the partial-area taxonomy of the *Neoplasm* subhierarchy. The largest area contains 137 partial-areas, 463 concepts, and 27 overlapping concepts. These overlapping concepts are distributed over 18 partial-areas. The second-largest area contains 100 partial-areas, 321 concepts and 25 overlapping concepts. These overlapping concepts are distributed over

24 partial-areas. These two areas contain the two largest sets of overlapping concepts among all areas in the *Neoplasm* subhierarchy.

Figure 5.1 shows the disjoint partial-area taxonomy for the area with the six role types *Disease Excludes Abnormal Cell*, *Disease Excludes Finding*, *Disease Has Abnormal Cell*, *Disease Has Finding*, *Disease Has Normal Cell Origin*, and *Disease Has Normal Tissue Origin* that summarizes 98 concepts in 26 partial-areas. Of these 98, 20 concepts are overlapping concepts. The overlapping concepts appear in nine partial-areas. An excerpt of this disjoint partial-area taxonomy was also shown in Figure 2.6(c). In Figure 5.1, Level 2 had to be distributed over two rows, as there are 15 disjoint partial-areas at this level that do not fit into one row.

After the three-step QA study, the two domain expert reviewers agreed that 71 concepts (12.3% = 71/575) had errors with a moderate or severe error type. Among the 71 erroneous concepts, 36 concepts (16% = 36/225) were overlapping concepts in 16 areas, with 48 errors (1.33 errors per erroneous overlapping concept) and 35 (10% = 35/350) were non-overlapping concepts with 39 errors (1.11 errors per erroneous non-overlapping concept).

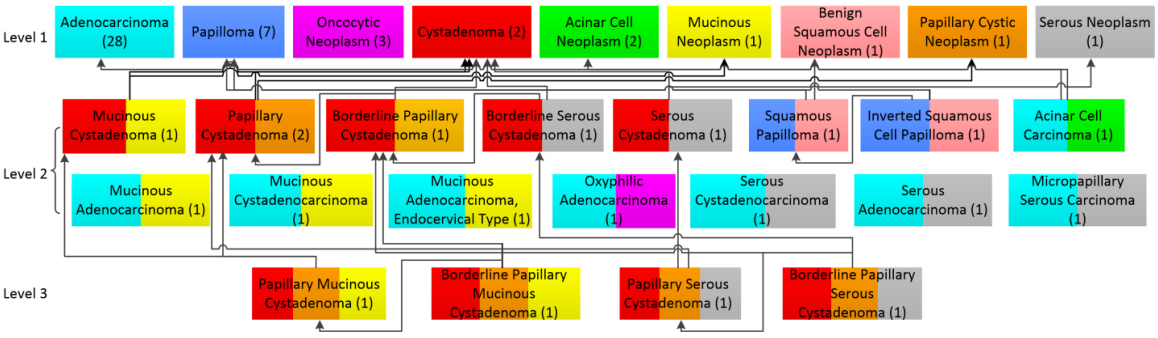

Figure 5.1 The disjoint partial-area taxonomy of the area with the six role types *Disease Excludes Abnormal Cell*, *Disease Excludes Finding*, *Disease Has Abnormal Cell*, *Disease Has Finding*, *Disease Has Normal Cell Origin*, and *Disease Has Normal Tissue Origin*. To reduce the density of the figure, the *child-of* links for the disjoint partial-areas at the second row of Level 2 are not shown.

Table 5.1 shows the area distribution of overlapping concepts and erroneous overlapping concepts. Table 5.2 is the contingency table for the *p*-value calculation between erroneous overlapping concepts and erroneous non-overlapping concepts. The two-tailed *p*-value of Fisher's exact test [118] was calculated to evaluate the statistical significance of the study. The *p*-value is 0.0377 ($p<0.05$), which means the study result has statistical significance. In other words, the overlapping concepts are likely to exhibit significantly more errors than non-overlapping concepts. Thus, Hypothesis 5.1 was supported by the results.

Table 5.1 The Distribution of Overlapping Concepts and Erroneous Overlapping Concepts

# of Overlapping Concepts in an Area	# of Areas	# of Areas with Errors	# of Erroneous Concepts
1	15	5	5
2	5	1	2
3	6	1	2
4	3	1	1
5	4	1	5
6	3	1	5
7	2	1	2
10	1	1	1
12	3	3	12
20	1	0	0
25	1	0	0
27	1	1	1
Total:	45	16	36

Table 5.2 The 2x2 Contingency Table for Erroneous Overlapping Neoplasm Concepts and Non-overlapping Neoplasm Concepts in NCIt

	# Erroneous Concepts	# Concepts w/o Errors
Overlapping concepts	36	189
Non-overlapping concepts	35	315

Of the 225 overlapping concepts, 195 came from disjoint partial-areas containing only one concept. The remaining 30 overlapping concepts came from disjoint partial-areas with at most four concepts. Altogether, only 18 overlapping concepts were not overlapping roots. Out of the 36 erroneous overlapping concepts, two concepts (11.1% = 2/18) were not overlapping roots and the other 34 concepts (16.4% = 34/207) were overlapping roots. In addition, only three concepts (10% =3/30) were from a disjoint partial-area with three concepts. The remaining 33 concepts (16.9% = 33/195) were from singleton disjoint partial-areas (disjoint partial-areas with only one concept).

There were two main error types of the overlapping concepts, 14 concepts with missing roles and 23 concepts with incorrect roles. The concept *Pancreatic Vipoma* has a missing role error and an incorrect role error at the same time. Table 5.3 illustrates five examples of errors found in overlapping concepts with suggested corrections and reasons. Besides 21 non-overlapping concepts with missing role errors and 12 non-overlapping concepts with incorrect role errors, the two domain experts also found incorrect parent, missing parent and incorrect neoplastic status for three non-overlapping concepts, which is illustrated in Table 5.4. “Neoplastic status” is a data property for neoplasm concepts in NCIt [119], with possible values “Benign,” “Malignant,” “Precancerous,” “Uncertain Malignant Potential,” and “Undetermined.” It defines a neoplastic growth as non-cancerous, cancerous, or of uncertain cancerous potential. The concept *Basophilic Adenocarcinoma* has both a missing parent error and a missing role error.

Figure 5.2 shows an interesting error case, in which the corrections of three erroneous overlapping concepts transform them into non-overlapping concepts *in another area*, by adding a new role *Disease Has Primary Anatomic Site* suggested by the two

domain experts. Figure 5.2(a) shows an excerpt of the disjoint partial-area taxonomy consisting of three disjoint partial-areas for the area with the three role types *Disease Excludes Primary Anatomic Site, Disease Has Abnormal Cell, Disease Has Associated Anatomic Site* and the area with an additional role type *Disease Has Primary Anatomic Site* (italic and underline). Notably, there is a *child-of* link between the two partial-areas *Recurrent Anterior Pituitary Gland Neoplasm (1)* and *Pituitary Gland Neoplasm (3),* because the concept *Recurrent Anterior Pituitary Gland Neoplasm* is a child concept of *Anterior Pituitary Gland Neoplasm* in the partial-area *Pituitary Gland Neoplasm (3)*.

Table 5.3 Five Examples of Errors in Overlapping Concepts Identified in the QA Study

Concept	Error Type	Correction	Reason
Childhood Central Nervous System Mature Teratoma	Incorrect role	Remove the role *Disease Has Abnormal Cell* with the target *Malignant Cell*	*Mature Teratoma* is a benign neoplasm
Occult Adenosquamous Lung Carcinoma	Incorrect role	Remove the role *Disease Excludes Finding* with the target *No Evidence of Radiologic Finding* or change the role to *Disease Has Finding* with the same target	According to the definition "*The primary tumor is undetectable radiographically or during bronchoscopy*"
Testicular Granulosa Cell Tumor	Missing role	Add the role *Disease Has Normal Cell Origin* with a more refined target *Granulosa Cell*	According to the definition "*It is characterized by the presence of granulosa-like cells*"
Pancreatic Vipoma	Missing role	Add the role *Disease May Have Associated Disease* with the target *Multiple Endocrine Neoplasia Type 1*	This concept has the role *Disease Mapped To Gene* with the target *MEN1 Gene*
Stage IVA Oral Cavity Cancer	Missing role	Add the role *Disease Is Stage* with the target *AJCC v7 Stage*	According to the definition, it is an AJCC 7th stage concept

Table 5.4 Three Other Error Types Identified in Non-overlapping Concepts of the QA Study

Concept	Error Type	Correction	Reason
Basophilic Adenocarcinoma	Missing parent	Add an IS-A link directed to *Anterior Pituitary Gland Neoplasm*	According to the definition "*A malignant epithelial neoplasm of the anterior pituitary gland*"
Papillary Hidradenoma	Incorrect parent	Replace the parent *Benign Sweat Gland Neoplasm* with *Hidradenoma*	*Hidradenoma* is more relevant
Gallbladder Goblet Cell Carcinoid	Incorrect neoplastic status	Change the value "Undetermined" to "Malignant"	According to the definition "An invasive mixed adenoneuroendocrine carcinoma of the gallbladder"

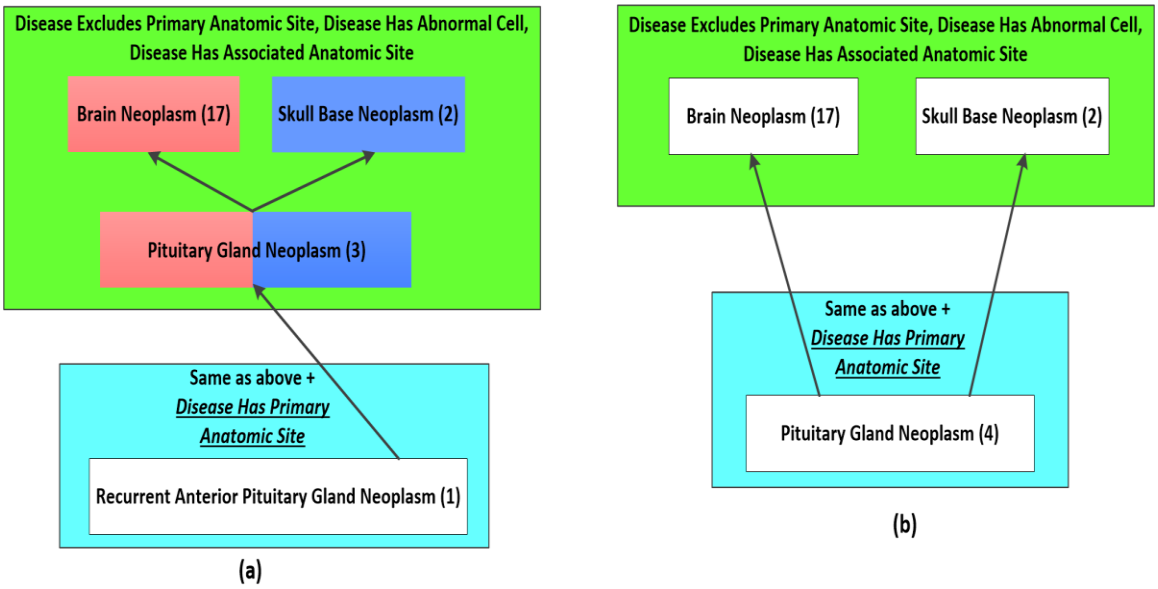

Figure 5.2 Simplification of the complexity of the disjoint partial-area taxonomy due to correction of overlapping concepts: **(a)** Excerpt from disjoint partial-area taxonomy before correction of three erroneous overlapping concepts in the partial-area *Pituitary Gland Neoplasm (3)* with the error "missing the role *Disease Has Primary Anatomic Site*"; **(b)** after correction by adding the missing role (italic and underline) to the three erroneous overlapping concepts. The two partial-areas in Figure 5.2(a) *Pituitary Gland Neoplasm (3)* and *Recurrent Anterior Pituitary Gland Neoplasm (1)* are merged together to become a new partial-area *Pituitary Gland Neoplasm (4)*, because *Recurrent Anterior Pituitary Gland Neoplasm (1)* is *child-of Pituitary Gland Neoplasm (3)*. All three partial-areas are not colored, since they do not contain overlapping concepts.

The two concepts *Pituitary Gland Neoplasm* and its child *Posterior Pituitary Gland Neoplasm* were two overlapping concepts in the audited sample (Figure 5.2(a)). The domain experts reported that both concepts missed the role *Disease Has Primary Anatomic Site* with the target *Pituitary Gland*. Hence, after correction by adding the missing role in NCIt, these two concepts (in fact three concepts, including the other child *Anterior Pituitary Gland Neoplasm* due to inheritance) in the newly derived corresponding disjoint partial-area taxonomy appear in the bottom area with four role types in Figure 5.2(b). The name of the added role in NCIt is again italicized. Furthermore, the two partial-areas in Figure 5.2(a) *Pituitary Gland Neoplasm (3)* and *Recurrent Anterior Pituitary Gland Neoplasm (1)* are merged into a new partial-area *Pituitary Gland Neoplasm (4)* in Figure 5.2(b). The three concepts of the partial-area *Pituitary Gland Neoplasm (3*) in Figure 5.2(a) are not overlapping concepts anymore in the new area in Figure 5.2(b). Specifically, *Pituitary Gland Neoplasm* became an area root in the new area. That is, after the correction these three concepts are not "complex" anymore, because they are not overlapping concepts, since they are in a separate area with one root.

Figure 5.2 demonstrates that the corrections of erroneous overlapping concepts may transform overlapping concepts into non-overlapping concepts. Thus, the complexity of the disjoint partial-area taxonomy is reduced. For example, in Figure 5.2(b) this is expressed by the elimination of one disjoint partial-area (*Pituitary Gland Neoplasm*) in the disjoint partial-area taxonomy, leading to a simpler summary. Hence, correcting erroneous overlapping concepts may reduce the complexity of the ontology. The simplification in Figure 5.2 is expressed by eliminating the "striped" node of Figure

5.2(a) when generating Figure 5.2(b). This reduces the total number of boxes and makes it unnecessary to color any of the partial-area nodes. This phenomenon shown in Figure 5.2 is a novel, important, and useful one during quality assurance of "overlapping concepts."

To conclude, in this study, the partial-area taxonomy and the disjoint partial-area taxonomy for the *Neoplasm* subhierarchy of the *Disease, Disorder or Finding* subhierarchy of NCIt were derived. A three-step manual QA study was performed on a sample of 575 neoplasm concepts consisting of overlapping concepts and non-overlapping concepts selected from the *Neoplasm* disjoint partial-area taxonomy. The results of the QA study show that overlapping concepts have a statistically significantly higher error rate than non-overlapping concepts (16% vs. 10%), making the *Neoplasm* subhierarchy in NCIt became the fourth ontology in its BioPortal family, for which the methodology of reviewing overlapping concepts was successfully demonstrated.

5.1.2 Quality Assurance of NCIt Gene Hierarchy by Role-subset Partial-area Sub-taxonomy

Overlapping concepts existing in partial-area taxonomies of ontologies are more likely to have errors than control concepts. This effective QA technique has been successfully demonstrated on the four ontologies of the same BioPortal ontology family, i.e., the *Specimen* hierarchy and the *Bleeding* subhierarchy of SNOMED CT, Uberon, and the *Neoplasm* subhierarchy of NCIt. The reason is as follows. All the concepts of a partial-area share the same semantics. A concept that simultaneously belongs to multiple partial-areas has a compound semantics combining the "simple" semantics of each of those partial-areas. Such a concept is thus more complex than a concept that belongs only to one partial-area.

However, the number of such complex concepts of compound semantics in the NCIt *Gene* hierarchy introduced in Section 2.1.3 is small (96). Hence, in spite of the fact that such concepts have been shown to have a statistically significantly higher error rate than control concepts, reviewing all such gene concepts will have a very limited impact on the quality of the *Gene* hierarchy. A new innovative QA methodology is needed to discover additional complex concepts that display similar properties as the concepts with compound semantics, even though these additional concepts have simple semantics in the *Gene* hierarchy as seen through the prism of a partial-area taxonomy. This section demonstrated that the *role-subset partial-area sub-taxonomy* for the *Gene* hierarchy contains more complex concepts than the original partial-area taxonomy and such additional complex concepts were statistically significantly more likely to have errors than "simple" concepts in the role-subset partial-area sub-taxonomy.

5.1.2.1 Methods. In the NCIt *Gene* hierarchy, overlapping concepts are manifested as genes that are simultaneously related to multiple processes. Therefore, these concepts have the compound semantics of relating to different processes. For example, the *PARP2 Gene* is only involved in the process of *DNA Repair*, while the *RAD9A Gene* plays roles in three processes, *Cell Cycle*, *Hydrolysis*, and *DNA Repair*. The compound semantics makes overlapping concepts more difficult to model. Hence, in this study, the disjoint partial-area taxonomy for the *Gene* hierarchy was derived and the following hypothesis was investigated.

Hypothesis 5.2: Overlapping concepts are more likely to have errors than non-overlapping concepts in the disjoint partial-area taxonomy derived from the *Gene* hierarchy of NCIt.

Validity of Hypothesis 5.2 implies the following QA methodology: the disjoint partial-area taxonomy can be utilized to identify overlapping concepts. These concepts are likely to have a high error yield, measured by the ratio of the number of discovered errors to the number of reviewed concepts, compared to non-overlapping concepts. This auditing methodology is a complement to the methodology described by Cohen et al. [83].

To test Hypothesis 5.2, a QA study was conducted on a random sample consisting of 50 overlapping concepts in the disjoint partial-area taxonomy of the *Gene* hierarchy as the study sample and 50 non-overlapping concepts from the same partial-areas as the study sample, as the control sample. The study sample and control sample were combined into one list in randomized order. The randomized list was presented to the domain expert Dr. Hua Min for review. The domain expert reviewed each concept using the NCI term browser and focused on commission errors (of wrong features) and omission errors of hierarchical relationships and roles (missing features). She generated a report in which she marked which concepts have what kinds of errors and she suggested corrections for these errors. Her report was reviewed by the NCIt team. Based on the errors confirmed by the NCIt team, the two-tailed *p*-value of Fisher's exact test [118] was calculated to evaluate the statistical significance of the difference between the error rate of overlapping concepts and that of non-overlapping concepts.

However, as mentioned before, the number of overlapping concepts is low (1%). The practical impact of auditing all 96 overlapping concepts would be small, even if the error rate turns out to be high. The reason that the number of overlapping concepts turned out to be low is that there are relatively many role types (16). The impact of each additional role type R is that it has the potential of dividing an area into two smaller areas,

separating those concepts with R from those concepts without R. The probability of having several roots and overlapping concepts in a small area is reduced, compared to a large area.

It should be noted that all 96 overlapping concepts are in the same area, containing only the role "*Gene Plays Role In Process*." That is, all the overlapping concepts in the *Gene* hierarchy are deriving their extra complexity from belonging to two or three partial-areas referring to two or three different kinds of processes. Thus, a "role-reduced *Gene* hierarchy" was derived by eliminating all roles except for "*Gene Plays Role In Process*." Then a new disjoint partial-area taxonomy for the reduced *Gene* hierarchy was constructed. The Ontology Abstraction Framework (OAF) software system [67] could easily derive such a *role-subset partial-area sub-taxonomy*. This new disjoint partial-area taxonomy with only two areas, denoted T1 ("Taxonomy 1"), was found to contain 376 overlapping concepts in the area {*Gene Plays Role In Process*}. No overlapping concepts exist in the other area. Figure 5.3 shows the flowchart of obtaining overlapping concepts for the original *Gene* hierarchy (Figure 5.3(a)) and for the role-reduced *Gene* hierarchy (Figure 5.3(b)).

This increase of overlapping concepts (from 96 to 376) is expected and desired. As a simplified example of the observed effect, consider the case of another role-reduced

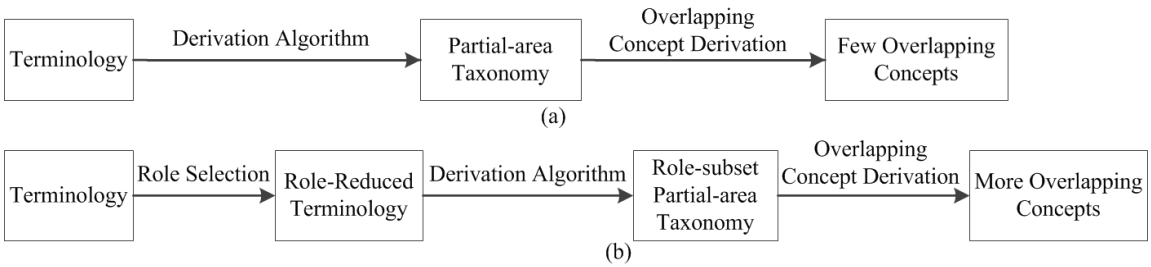

Figure 5.3 Flowchart for finding overlapping concepts **(a)** for the original *Gene* hierarchy **(b)** for the role-reduced *Gene* hierarchy.

Gene hierarchy with exactly two *roles Gene Plays Role In Process* ("*Process*" for short) and *Gene Associated With Disease* ("*Disease*" for short). Using the OAF software with this assumption, the resulting disjoint partial-area taxonomy, denoted T2 ("Taxonomy 2"), has four areas, and all the overlapping concepts are again concentrated in the area {*Process*}. However, the number of overlapping concepts in this area is now only 298. Why did the addition of the second role decrease the number of overlapping concepts in the area of concepts with only the "*Process*" role (Figure 5.4)?

The total number of concepts in the {*Process*} area in T1 is 8,775, but the number of concepts in this area in T2 is only 7,571. The reason is that 8775–7571=1204 of the concepts in the {*Process*} area in T1 have both roles in T2. Hence, adding an extra role is increasing the number of areas and some concepts of the previous {Process} area appear now in a new area {*Process, Disease*} with both roles. Areas are always, by definition, disjoint. Hence, some of the roots and thus the corresponding partial-areas, (e.g., *NAT2 Gene (6)* of the {*Process*} area in T1) are now in the area {*Process, Disease*} in T2. As a result, some overlapping concepts in T1 belong to the {*Process*} area in T1, but appear in the {*Process, Disease*} area in T2 where they are not overlapping concepts, since the two partial-areas that contained them in T1 are in the {*Process*} area in T2. This somewhat complex reasoning chain is elucidated by Figure 5.4.

The following example, shown in Figure 5.4, demonstrates that adding a role to a hierarchy decreases the number of overlapping concepts belonging to both partial-areas *Ligand Binding Protein Gene* and *Phosphotransferase Gene* in the {*Process*} area as a result of splitting the area into two smaller areas. The number of overlapping concepts belonging to both partial-areas in T1 decreased by 40 (from 126 to 86) in T2. *Ligand*

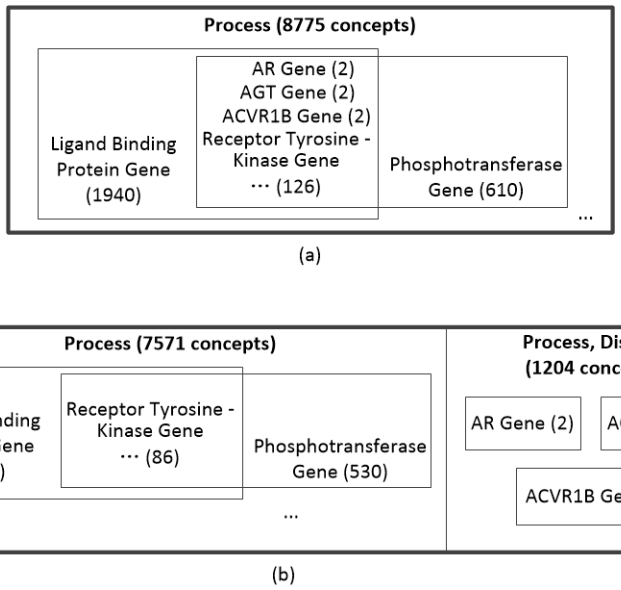

Figure 5.4 (a) Two overlapping partial-areas in T1. **(b)** An excerpt of T2 shows the effect of the addition of one role.

Binding Protein Gene (1940) and *Phosphotransferase Gene (610)* are two partial-areas in T1 and they have 126 common concepts (overlapping concepts), e.g., *AR Gene* and its corresponding wild-type allele *AR wt Allele* represented by "AR Gene (2)" in Figure 5.4(a). The addition of the role *Disease* transforms 40 (=126–86) overlapping concepts in T1 into non-overlapping concepts in T2, since these 40 concepts now have both roles *Process* and *Disease*, and thus they are in the new area {*Process, Disease*} in T2 in Figure 5.4(b). For example, the two former overlapping concepts *AR Gene* and *AR wt Allele* in T1 are now in their own partial-area *AR Gene (2)* in the area {*Process, Disease*} in T2, thus they are not overlapping in T2.

The concepts that become overlapping in a partial-area taxonomy – by reducing the number of roles that are considered – are called *extra overlapping concepts*. The interesting questions with regard to the 376–96=280 extra overlapping concepts are, (1) are they as complex as the 96 original overlapping concepts? (2) do they have a

comparable higher error rate? It is important to note that when auditing those concepts all their roles in the original *Gene* hierarchy are taken into account. The "role-reduced hierarchies" above were only used to derive T1 and T2, not to permanently change the *Gene* hierarchy.

The complexity of concepts in the *Gene* hierarchy is caused by their belonging to multiple partial-areas reflecting their participation in multiple different biological processes and not by their roles. This kind of complexity was already evident in T1. The addition of extra roles definitely does not decrease the complexity of concepts. To the contrary, in a recent paper investigating the *Neoplasm* subhierarchy of NCIt [14], concepts with more roles have been shown to have higher error rates due to their higher complexity measured in that case by the number of role types. Hence, the extra overlapping concepts are expected to have at least similar error rates as the original overlapping concepts.

In order to find the answer to the above two questions, the role-subset partial-area sub-taxonomy T3 [120] was derived using the subset of roles {*Gene Has Abnormality*, *Gene Involved In Pathogenesis Of Disease*, *Gene Is Biomarker Type*, *Gene Plays Role In Process*} in the *Gene* hierarchy. The taxonomy T3 with its four roles has 12 areas and 874 partial-areas. A QA study on a second random sample was conducted to test Hypothesis 5.3.

Hypothesis 5.3: Extra overlapping concepts in a role-subset partial-area sub-taxonomy derived from the *Gene* hierarchy of NCIt are more likely to have errors than non-overlapping concepts, when reviewed in the *Gene* hierarchy itself.

The reason why this subset of roles was chosen is that in T3, for this subset of roles, there are 340 extra overlapping concepts that are all in the area {*Process*}. This number of extra overlapping concepts is the closest to the 376 overlapping concepts in T1 for all role-subsets with four role types, which is an option with a balance between the number of overlapping concepts and the number of role types. The study sample for this second study was composed of 50 concepts randomly selected from the 340–96=244 extra overlapping concepts in T3, excluding the 96 overlapping concepts from the original *Gene* partial-area taxonomy.

The control sample consists of 50 non-overlapping concepts randomly selected from the same partial-areas in T3 as the study concepts. After the study sample and the control sample were randomly mixed, this random list was reviewed by the same domain expert Dr. Hua Min. The domain expert's error report was reviewed by the NCIt team who confirmed some of the errors. Based on the errors upheld by the NCIt team for the second study, the two-tailed *p*-value of Fisher's exact test [118] was calculated to evaluate the statistical significance for the difference between the error rate of extra overlapping concepts and the error rate of non-overlapping concepts in the partial-area taxonomy for the *Gene* hierarchy.

5.1.2.2 Results. The partial-area taxonomy of the NCIt *Gene* hierarchy, derived for the September 2016 release, is composed of 5,318 partial-areas in 140 areas, with 96 overlapping concepts. All these overlapping concepts are in the area {*Process*}, which summarizes 3,232 concepts (33.88% = 3,232/9,540) by 417 partial-areas (7.84% =417/5,318). Two overlapping concepts are simultaneously in three partial-areas and the

other 94 overlapping concepts are simultaneously in two partial-areas. Figure 5.5 shows an excerpt of the disjoint partial-area taxonomy for the area {*Gene Plays Role In Process*}, which includes 75 overlapping concepts (78% = 75/96).

After auditing the 100 concepts of the first sample using the NCIt term browser, the domain expert found 76 concepts having errors, distributed over 32 (64%) non-overlapping concepts and 44 (88%) overlapping concepts. There were two kinds of errors, redundant *Process* roles (i.e., redundant role targets) and missing roles. One concept may have both kinds of errors.

After reviewing the errors reported by the domain expert, the NCIt team confirmed 65 erroneous concepts, including 23 non-overlapping concepts (46%) and 42 overlapping concepts (84%). The two-tailed *p*-value for the errors confirmed by the NCIt team, using Fisher's exact test [118], is p=0.0001, meaning the error rate of overlapping concepts is statistically significantly higher than that of non-overlapping concepts. Thus, Hypothesis 5.2 was supported by the confirmed errors. Table 5.5 shows four examples of errors that were confirmed by the NCIt team.

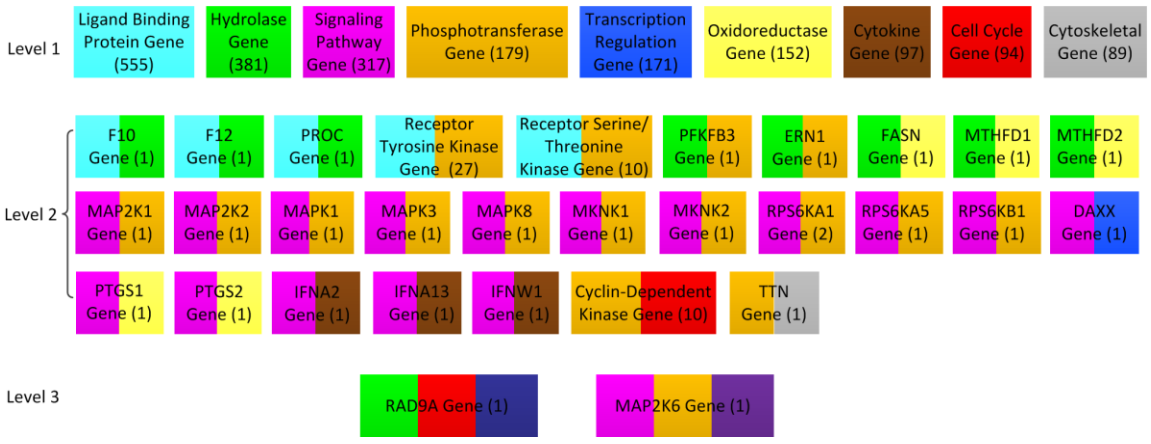

Figure 5.5 An excerpt of the disjoint partial-area taxonomy for the {*Gene Plays Role In Process*} area, which only shows the nine largest partial-areas and all disjoint partial-areas derived from these nine partial-areas. *Child-of* links are omitted for readability, since they are implied by the color coding.

The NCIt team confirmed all 53 concepts with redundant *Process* role errors, including 37 overlapping concepts and 16 non-overlapping concepts, as well as 14 concepts with missing roles. Table 5.6 summarizes the distribution of confirmed erroneous concepts with missing role errors by role type. All missing *Gene Has Physical Location* role errors were not accepted, because these errors were reported for wild-type allele concepts and the NCIt team does not require the addition of such a role for this case.

Table 5.5 Four Examples of Confirmed Errors by the NCIt Team

Concept	Overlapping Concept	Error	Suggested Correction
Cyclin-Dependent Kinase Gene	Y	Redundant target: The target of *Process* role *Phosphorylation Process* is the ancestor of another target *Serine/Threonine Phosphorylation*	Remove the *Process* role with the target *Phosphorylation Process*
RYK Gene	Y	Missing the *Process* role with the target *Signal Transduction* according to its definition	Add the role
RPS6KA1 wt Allele	Y	Missing the role *Gene In Chromosomal Location* with the target *1p36.11*	Add the role
CCNA1 wt Allele	N	Missing the role *Gene Found In Organism* with the target *Human*	Add the role

Table 5.6 The Distribution of Confirmed Erroneous Concepts with Missing Role Errors by Role Type

Role Type	# of Confirmed Erroneous Concepts	# of Confirmed Overlapping Concepts	# of Confirmed Non-overlapping Concepts
Gene Plays Role In Process	8	6	2
Gene In Chromosomal Location	3	1	2
Gene Found In Organism	3	0	3

For the concepts with missing *Gene Found In Organism* errors, the NCIt team accepted only three error reports, because the NCIt editor thought that such a role should be instantiated at a more general concept not at a specific gene, since this role is suitable for all non-human genes. That is, these errors reported by the domain expert are indeed errors, but should be corrected at other concepts.

For the missing *Gene In Chromosomal Location* errors, the NCIt team only accepted errors with chromosomal band positions that already exist in NCIt (for example, 2q35), since they do not wish to create new specific chromosomal band positions, unless a user requests them. There are two reasons for this. First, the NCIt team is not notified when these values change, as it happens when experimental evidence refines the locations. Secondly, they are considering modeling such information differently in the future. Therefore, even though the suggested value is correct, if it does not currently exist as a concept in NCIt, they will not add this role.

For the missing *Gene Plays Role In Process* errors, as a rule, they only model gene concepts with this role but do not model wild-type allele concepts with it. More specifically, they add the role only to such gene concepts for which the associated Gene Ontology annotation evidence codes are either experimental or based on authors' statements.

Based on the analysis of errors not accepted by the NCIt team, it is observed that the problems reported by the domain expert are indeed errors, but due to various internal NCIt rules that were not known to the external auditor they were not corrected in NCIt.

For the second sample of 100 concepts, 78 concepts (78%) were found to have errors consisting of 45 extra overlapping concepts (90% of 50) and 33 non-overlapping

concepts (66% of 50) by the domain expert. The kinds of errors for this sample are similar to those of the first sample, namely, redundant targets of *Process* roles and missing roles. The NCIt team confirmed 26 erroneous concepts, including 22 extra overlapping concepts (44%) and four non-overlapping concepts (8%). The *p*-value for the second QA study, based on the NCIt team's confirmed errors, is less than 0.0001, meaning the error rate of extra overlapping concepts is significantly higher than that of non-overlapping concepts. These results confirmed Hypothesis 5.3.

In conclusion, a QA study of complex concepts discovered with the help of the partial-area taxonomy of the *Gene* hierarchy in NCIt was conducted. The results show that complex concepts are more likely to have errors than simple concepts (84% vs. 46%). To extend the practical impact of "complex" concepts on the QA process of the *Gene* hierarchy, a new QA methodology was introduced by deriving the partial-area **sub**-taxonomy using a subset of roles defined for the *Gene* hierarchy. In other words, the partial-area taxonomy for the role-reduced *Gene* hierarchy was derived. This new methodology identified an additional set of complex concepts that also exhibited a statistically significantly higher error rate. The error rate for the additional complex gene concepts (44%) was about five times as large as the error rate for control concepts (8%). Thus, this study is the fifth study that confirmed the usefulness of QA based on partial-area taxonomies, with a focus on complex concepts, and constitutes an important building block towards the goal of showing the effectiveness of family-based QA for biomedical ontologies.

5.1.3 Quality Assurance of Complex Infectious Disease Concepts in SNOMED CT

This section presents an overlapping concept-based QA study on the sixth ontology, the *Infectious Disease* subhierarchy of SNOMED CT, which is in the same BioPortal ontology family as the five ontologies, the *Specimen* hierarchy and the *Bleeding* subhierarchy of SNOMED CT, the Uberon ontology, and the two ontologies in the previous two sections, to which the overlapping complex concepts-based QA methodology has been successfully applied.

During the year 2015, editors of SNOMED International conducted a project of remodeling the *Infectious Disease* subhierarchy of SNOMED CT. Details of this work were published by Ochs et al. [121]. Due to scheduling difficulties, the project was not completed. In the process they remodeled the stated concepts, and by using a classifier [122] the inferred view of the subhierarchy was generated.

The QA study on the *Infectious Disease* subhierarchy took advantage of the remodeling project initiated by SNOMED International. The study concentrated on all the inferred changes made to the *Infectious Disease* subhierarchy between the January 2015 release and the July 2015 release. During this period, 4,308 concepts were changed. Any time a concept was changed during such a remodeling process it is apparent that this concept was previously erroneous. A similar idea was extensively used by Ceusters et al. [123] and by Zhang et al. [124]. This approach is substantially different from the studies in the previous sections, in which several domain experts reviewed a sample of overlapping concepts and a sample of non-overlapping concepts for errors. The following hypothesis was investigated in this study.

Hypothesis 5.4: Overlapping concepts are more likely to have errors than non-overlapping concepts for the SNOMED CT *Infectious Disease* subhierarchy.

In evaluating Hypothesis 5.4, only "severe" and "moderate" errors were considered, just as was done for the NCIt *Neoplasm* subhierarchy. Since there was no domain expert involved in determining what is considered a severe or moderate error, the judgment of what makes an error "severe" or "moderate" had to be arrived at indirectly. Previous feedback of ontology curators has indicated that commission errors are considered more severe than omission errors, because commission errors indicate that some part of the modeling of a concept is outright wrong. Omissions are sometimes done on purpose by ontology curators, because there is no use case for the omitted information. Such errors are generally considered non-critical.

For this study, a sample was generated containing all the overlapping *Infectious Disease* concepts and a random control sample consisting of an equal number of non-overlapping concepts from the *Infectious Disease* subhierarchy. To assure a fair comparison, the control concepts were randomly taken from the same areas as the overlapping concepts. Since concepts in small partial-areas are prone to have more errors, the control population excluded such concepts as a confounding factor. The two-tailed *p*-value of Fisher's exact test [118] was calculated to evaluate the statistical significance of the different error rates for overlapping *Infectious Disease* concepts and for non-overlapping *Infectious Disease* concepts.

The SNOMED CT *Infectious Disease* subhierarchy contained 6099 concepts in the January 2015 release. Its partial-area taxonomy contains 80 areas and 1305 partial-areas with 196 overlapping concepts distributed over eight areas. The area with the most

overlapping concepts has three role types *Associated morphology*, *Finding site*, and *Pathological process* with 665 concepts among which there are 83 overlapping concepts.

The overlapping concepts were found by the Ontology Abstraction Framework (OAF) software tool [67]. The concepts that underwent a change between two releases were found by the SNOMED CT Visual Semantic Delta tool [125]. The concepts with commission errors were obtained from the sample of 196 overlapping concepts and from the control group of 196 randomly chosen non-overlapping concepts.

Table 5.7 is the contingency table for the *p*-value calculation distinguishing between erroneous overlapping and erroneous non-overlapping concepts. Erroneous concepts that have commission errors, such as wrong parent, wrong role type, or wrong role target were counted. A sample of commission errors of different kinds appears in Table 5.8. The two-tailed *p*-value of Fisher's exact test [118] was calculated to evaluate the statistical significance of the study. The *p*-value is 0.0067 ($p<0.05$), which means the study result has statistical significance. Thus, Hypothesis 5.4 was supported by the results.

To summarize, the hypothesis that overlapping concepts are more likely to have errors than non-overlapping concepts was supported for this sixth ontology, the *Infectious Disease* subhierarchy of SNOMED CT, in addition to the other five ontologies in the same family, the *Specimen* hierarchy and the *Bleeding* subhierarchy of SNOMED CT, Uberon, the *Neoplasm* subhierarchy and the *Gene* hierarchy of NCIt. Thus, the "six out of

Table 5.7 The 2x2 Contingency Table for Erroneous Overlapping versus Non-overlapping *Infectious Disease* Concepts in SNOMED CT

	# Erroneous Concepts	**# Concepts w/o Errors**	**% Errors**
Overlapping concepts	76	120	38.8
Non-overlapping concepts	50	146	25.5

Table 5.8 Different Kinds of Commission Errors for Overlapping versus Non-overlapping Concepts

	Overlapping concepts	**Non-overlapping concepts**
Wrong Parent	*Tuberculous enteritis*	*Tuberculous ascites*
Wrong Parent	*Oculoglandular tularemia*	*Mumps nephritis*
Wrong role type	*Tuberculous peritonitis*	*Anal candidiasis*
Wrong role type	*Bullous staphylococcal impetigo*	*Bacterial peritonitis*
Wrong target	*Beta lactam resistant bacterial infection*	*Infection by Diplodinium*
Wrong target	*Superficial foreign body of anus without major open wound but with infection*	*Infection by Theileria parva*

six" requirement for this family is fulfilled.

Among the six ontologies, there are two from NCIt and three from SNOMED CT. However, there are differences between them. The NCIt *Gene* subhierarchy is different from the *Neoplasm* subhierarchy, since all the genes are modeled as leaves or as parents of leaves in cases where they have alleles. In contrast, diseases can appear anywhere in the *Neoplasm* subhierarchy. Regarding SNOMED CT, *Specimen* is a small subhierarchy, while *Clinical finding* is the largest subhierarchy of SNOMED CT. It is two magnitudes larger than *Specimen.* Because it is so large, two subhierarchies of it were reviewed, the small *Bleeding* subhierarchy and the medium-sized *Infectious Disease* subhierarchy, to assess the validity of the overlapping concepts-based QA technique for ontologies of different sizes.

The implication of confirming the efficacy of the above uniform QA methodology for six ontologies is that for at least half of the other ontologies in the substantial BioPortal family studied in this paper the error rate for overlapping concepts will be significantly higher than the error rate for non-overlapping concepts [8]. Hence, by

concentrating QA efforts on overlapping concepts in the ontologies of that family, a higher QA yield is expected in terms of the number of concepts identified as erroneous for a given number of reviewed concepts, exercising the best possible use of scarce human resources. Thus, when embarking on quality assurance for members of this family under resource constraints, overlapping concepts should be audited first. At the very least, all overlapping concepts should be audited for every member of this family of ontologies.

Besides higher yield, another advantage of the family-based QA approach [8] is that it is supported by the OAF software tool [67] that finds the overlapping concepts for each ontology of the family, rather than having to develop algorithms separately for each member of the family. Hence, this methodology is semi-automatic, because the overlapping concepts are found automatically by the OAF software and the manual review is only performed for those concepts. Finding a method that prioritizes among the overlapping concepts would be beneficial for ontologies with many such concepts.

Hence, the results in this study suggest that the methodology of reviewing overlapping concepts is an effective QA methodology for ontologies of one family in BioPortal, as this methodology has been demonstrated successfully for six out of six ontologies in the chosen BioPortal family. This means that the overlapping concept methodology can be applied to the whole BioPortal family of 76 similar ontologies and is likely to be successful for at least half of the members of this family.

5.1.4 Quality Assurance of Complex Concepts in NCIt Biological Process Hierarchy

This section reports a QA study [14] on another kind of complex concepts in the NCIt *Biological Process* (*BP*) hierarchy from the perspective of a relatively straightforward

characterization of lateral complexity of concepts, namely, their overall numbers of role types.

5.1.4.1 Methods. This section intends to explore whether more laterally complex concepts, where "complexity" is defined in terms of the number of exhibited role types, are more prone to errors than less complex concepts. Roles play a central part in logically modeling concepts, and thus it is natural to focus on them as a measure of complexity. As an example, consider the concept *G1 to S Transition Process* in NCIt's *Biological Process* hierarchy, with the five role types *Location*, *Initiator Chemical or Drug*, *Initiator BP*, *Resulting BP*, and *Part of Process* (the full names were given in Table 2.2). It is one of the four concepts in the bottom area of the area taxonomy in Figure 5.6. This concept elaborates five different aspects of a biological process and can be considered more complex than *Neuronal Transmission* with only three of those aspects. *Neuronal Transmission*, in turn, is more complex than its parent *Intercellular Communication Process*, which has only the one role type *Initiator Chemical or Drug*.

In this study, the area taxonomy of the NCIt *Biological Process* hierarchy was divided into two halves, based on a level that forms the boundary between more laterally complex and less complex concepts. Specifically, let *r* be the maximum number of different role types exhibited by any actual concept in a given hierarchy. (The value *r* is obviously less than or equal to the number of predefined kinds of role types for the hierarchy.) For this study, *r* serves as a lateral complexity measure of the concepts. Following the principle of **Divide-and-Conquer** [126], the straightforward application of this principle, according to this complexity measure, is to divide the range into two equal-sized parts. That is, let $h = \left\lfloor \frac{r}{2} \right\rfloor$, i.e., *h* is half of *r*, rounded down to the nearest integer.

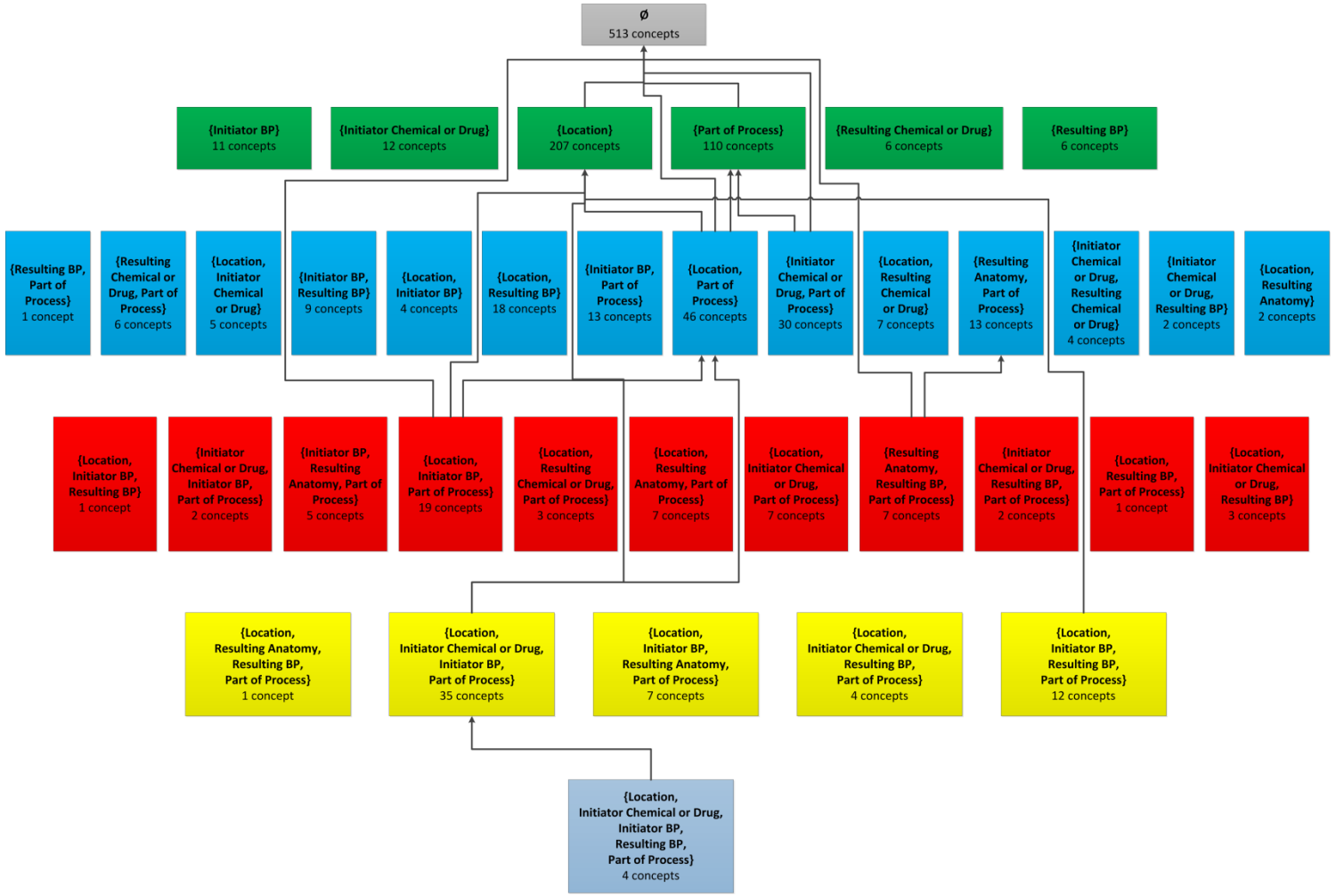

Figure 5.6 Complete area taxonomy of the *Biological Process* hierarchy. Most *child-of*s have been omitted to avoid overload. Note how the importance of the role *Location* is reflected in the area taxonomy. The area {*Location*} has 207 concepts, and *Location* appears in 20 of 37 area names.

The difference between this study and the study in Section 5.2.2 is that this study targeted only concepts with roles and excluded the concepts in the top area (Level 0), while the study in Section 5.2.2 was about the concepts in the top area.

Concepts residing on Levels 1, 2, …, h (the lower-half levels) of the taxonomy were taken to be simpler concepts. The concepts on Levels $h + 1$, $h + 2$, …, r (the upper-half levels) were taken to be more complex. In the case of the *Biological Process* hierarchy, $r = 5$ (though there are seven predefined role types), and the partition of the area taxonomy of Figure 5.6 was between Levels 1 and 2 versus Levels 3, 4, and 5. For the *Disease, Disorder or Finding* hierarchy with 20 levels in its area taxonomy, the levels would be divided into Levels 1–9 and Levels 10–19. It was postulated that concepts in the upper-half levels would have on average a higher number of modeling errors than concepts in the lower-half levels.

In this study, various concepts in the *BP* hierarchy were subjected to a thorough QA analysis by the subject-domain expert Dr. Hua Min. Since the number of concepts in the upper-half levels of an area taxonomy is typically much smaller than that in the lower-half levels (e.g., in the taxonomy of Figure 5.6, only four concepts are on Level 5), all concepts in the upper-half levels were analyzed. As a second group, a random sample comprising the same number of concepts from the lower-half levels underwent a QA analysis. The random sample was chosen in such a way that there was proportional representation according to the number of concepts on each of the lower-half levels.

The domain expert was looking for all types of errors, including errors of omission (e.g., an omitted role from the predefined set for the hierarchy) and commission (e.g., an incorrect target concept for a defined role). After the initial phase of QA, a

review phase was carried out by a curator of NCIt, who was asked to re-analyze and verify the discovered errors.

The following hypothesis is central to this study:

Hypothesis 5.5: For a given hierarchy, concepts on the lower-half levels (1, 2, …,$\left\lfloor\frac{r}{2}\right\rfloor$) of its area taxonomy of *r* levels have a lower average number of errors than concepts on the upper-half levels ($\left\lfloor\frac{r}{2}\right\rfloor + 1$, …, *r*).

The statistical significance for the error rates between the lower-half levels and the upper-half levels was evaluated using a two-tailed Fisher's exact test [118]. The implication of verifying Hypothesis 5.5 is that the set of concepts with more than $\left\lfloor\frac{r}{2}\right\rfloor$ kinds of roles denote a characterization of concepts where more errors are expected. Concentrating a QA analysis on such a set of concepts is expected to yield more corrections than a QA analysis of a random set of the same number of concepts with at most $\left\lfloor\frac{r}{2}\right\rfloor$ kinds of roles. The Ontology Abstraction Framework (OAF) tool [67] can automatically extract concepts at the levels where higher error rates are expected and its Neighborhood Auditing Tool (NAT) [127] can support the review of the auditor.

5.1.4.2 Results. The QA analysis was carried out on the NCIt's *Biological Process* hierarchy, consisting of 1,145 concepts (15.02d release). For this hierarchy, $r = 5$, i.e., no concept exhibits more than five role types, though there are seven possible predefined role types. Level 0 (concepts with no roles) contains 513 concepts, so the pool of concepts for the study (i.e., those with roles) is 632 concepts (55.2% of the overall hierarchy). Out of these 632 concepts, 393 concepts (62.2%) are defined in terms of the role *Location*. The analysis was done on the OWL version of NCIt by the domain expert

using the NCI Term Browser [128]. The domain expert reviewed for each concept all hierarchical relationships as well as all roles. Verification of the results was done by a curator of the NCIt.

On the upper-half Levels 3, 4, and 5 of *Biological Process*, there are 57, 59, and 4 concepts, respectively, totaling 120 concepts (10.5% of the hierarchy). Correspondingly, 120 concepts were randomly selected from the lower-half Levels 1 and 2. These 120 concepts were divided between the Levels 1 and 2, approximately in the same ratio as their numbers of concepts.

Table 5.9 shows the primary results of the initial phase of the QA analysis, with the numbers of erroneous concepts given for each of the levels of the area taxonomy. For example, on Level 1, seven erroneous concepts were discovered among the 80 concepts that were analyzed, for an error rate of 8.75%. On Level 3, nine of the 57 concepts were deemed erroneous, for a 15.79% error rate. It should be noted that the error rates are in the single digits for the lower-half levels and in the double digits for the upper-half levels. In total, there were 43 errors for the 40 erroneous concepts, among which 22 errors (for 20 concepts) were missing-role errors. Another prominent error type involved conflicting semantics between IS-A hierarchical relationships and *Part of Process* roles. In particular, the reviewer deemed that these two relationships should not target the same concept, directly or transitively, from a single source. For example, it was found that the concept *Anaphase* has the role *Part of Process* with the target *Cell Cycle Process*, while at the same time *Cell Cycle Process* is the grandparent of *Anaphase* (i.e., *Anaphase* is transitively connected to *Cell Cycle Process* via IS-A relationships). This was considered a conflict, and *Anaphase* was marked as erroneous. Twenty occurrences of such modeling

were found with respect to 20 different concepts.

Additionally, the concept *Negative Regulation of G0 to G1 Transition* was reported as having a missing *Resulting BP* role error and an incorrect *Part of Process* role error. The remaining error discovered in the first phase of QA analysis was for the concept *Tumor Immunity* with an incorrect target of the role *Resulting BP*, which required a change of its target from *Cancer Progression* to *Tumor Progression*.

Table 5.9 Distribution of Erroneous Concepts in the *Biological Process* Hierarchy

Level (# Role Types)	# Concepts	# Concepts Analyzed	# Erroneous Concepts	Error Rate
1	352	80	7	8.75%
2	160	40	3	7.50%
3	57	57	9	15.79%
4	59	59	19	32.20%
5	4	4	2	50.00%
Total:	**632**	**240**	**40**	**16.67%**

The 2x2 contingency table (Table 5.10) was calculated for comparing the probability of erroneous concepts in the lower-half levels and upper-half levels. The results are statistically significant, since the *p*-value for the two-tailed Fisher's exact test equals 0.0008 ($p < 0.05$). Therefore, the results confirm Hypothesis 5.5 that concepts in the upper-half levels are more likely to have errors than concepts in the lower-half levels.

Table 5.10 The 2x2 Contingency Table for the Lower-half Levels and the Upper-half Levels

	# Erroneous Concepts	# Concepts w/o Errors	Error Rate
Lower-half (≤ 2 role types)	10	110	8.33%
Upper-half (≥ 3 role types)	30	90	25.00%

There were errors among four out of the seven kinds of roles in the *Biological Process* hierarchy. Table 5.11 shows the distribution of erroneous concepts for each of these four role types. The major issues were concepts with missing *Location* and *Part of Process* roles. For example, the concept *Erythrocyte Differentiation* is missing *Location* with a target value of *Bone Marrow*. In total, 16 concepts (40%) with errors were found for the *Location* role, and 20 concepts (50%) for the *Part of Process* role.

Table 5.11 The Number of Concepts Reported with Errors for Each Role Kind

Role	# Erroneous Concepts	Example Concept	Suggested Correction
Location	16	*Erythrocyte Differentiation*	Add the role with the target *Bone Marrow*
Resulting Anatomy	4	*Megakaryopoiesis*	Add the role with the target *Megakaryocyte*
Resulting BP	3	*T-Cell Activation*	Add the role with the target *T Cell Proliferation*
Part of Process	20	*Mitosis*	Remove the role with the target *Cell Cycle Process*

The secondary review phase of this study led to the confirmation of 33 errors for 32 concepts (80% = 32/40). These included nine errors for nine concepts concerning missing *Location* roles, four errors for four concepts concerning missing *Resulting Anatomy* roles, and 20 errors for 20 concepts with incorrect *Part of Process* roles that should be removed. One of the confirmed erroneous concepts *Megakaryopoiesis* is missing both *Location* and *Resulting Anatomy*.

Table 5.12 shows the distribution of confirmed erroneous concepts according to the lower-half levels and upper-half levels. For example, the secondary review phase by the NCIt curator confirmed 27 concepts out of 30 concepts (90% = 27/30) in the upper-half levels reported in the initial phase of QA as erroneous. The two-tailed *p*-value by

Fisher's exact test is less than 0.0001, meaning there was also statistical significance in the difference between the numbers of confirmed erroneous concepts in the lower-half levels and the upper-half levels.

Table 5.12 Erroneous Concepts in the Lower-half and Upper-half Levels Confirmed by the NCIt Curator

	# Erroneous Concepts	**# Concepts w/o Errors**	**Error Rate**
Lower-half (≤ 2 role types)	5	115	4.17%
Upper-half (≥ 3 role types)	27	93	22.50%

Some of the results of the secondary review phase are summarized in Tables 5.13 and 5.14. Table 5.13 lists examples of errors that were confirmed on review by the curator of NCIt. For example, *Megakaryopoiesis* is indeed missing the role *Location* with the target *Bone Marrow* as well as the role *Resulting Anatomy* with the target *Megakaryocyte*. An internal modeling rule used by the NCIt team expressly forbids the target of a *Part of Process* role from simultaneously being an ancestor of the source concept. As noted, this conflicting semantics was observed during the first phase of the QA analysis. Thus, all such errors were confirmed during the secondary review phase. Three examples of this error are given in Table 5.13. Table 5.14 shows examples of errors for each kind of role that were rejected by the curator along with the reasons for the rejection. For example, the suggestion that the concept *Expiration* be given the role *Location* with the target *Lung* was rejected, because of the fact that expiration can involve other locations besides the lung. Table 5.15 shows the breakdown of the errors according to the various types of errors.

Table 5.13 Example Concepts with Confirmed Errors in the *Biological Process* Hierarchy

Role	Concept with Confirmed Error	Target of Role	Corrective Action
Location	*Megakaryopoiesis*	*Bone Marrow*	Add the role
Location	*Mismatch Repair*	*Chromosome*	Add the role
Resulting Anatomy	*Epithelial Cell Proliferation*	*Epithelial Cell*	Add the role
Resulting Anatomy	*Megakaryopoiesis*	*Megakaryocyte*	Add the role
Part of Process	*Antigen Presentation*	*Immune Response Process*	Remove the role
Part of Process	*Anaphase*	*Cell Cycle Process*	Remove the role
Part of Process	*Positive Regulation of Mitosis*	*Cell Cycle Process*	Remove the role

Table 5.14 Example Concepts with Rejected Errors in the *Biological Process* Hierarchy

Role	Reported Example of Concept Missing Role	Proposed Target of Missing Role	Reason for Rejection
Location	*Expiration*	*Lung*	Other locations can involve *Expiration*
Resulting BP	*T-Cell Activation*	*T-Cell Proliferation*	Incorrect

Table 5.15 Erroneous Concept Distribution by Error Types for Concepts in Each Level and for the Lower-half Levels (Levels 1-2) and the Upper-half Levels (Levels 3-5) of the Area Taxonomy

Level (# Role Types)	# Concepts Missing Role	# Concepts with Incorrect Role	# Concepts with Incorrect Role Target	Total
1	6	0	1	7
2	3	0	0	3
3	9	0	0	9
4	1	18	0	19
5	0	2	0	2
1-2	9	0	1	10
3-5	10	20	0	30

The disagreements between the domain expert and the curator in the two phases of the study can partially be explained by their different perspectives. In the initial phase of QA analysis, the work was carried out by (HM), who is outside the ontology's curatorial organizational structure. As such, her analysis was not influenced by any prescribed modeling approaches that may have been utilized in the ontology's original design and ongoing maintenance. Her job was to use her own judgment to point out any potential errors or inconsistencies and, from that analysis, to suggest changes (e.g., additions, corrections) to improve the ontology.

The secondary phase reviewer (the curator of NCIt) was obliged to work with an eye toward established protocols of the organization. For example, as noted, an internal NCIt rule says that a concept *A* cannot simultaneously be IS-A and *Part of Process* with respect to another concept *B.* Moreover, user-driven decisions are important to the curatorial staff. For example, in NCIt, the completeness of neoplasm concepts in the *Disease, Disorder or Finding* hierarchy is more important than that of non-neoplasm concepts due to the overall focus of the ontology. The lack of sufficient resources is also a factor. For example, additional, correct ontological elements are not necessarily included in NCIt unless there are compelling use-cases, to avoid the maintenance overhead involved as a result of such additions.

In summary, this study was performed to determine whether a measure of lateral complexity could be used as a guiding factor in QA. In particular, it was investigated whether more complex concepts are more prone to errors than simpler concepts. The foundational ontological unit of "role type" was used as the basis for the distinction between a complex and a simple concept. The outcomes of the two-phase QA study on

the NCIt's *Biological Process* hierarchy indeed showed a statistically significant difference between the error rate of the more laterally complex concepts *vis-à-vis* the error rate of simpler concepts. As such, this distinction can be used to guide ongoing efforts in ontology QA.

5.1.5 Quality Assurance of Complex Concepts in ChEBI

ChEBI was introduced in Section 2.1.4. Because it comprises a large collection of concepts and their interconnections and it undergoes frequent changes, it is not reasonable to expect that ChEBI would be completely free of modeling errors and inconsistencies. In fact, its curatorial team maintains a GitHub issue tracking system [53] to allow the user community to report problems as well as request various modifications to the ontology. Any modeling problems persisting in ChEBI could have an adverse impact on the applications dependent on it. As such, quality assurance (QA) of ChEBI's content is a critical maintenance task. Due to ChEBI's magnitude, repeated comprehensive QA reviews are not practical.

This section describes a semi-automated approach that concentrates QA efforts on complex concepts in ChEBI expected to harbor modeling problems with a higher likelihood. Similar as in the study in Section 5.1.4 on the NCIt *Biological Process* hierarchy, the number of lateral relationship types that a concept exhibits was considered as a measure of concept complexity. The more aspects to a concept's definition—from an interconnectedness perspective—the more involved and complex such a concept is and the more modeling errors can be expected. A structural artifact that is very helpful in classifying concepts along these complexity lines is the *area taxonomy*.

5.1.5.1 Methods. The complete area taxonomy for ChEBI's inferred version has a total of 135 areas, spanning nine levels. (The asserted version of ChEBI is released by the EMBL-EBI and includes all explicitly defined knowledge, while the inferred version was obtained by running a reasoner on the asserted version.) Figure 5.7 shows an excerpt of the area taxonomy consisting of 62 areas, each of which contains at least 10 ChEBI concepts. To save space, relationship names have been letter-coded, with the legend appearing in the figure. For example, the area {B, C} is {*has parent hydride*, *has part*}. At the left side of the figure, the total number of areas and the total number of concepts at each level are for the complete area taxonomy, not the excerpt shown. *Child-of* links have all been omitted from the figure. Note that the inclusion of the most prominent root (with most descendants) serves as an illustration of the semantics elaborated in an area. For example, for the area {*has parent hydride*, *has part*} with 918 concepts at the left-most position on Level 2 in blue, the root *organic amino compound* (385) gives an idea about the nature of the chemical concepts in the collection, which happens to include 92 cyanides. The areas in Figure 5.7 represent a total of 60,786 ChEBI concepts (98.2%).

The goal of this study was to determine whether or not ChEBI concepts with more relationship types have a higher expectation of being in error vis-à-vis concepts with fewer relationship types. With relationships representing the most critical components of concepts' logical definitions, it is reasonable to rely on them to measure a form of complexity. As an example, consider the ChEBI concept *L-alanine* defined with the following eight relationship types: *has functional parent*, *has parent hydride*, *has part*, *has role*, *is conjugate acid of*, *is conjugate base of*, *is enantiomer of*, and *is tautomer of*. It

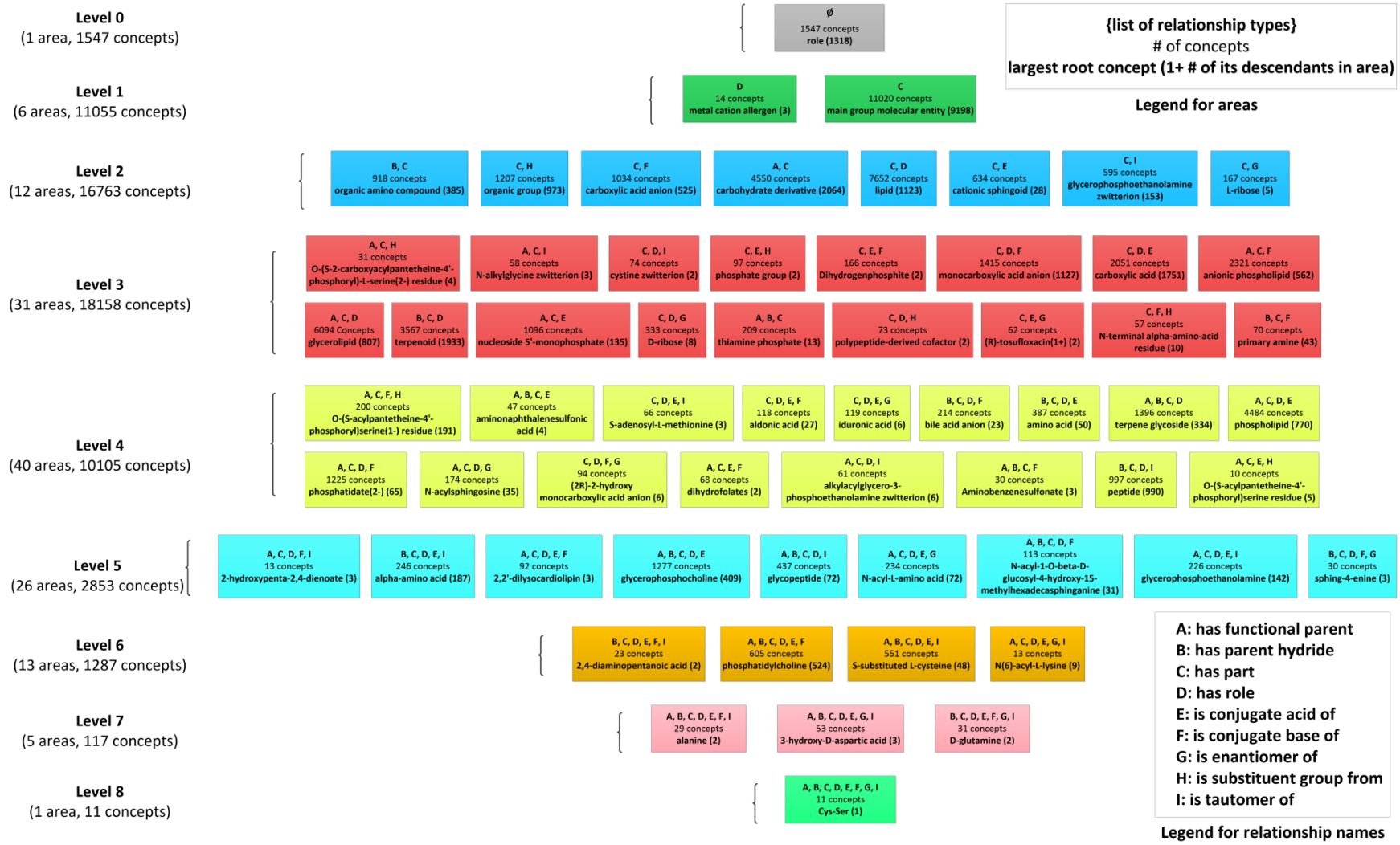

Figure 5.7 A 62-area excerpt of ChEBI's area taxonomy (which has a total of 135 areas).

is one of the 11 concepts in the highest numbered level (Level 8) area of the area taxonomy of Figure 5.7. This concept elaborates eight different aspects of a chemical entity and can be considered more complex than, say, *polypeptide-derived cofactor* exhibiting three aspects. *Polypeptide-derived cofactor*, in turn, is more complex than its parent *organic group*, which has only the two relationship types, *has part* and *is substituent group from*.

To make the determination about concept-error likelihood, a QA-analysis study of a random sample of the concepts in ChEBI was performed. In the study, 300 ChEBI concepts (about 0.5% of the entire ontology) were sampled based on the level arrangement of the area taxonomy in Figure 5.7. In particular, concept selection from the various levels was based on the number of concepts in each level. From Figure 5.7, it can be seen that the number of concepts in each of the Levels 1, 2, 3, and 4 (more than 10,000 concepts) is significantly greater (by orders of magnitude) than that of the other levels. For Levels 1, 2, 3 and 4, the numbers of concepts were randomly selected approximately proportional to the respective numbers of concepts on each level. Due to the small number of concepts at Level 8 (highest numbered level, most complex concepts) and its importance for this study, all its 11 concepts were included for QA analysis. Similarly, for Levels 6 and 7 with relatively small sizes, 11 concepts each were randomly selected to match Level 8's contribution. For Level 5 (2,853 concepts), the number of concepts selected was 20, reflecting a percentage that is between the higher percentage of Level 6 and the lower percentages for the Levels 1–4. The concepts completely lacking relationships (i.e., those on Level 0 in area ∅ of Figure 5.7) were ignored in the study

since they tend to be more general and abstract concepts, such as *chemical entity*, *molecular entity*, and *mineral*.

The actual QA analysis of the 300 ChEBI concepts was performed by two chemistry subject-domain experts using a multi-step process. In the first step, every sample concept was analyzed by the two experts independently, with no communication permitted between them. Their respective results were tabulated in two error reports. Each error finding was accompanied by the rationale for the judgment plus a suggested means of remediation.

In the second step, a combined error report, listing both experts' respective error findings for all sample concepts, was shared with the two experts. Each was separately asked to mark their agreement or disagreement with the findings of the other person. Furthermore, they were asked to review their own findings in light of the other expert's decisions. In this phase, each expert was permitted to change their mind regarding their own original judgment of a discovered error. A concept previously deemed to be exhibiting a modeling error could instead be deemed correct, and vice versa. In the final step of the analysis, a concept was marked *erroneous* if the two subject-domain experts agreed on that conclusion. Such findings collectively formed the *consensus QA report*, upon which the results are based.

The subject-domain experts were requested to look for *errors of commission* and *errors of omission*. Errors of commission included problems such as incorrect hierarchical relationships, incorrect lateral relationships, and incorrect relationship targets. Errors of omission included missing hierarchical relationships and missing lateral relationships. At the conclusion of the QA study, the errors of commission, the more

severe kind, were first submitted to ChEBI's curators for review, and later the errors of omission were submitted.

In regard to the QA analysis, the validity of the following hypothesis expressed in terms of area taxonomy levels was investigated:

Hypothesis 5.6: For a given ontology, concepts on the lower numbered levels of its area taxonomy have a lower average number of errors than concepts on the higher numbered levels.

Note that Hypothesis 5.6 does not specify the boundary between the lower numbered levels and the higher numbered levels. For confirmation of Hypothesis 5.6, it is sufficient that there exists a level *m* such that the average error rate for Levels 1, 2, …, *m* is lower than the average rate for Levels *m* + 1, *m* + 2, …, *n*, where *n* is the highest numbered level in the area taxonomy. The selection of the level *m* will be done in a way to maximize the difference between the two averages. The implication of verifying Hypothesis 5.6 is that the set of concepts with more relationship types offers a characterization of concepts for which more errors are expected to be found. A methodology focusing QA efforts on such a set is expected to yield more corrections than auditing a random set of the same number of concepts with fewer relationship types. The two-tailed Fisher's exact test [118] was used to analyze the results and judge the statistical significance of the difference between the error rates for the lower numbered levels and the higher numbered levels of the area taxonomy.

5.1.5.2 Results. The sample of 300 ChEBI concepts was taken from the February 2016 release based on the levels of its area taxonomy. The first-step QA analysis of the sample concepts was done by two chemistry domain experts Dr. Ling Chen (LC) and Dr.

Hasan Yumak (HY). Out of the 300 concepts analyzed, 155 of them (51.7%) made it into the consensus report, i.e., were deemed by both experts in the second-phase analysis to be erroneous. In the following, a ChEBI concept will often be referred to by its name together with its unique ChEBI ID, written in a format such as "CHEBI: 63667" (which happens to be the concept with the name *dipyridodiazepine*).

Table 5.16 shows the distribution of all ChEBI concepts with respect to the levels in the area taxonomy, the number of those that underwent QA analysis, and the number of erroneous concepts. For example, in the area taxonomy, there are 10,105 concepts at Level 4 (Figure 5.7), of which 44 (0.44%) were randomly selected for QA analysis. The domain experts found 22 concepts (50.0%) of the analyzed concepts on Level 4 to be erroneous. Note that as the level number increases, the percentage of erroneous concepts at each level (last column) does not decrease, i.e., the error rate shows a monotonic trend.

There are two cases in Table 5.16 where the error rate increases significantly, between Level 2 and Level 3 and between Level 4 and Level 5. Hence, in this case there are two options of how to divide the concepts by their levels into simple and complex concepts. Table 5.17, the 2x2 contingency table, presents the comparison of the cumulative error rates of Levels 1–4 and Levels 5–8. For example, 44.5% of the sample concepts on Levels 1–4 have errors in the consensus report. These results are statistically significant, because the *p*-value (two-tailed Fisher's exact test) is less than 0.0001 ($p < 0.05$). Table 5.18 gives the comparison between the cumulative error rates of Levels 1–2 and Levels 3–8. The corresponding *p*-value is 0.0003. The result for this division is also statistically significant even though the *p*-value is slightly higher. Hence, the results of the study confirm the hypothesis that concepts with more relationship types are more

Table 5.16 Distribution of Erroneous Concepts According to Levels in the Area Taxonomy

Level	# Concepts	# Analyzed Concepts	# Erroneous Concepts	% of Erroneous Concepts
1	11,055	49	19	38.8%
2	16,763	74	29	39.2%
3	18,158	80	40	50.0%
4	10,105	44	22	50.0%
5	2,853	20	14	70.0%
6	1,287	11	9	81.8%
7	117	11	11	100.0%
8	11	11	11	100.0%
Total:	60,349	300	155	51.7%

Table 5.17 The 2x2 Contingency Table for the Lower Numbered Levels (Levels 1–4) and the Higher Numbered Levels (Levels 5–8) with *m*= 4

	# Erroneous Concepts	# Concepts w/o Errors	Error Rate
Level 1 – Level 4	110	137	44.5%
Level 5 – Level 8	45	8	84.9%

Table 5.18 The 2x2 Contingency Table for the Lower Numbered Levels (Levels 1–2) and the Higher Numbered Levels (Levels 3–8) with *m*= 2

	# Erroneous Concepts	# Concepts w/o Errors	Error Rate
Level 1 – Level 2	48	75	39.0%
Level 3 – Level 8	107	70	60.5%

likely to exhibit errors than concepts with fewer relationship types, for both of the above optional dividing points.

Table 5.19 shows the different kinds of errors encountered from the ontological perspective. For example, there were 48 concepts (16.0%) having incorrect relationship

targets, an error of commission, and 105 concepts (35.0%) with missing hierarchical relationships, an error of omission.

Table 5.19 Error Distribution from the Ontological Perspective

Error Type	# Erroneous Concepts	% (/300)
Incorrect relationship target	48	16.0%
Incorrect hierarchical relationship	42	14.0%
Missing hierarchical relationship	105	35.0%
Missing lateral relationship	7	2.3%

An interesting question was whether the more complex concepts are also exhibiting a higher rate of errors of commission than the simpler concepts. Among the 88 erroneous concepts with errors of commission, 60 (24.3% = 60/247) were from Levels 1–4 and 28 concepts (52.8% = 28/53) were from Levels 5–8. Furthermore, the difference in the rates of errors of commission between Levels 1–4 and Levels 5–8 has statistical significance, because the *p*-value (for the two-tailed Fisher's exact test) is less than 0.0001. Another observation is that out of the 48 concepts with the error of *Incorrect charge difference between conjugate acids and bases* (Row 2 in Table 5.20), 26 concepts are from Levels 1–4 and 22 concepts are from Levels 5–8. So for this special kind of error, the error rate in Levels 1–4 is 10.5% (= 26/247) and the error rate in Levels 5–8 is 41.5% (= 22/53). This difference also has statistical significance with $p < 0.0001$. Hence, analyzing the more complex concepts for errors is a more efficient way.

Table 5.20 presents the typical kinds of chemistry-based modeling errors in the consensus report along with their numbers and sample percentages. For example, 11 concepts (3.7% of the sample) were found to exhibit incorrect amide classifications. Note that the kinds of errors in the table are not necessarily disjoint, meaning some concepts

may have several kinds of errors. As an example, *piperidine* (CHEBI: 18049) has both an incorrect conjugates charge error and an incorrect chemical classification error.

Table 5.20 Typical Chemistry-based Errors

Error Type	**# Erroneous Concepts**	**% (/ 300)**
Missing chemical classification	105	35.0%
Incorrect charge difference between conjugate acids and bases	48	16.0%
Incorrect chemical classification	21	7.0%
Incorrect amide classification	11	3.7%
Incorrect number of cyclic units	10	3.3%
Unmatched chemical name and structure	2	0.7%

Errors of commission are considered more severe than errors of omission, since they reflect incorrect modeling with respect to at least one aspect of a defined concept. On the other hand, there are more degrees of freedom regarding decisions about errors of omission, as it may have been a conscious editorial decision not to include some conceptual modeling details. An ontology's editorial policy may in fact dictate that some modeling elements be omitted, or it may simply be a matter of personal taste of an editor.

To further validate the findings, 62 concepts exhibiting errors of commission were first submitted to the curators for consideration. These were from among all the kinds of errors reported in Table 5.20 except for the first kind, "missing chemical classification," which is an error of omission. To date, ChEBI's curators have reviewed 49 concepts and confirmed 21 of them as being in error (42.9%). Some of the review details by ChEBI's curators are summarized in Tables 5.21 and 5.22. Table 5.21 lists examples of errors that were confirmed upon review and have subsequently been corrected in a new release of ChEBI. Table 5.22 shows examples of errors rejected by ChEBI's curators, along with

their reasons for this judgment. The errors of omission were submitted later via ChEBI's GitHub and are still awaiting review.

In summary, this study was carried out by two chemistry experts using the area taxonomy Abstraction Network to determine whether ChEBI concepts having more relationships—and in this sense higher complexity—warrant special attention in QA efforts. From the QA analysis of a random sample of ChEBI concepts consisting of both complex and simple concepts, it was confirmed with statistical significance that more complex concepts are more likely to harbor modeling errors than simpler concepts.

Table 5.21 Example Concepts with Confirmed Errors by ChEBI Curators

Concept with Confirmed Error	Confirmed Error	Error Explanation	Corrective Action
uric acid (ChEBI: 27226)	Incorrect target of the relationship *is conjugate acid of*: *urate anion*	Charge difference between conjugates should be 1	Add a new relationship *is conjugate acid of* with the target *urate(1−)*
trans-vaccenic acid (ChEBI: 28727)	Incorrect target of the relationship *is conjugate acid of*: *trans-vaccenate*	Charge difference between conjugates should be 1	Replace the target *trans-vaccenate* with *trans-vaccenate(1−)*
Malaoxon (CHEBI:6649)	Incorrect classifications: *dicarboxylic acid, carboxylic acid, hydroxides*	Chemical does not contain carbocyclic acid structure and hydroxyl group	Replace with the correct classification *organic thiophosphate*
Glucolepidiin (CHEBI: 5408)	Incorrect classifications: *glycosinolate, anion, polyatomic anion, ion*	No ion structure is shown in the structure	Replace with the correct classification *alkylglucosinolic acid*

Table 5.22 Example Concepts with Rejected Errors by ChEBI Curators

Concept with Reported Error	Reported Error	Reported Error Explanation	Reason for Rejection
pyrazolopyridazine (CHEBI:48383)	Incorrect classifications: *organic heterobicyclic compound, heterobicyclic compound*	Concept has 4 rings	Bicyclic, tricyclic, tetracyclic, etc., do not refer to the number of rings in a structure, but to the number of fused rings (i.e., rings that share one atom (spirocycles) or, more commonly, two atoms)
thermospermine (CHEBI:59564)	Incorrect target of the relationship *is conjugate base of*: *thermosperminium(4+)*	Charge of its conjugate base should be 1+ not 4+	Although the IUPAC definition of conjugate acid/base refers to a difference in charge of 1 unit only, for ChEBI, this is relaxed to include multiple charge differences
nystatins (CHEBI:59676)	Incorrect classification: *polyketide*	Concept does not contain ketone groups	Polyketide is structurally a very diverse group of compounds. For this reason, ChEBI denotes 'polyketide' as "*is a*" *organooxygen compound*, rather than "*is a*" *carbonyl compound*. The ChEBI definition of polyketide was taken from the IUPAC Gold Book

5.1.6 Auditing the Chemical Ingredient Hierarchy Based on the IAbN

The Abstraction-Network-based QA framework can be summarized as follows. First, an Abstraction Network is developed to summarize the specific terminology [4]. An algorithm is developed and implemented to computationally derive this Abstraction Network from the terminology. Based on the Abstraction Network, characterizations of sets of concepts of the terminology that are expected to display a higher percentage of errors are identified, compared to a control sample [8, 60]. Those sets of concepts can be computationally retrieved [67, 129], because the characterizations of such sets of concepts are based on structural features.

One of the recurring themes in such characterizations has been that there are concepts that are more complex than "arbitrary" concepts of the terminology. Examples of characterizations of complex concepts include overlapping concepts [91, 92, 130] and multiple inheritance regions [6, 57]. Complex concepts are typically more error-prone. While those characterizations were based on deriving a partial-area taxonomy [5, 6, 59] their complexity stems from concepts having multiple generalizations through multiple parents, reflecting an entity that is simultaneously "this and that." Not surprisingly, the modeling of such concepts is more challenging and a higher ratio of errors can be expected for them.

The characterization of concepts that were tested in this study on the NDF-RT *CI* hierarchy is "drug ingredients belonging to only one ingredient group with multiple parent ingredient groups" in the IAbN. Such concepts fit the above theme of complex concepts being "this and that" and are expected to have higher error rates.

Hypothesis 5.7: Among drug ingredients belonging to only one ingredient group, those in an ingredient group with multiple parent ingredient groups are more likely to have errors than those in an ingredient group with only one parent ingredient group.

The drug ingredients from those ingredient groups that have multiple *parent ingredient groups* inherit multiple classifications. The more classifications the drug ingredients belong to, the more complex those ingredients are, which increases the possibility that the classifications may have errors.

Hypothesis 5.8: Among drug ingredients belonging to only one ingredient group, those in an ingredient group with more than two parent ingredient groups are more likely to have errors than those with exactly two parent ingredient groups.

To test the above hypotheses, a sample of drug ingredient concepts within only one ingredient group was reviewed by two chemistry domain experts Dr. Ling Chen (LC) and Dr. Hasan Yumak (HY). Table 5.23 shows the distribution of NDF-RT's drug ingredients appearing in exactly one ingredient group according to their group's number of parent ingredient groups. There were a total of 263 drug ingredients as study concepts picked from the ingredient groups that have multiple parent ingredient groups. The study concepts included 118 randomly selected drug ingredients with two parent ingredient groups plus all drug ingredients with three (118), four (25) or five (2) parent ingredient groups. Thus, in total there were 263 study concepts. The control concepts consisted of 170 drug ingredients randomly chosen from the ingredient groups that have only one parent ingredient group. Hence, the total number of reviewed drug ingredients in the study was 433.

Table 5.23 The Distribution of the Drug Ingredients in Exactly One Ingredient Group Based on their Number of Parent Ingredient Groups

# of Parent Ingredient Groups	# of Drug Ingredients	Percentage (Column 2/1851)
0	1	0.05%
1	1136	61.37%
2	569	30.74%
3	118	6.37%
4	25	1.35%
5	2	0.11%
Total:	1851	100.00%

The two domain experts were blind to the hypotheses and the sampling methodology. The concepts were presented in alphabetical order. There were three steps of the review process. First, each of the reviewers studied the sample individually and submitted an error report that consisted of identified errors with corresponding corrections.

The domain experts were instructed to review the hierarchical relationships of each concept for correctness and to mark those they considered incorrect. The individual error reports from the domain experts were combined into a single anonymized list of unique errors. In the second step, the list of combined errors was sent back to the domain experts who had to obtain a consensus. Each reviewer marked ‘agree’ or ‘disagree’ for each error in the list.

In the third step, an additional evaluation of the consensus result was performed by Joan Kapusnik-Uner (JKU), a pharmacologist who is leading First DataBank’s drug vocabulary standards initiatives. Only the errors agreed upon by both LC and HY were sent to JKU for the third round review. JKU recorded those concepts for which she agreed that there was an error in a hierarchical relationship. Thus, in this study a concept

was considered erroneous only if all three domain experts (LC, HY, and JKU) agreed on the error.

In the two initial auditing reports generated at the first step, the two reviewers agreed on 100 erroneous drug ingredients; 19 drug ingredients were judged as erroneous by one or the other reviewer. A new data set including all the errors reported by any of the two auditors, without the name of the originator of the error, was compiled and sent back to the two auditors for generating a consensus report. The two auditors gave their responses (agree or do not agree) to all the errors listed in the new dataset (in fact, they agreed on all errors at the second step), which were compiled into a consensus report including all 119 erroneous drug ingredients that both reviewers agreed to. Then the consensus report was reviewed by JKU. Only when an error of a concept listed in the consensus report of LC and HY was confirmed by JKU, then this concept was labeled "erroneous," i.e., a consensus of three reviewers was achieved for these concepts in this three step study. In fact, JKU confirmed all consensus errors reported by LC and HY.

Table 5.24 shows the error distribution of the 433 audited drug ingredients. The percentage of erroneous concepts increases with the number of parent ingredient groups (except for the small number with five parents).

Table 5.25 shows the contingency table for the control and study concepts to calculate the *p*-value for Hypothesis 5.7. The two-tailed *p*-value is less than 0.0001 by Fisher's exact test [118], which means that the drug ingredients from the ingredient groups that have multiple parent ingredient groups are statistically significantly more likely to have errors than those from the ingredient groups that have one parent ingredient group.

Table 5.24 The Statistical Analysis of the Auditing Results of the 433 Drug Ingredients

# of Parent Ingredient Groups	# of Audited Concepts	# of Erroneous Concepts	Error Percentage
1	170	22	12.9%
2	118	29	24.6%
3	118	55	46.6%
4	25	13	52.0%
5	2	0	0.0%
Total:	433	119	27.5%

Table 5.25 The 2x2 Contingency Table for the Control and Study Concepts

# of Parent Ingredient Groups	# of Erroneous Concepts	# of Concepts w/o Errors
1	22	148
>1	97	166

In order to test Hypothesis 5.8, the error counts of drug ingredients from the ingredient groups that have more than two parent ingredient groups were compared with those that have exactly two parent ingredient groups. Table 5.26 shows the contingency table for the concepts with two and more than two parent ingredient groups to calculate the *p*-value. The two-tailed *p*-value equals 0.0002 by Fisher's exact test, which means that Hypothesis 5.8 was confirmed.

Table 5.26 The 2x2 Contingency Table for the Concepts with Two and More Than Two Parent Ingredient Groups

# of Parent Ingredient Groups	# of Erroneous Concepts	# of Concepts w/o Errors
2	29	89
>2	68	77

Overall, there were 119 concepts (119/433 =27.5%) with errors. Some errors appeared at the parent level, while other errors were introduced at higher levels (up to several levels above the erroneous concept). The types of errors are summarized in Table 5.27. The sets of erroneous concepts for the different error types in Table 5.27 are not disjoint, since one concept may have multiple errors. Row 1 in Table 5.27 shows that eight concepts in this study are assigned wrong parents, e.g., *Loracarbef* was erroneously defined as child of *Cephalosporins,* while the direct parent of *Loracarbef* should be *Carbacephem.*

Row 2 and Row 3 show two most common errors in the study that cover most of the erroneous concepts. In Row 2 an organic (or inorganic) concept is assigned to both *Organic Chemicals* and *Inorganic Chemicals*, due to the inheritance from its ancestor classification, which can be either organic or inorganic. For example, *Sulfur Compounds* appear as inorganic or organic compounds that contain sulfur as an integral part of the molecule according to the definition. For example, *Rabeprazole* actually is an organic chemical while it is classified under both *Organic Chemicals* and *Inorganic Chemicals*, because the parent of its grandparent is *Sulfur Compounds.* Row 3 indicates that the specified chemical ring structures of a concept which is a *Heterocyclic Compound[s]* are contradicting each other. For example, a concept is assigned several classifications out of the set *R* = {"*Heterocyclic Compounds, 1-Ring," "Heterocyclic Compounds, 2-Ring," "Heterocyclic Compounds, 3-Ring"* and *"Heterocyclic Compounds with 4 or More Rings"*}. That is due to inheritance from its ancestor classifications, i.e., a concept may have several ancestor classifications (at a very general level) and each of its ancestor classifications may be under one of the four choices in *R*. For example, *Alosetron* is a 4-

ring structure with three fused rings. Its parent is *Carbolines* with a 3-ring structure, the parents of which are, *Pyridines* with a 1-ring structure and *Indoles* with a 2-ring structure. Hence due to transitivity, *Alosetron* is a *Heterocyclic Compound(s), 1-Ring*, and also a *Heterocyclic Compound(s), 2-Ring*, and even a *Heterocyclic Compound(s), 3-Ring.*

Row 4 represents the other types of erroneous classifications, happening above the parent level, which cover 26 concepts (21.8%). For example, *Hydrogen Peroxide* does not belong to *Electrolytes*, because it is a molecule without ions, and it is not an electrolyte.

Table 5.27 Examples of Error Types with Counts

	Error Type	**# of Erroneous Concepts**	**Percentage (Column 2/119)**	**Examples**
1	Incorrect direct classification (= wrong parent)	8	6.7%	Bisacodyl, Ertapenem, Loracarbef
2	Organic/Inorganic Chemicals classification	81	68.1%	Cyclomethicone, Oxyphenonium, Rabeprazole
3	Heterocyclic Compounds, X-Ring(s) (X is one of {1, 2, 3, 4 or more})	34	28.6%	Alosetron, Bilirubin, Ramipril
4	Other types of erroneous classifications	26	21.8%	Hydrogen Peroxide, Loracarbef, Levodopa

5.2 Quality Assurance of Concepts with Uncommon Modeling

The following sections report two quality assurance studies focusing on concepts with uncommon modeling. Both studies utilized the “prism” constituted by Abstraction Networks for the detection of concepts with uncommon modeling, which were found to

contain relatively higher ratios of errors. An Abstraction Network offers an alternative compact visualization of an ontology's structure and content, which helps in detecting various anomalies not visible in the structure of the ontology itself. The first study is on concepts in small partial-areas within a partial-area taxonomy. The other study is on concepts in the area without any relationship within an area taxonomy.

5.2.1 Auditing NCIt Neoplasm Concepts in Groups of High Error Concentration

There are two major activities that lead to corrections of ontologies. (1) Curators of ontologies receive occasional requests of users to correct modeling errors they find, but such requests are *ad hoc* and do not constitute a rigorous QA process. (2) Ontology maintenance teams execute internal QA processes to test and verify the correctness of every new release of an ontology. See, for example, the internal QA process of NCIt in place at the National Cancer Institute (NCI) as described by De Coronado et al. [42]. Automated QA processes can only expose errors detectable by algorithms, such as redundant role assignments. Such QA algorithms can detect structural errors but not semantic errors, which are more difficult to uncover.

Hence, there is a need for a rigorous QA process as an integral part of the life cycle of an ontology that detects semantic errors as well [5]. As in finance, software verification, etc., such QA processes should not be the responsibility of the editorial team of an ontology, but be outsourced to an external department or even an external organization that has no emotional attachment to the modeling decisions of the ontology and thus can be objective in an ontology review.

Considering the fact that ontology errors are created as a result of unintentional human mistakes, rather than occurring as natural phenomena, one might think that they

will be distributed uniformly over the concepts of an ontology. However, this study refutes the assumption that errors are uniformly distributed in the investigated medical ontology. While this phenomenon is not so obvious when viewing an ontology with existing visualization tools that do not perform summarization, it becomes clear when viewing the same ontology through the prism of an Abstraction Network, which provides guidance for where to look for errors. Therefore, an economical approach to the QA of ontologies is to identify structural characterizations of sets of concepts for which a relatively high rate of errors is expected, compared to a random control sample. Reviewing such sets of concepts by domain experts is expected to provide a high QA yield, measured by the ratio of concepts confirmed as erroneous for a given number of reviewed concepts. This study explored the QA methodology concentrating on concepts in small partial-areas of the partial-area taxonomy for the *Neoplasm* hierarchy in NCIt.

5.2.1.1 Materials and Approach. A partial-area in a partial-area taxonomy represents a particular set of concepts in the ontology. These concepts are similar in their structure and semantics. That is, they all share the same roles and the same root. When encountering the partial-area *Neoplasm* with 403 concepts, the modeling of its concepts is considered "common" – there are many concepts with the same neoplasm semantics and structural modeling – with the role *Disease Has Abnormal Cell*. However, when encountering a small partial-area of, say, of two concepts only, their combination of structure and semantics is unique to them among the thousands of concepts in the hierarchy; then this would be a case of "uncommon modeling," since no other concepts have the same combination of structure and semantics.

It is, of course, possible that an ontology correctly contains only two concepts that

are represented with a specific structure and semantics. However, another option is that the reason for this uncommon modeling is that there is an error in how these two concepts are represented in the ontology. If so, once this error is corrected, say by adding a role or changing a parent link, then these concepts are likely to reappear in another partial-area to reflect the changes in their structures or semantics. It may well be the case that the new "home partial-area" is not small. This was an example scenario where the modeling was "uncommon" due to an error, and correcting the error(s) eliminated a small partial-area. In previous studies [5, 90, 120], small partial-areas were indeed found to be characterized by higher error rates.

As mentioned in Section 2.1.3, the *Neoplasm* subhierarchy of NCIt with 8,166 concepts is of special importance because of the priority given to modeling cancer-related concepts due to the mission of NCIt to support cancer research and care. Thus, the NCIt team is paying increased attention to the modeling of neoplasm concepts compared to many other concepts in NCIt.

Therefore, it is of interest to explore the problem whether a location where there is a higher concentration of erroneous concepts in the relatively large *Neoplasm* subhierarchy can be identified. Can the following hypothesis be confirmed with statistical significance?

Hypothesis 5.9: Small partial-areas in the *Neoplasm* sub-taxonomy of the *Disease, Disorder or Finding* hierarchy of NCIt harbor sets of concepts with higher rates of errors than large partial-areas.

Having the partial-area taxonomy available, 150 concepts from small partial-areas, defined here as having sizes between 1 and 10, were randomly selected as the study

sample. For the control sample, 40 concepts were randomly selected from large partial-areas of at least 20 concepts. The range 11-19 is considered to define medium-sized partial-areas. The study sample and the control sample were combined and the order of concepts was randomized.

The QA analysis on the 190 neoplasm concepts was carried out by three domain experts Dr. Yan Chen, Dr. Hua Min and Dr. Julia Xu who are domain experts in medicine with extensive experience in ontology QA. The review of the selected concepts involved two steps following the modified Delphi procedure [131]. Namely, first, each of three domain experts, being blind to the sampling technique that was used, independently reviewed all the concepts in the sample and reported all erroneous concepts. The three error reports were combined into a questionnaire, where each error identified by any of the experts was listed, without attributing it to the expert reviewer(s) who discovered it. For every reported error a suggested correction was included.

In the second step of the process, each of the reviewers marked whether she agreed with each error in the combined report or not. A concept is considered a "consensus erroneous concept" only if all three reviewers agreed that this concept has an error. Finally, the numbers and percentages of "consensus erroneous concepts" in small partial-areas and in large partial-areas were compared with each other.

5.2.1.2 Example of Group-based Auditing. The partial-area taxonomy of an ontology divides concepts into groups. The basic idea of the "group-based" auditing method is that if "many" concepts in a group are found to exhibit errors, then a conscientious auditor should review all the concepts of that group [91]. This is suported by work of He et al. [132], which showed that it is advisable to review the other concepts

of a group if some errors were found in the group. This is the abstract idea of group-based auditing.

Now follows a concrete example for illustrating group-based auditing, utilizing the partial-area taxonomy where the partial-areas function as groups. Figure 5.8 demonstrates a group-based auditing scenario. White boxes within colored boxes are partial-area nodes within area nodes. The indented format in each partial-area in Figure 5.8(a) represents the IS-A hierarchical structure inside of a partial-area node (e.g., *Benign Posterior Tongue Neoplasm* IS-A *Benign Tongue Neoplasm*, which IS-A *Benign Oral Cavity Neoplasm*). This detailed information is normally **not** shown in a partial-area taxonomy diagram, but is necessary for the demonstration of group-based auditing. The concept in bold (e.g., *Benign Oral Cavity Neoplasm*) in each partial-area node is the root, and the arrows denote the hierarchical *child-of* links between partial-area nodes. For example, the root concept *Oral Cavity Benign Granular Cell Tumor* IS-A *Benign Oral Cavity Neoplasm*, so there is a *child-of* link from the partial-area node *Oral Cavity Benign Granular Cell Tumor (2)* to the partial-area node *Benign Oral Cavity Neoplasm*

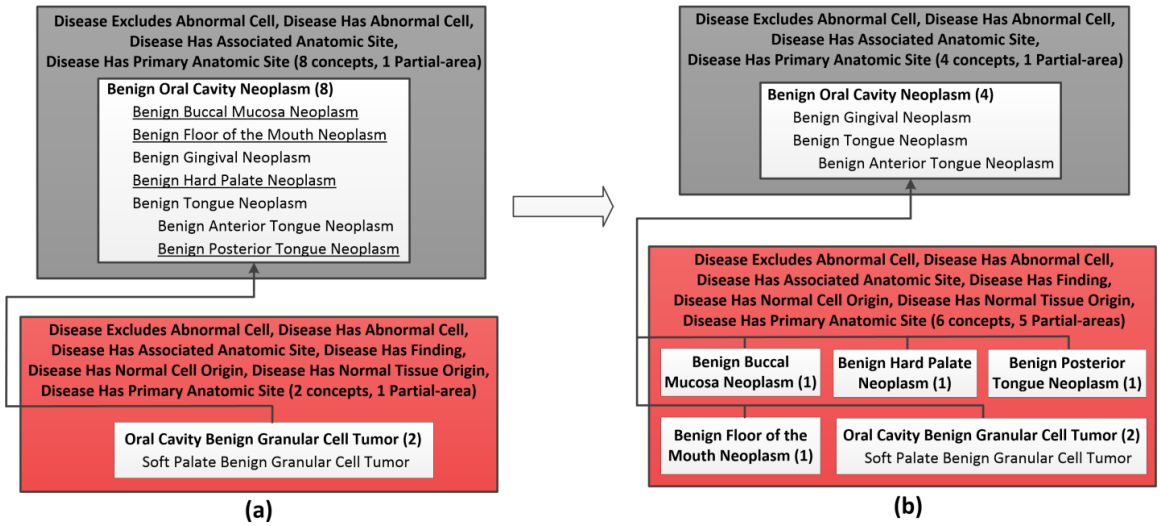

Figure 5.8 Example of the structure of the *Neoplasm* partial-area taxonomy **(a)** before and **(b)** after auditing.

(8). The number in parentheses () is the number of concepts summarized by a node.

Out of the eight concepts in the partial-area node *Benign Oral Cavity Neoplasm (8)* in the top area node (Figure 5.8(a)), only the four underlined concepts were in the random sample sent to the auditors (rows 1, 2, 4 and 7 under *Benign Oral Cavity Neoplasm (8)).* The auditors recommended adding the following three roles to these four concepts to improve the correctness of the modeling: *Disease Has Finding* with the target *Benign Cellular Infiltrate*, *Disease Has Normal Cell Origin* with the target *Connective and Soft Tissue Cell* and *Disease Has Normal Tissue Origin* with the target *Connective and Soft Tissue*. As noted above, this was a consensus decision.

Due to the suggested corrections, these four concepts should not remain in their current location. Rather they should appear in the lower area node {*Disease Excludes Abnormal Cell, Disease Has Abnormal Cell, Disease Has Associated Anatomic Site, Disease Has Finding, Disease Has Normal Cell Origin, Disease Has Normal Tissue Origin, Disease Has Primary Anatomic Site*}. When the partial-area taxonomy is re-derived, as shown in Figure 5.8(b), it is indeed the case that each of the four concepts now appears in the lower area node. Furthermore, each of the four concepts is a root in the area node, thus it gets its own partial-area.

After observing the above errors of concepts in the reviewed sample, the question whether the other four concepts in the same partial-area node might have the same errors as the four concepts in Figure 5.8(a) was raised. In other words, *group-based auditing of the remaining four concepts in the area node* at the top of Figure 5.9(a) was applied. These concepts were not in the random sample that was originally given to the auditors.

The consensus auditing result of the three domain experts confirmed that the four concepts *Benign Oral Cavity Neoplasm*, *Benign Gingival Neoplasm*, *Benign Tongue Neoplasm* and *Benign Anterior Tongue Neoplasm* were also missing the same three roles as the four concepts that were in the original sample. Thus, group-based auditing in this case doubles the number of errors found, with little extra effort.

Figure 5.9 Example of error correction propagation and the resultant partial-area taxonomy simplification; (**a**) shows the partial-area taxonomy before and **(b)** after the auditing/correction steps.

5.2.1.3 Error Correction Propagation and Partial-area Taxonomy Simplification.

Once correction of errors is achieved for a whole group, potentially through group-based auditing, one should consider the propagation of the correction of errors to descendant

groups. Errors of a parent group are typically inherited by descendant groups. Thus, one can easily examine the inheritance of the corrections suggested for such errors.

The method of *error correction propagation* will now be demonstrated, where the corrections of the above errors, discovered for concepts within a partial-area **A,** will be propagated to three descendant partial-areas of the partial-area **A**. This happens as follows.

During group-based auditing, the root concept *Benign Oral Cavity Neoplasm,* describing the semantics of the whole partial-area, was determined to miss the above three roles. Because of that, the 22 concepts in Figure 5.9(a) should all have the same roles, due to inheritance of the corrected set of relationships.

The concepts in the partial-area nodes of the area nodes in the second and third level from the top in Figure 5.9(a) originally had some, but not all, of the three missing roles discovered in Figure 5.8. However, those missing roles are now added due to inheritance from the corrected *Benign Oral Cavity Neoplasm* root concept. After adding these three roles, the 22 concepts in the eight partial-area nodes in Figure 5.9(a) will be summarized by only **one** partial-area node *Benign Oral Cavity Neoplasm (22)* in Figure 5.9(b), which is not small anymore.

Figure 5.9 demonstrates an interesting visual impact of the error correction propagation process described above, namely partial-area taxonomy simplification. The eight partial-area nodes in four different area nodes appearing in Figure 5.9(a) are unified into one single partial-area node of 22 concepts in Figure 5.9(b). That is, modeling errors are manifested when more small partial-areas “than needed” appear. The process described by Figure 5.9 shows the flip side of this observation, namely that the correction

of modeling errors may lead to a beneficial simplification of the partial-area taxonomy (the summary) of the ontology, by unifying several small partial-area nodes into one larger partial-area node. Furthermore, due to inheritance, the simplification occurred across several area nodes.

5.2.1.4 Results. The *Neoplasm* subhierarchy, containing 8,166 concepts (32.25% of the whole *Disease, Disorder or Finding* hierarchy) in the February 2015 release of NCIt, is summarized by 4,824 partial-area nodes in the *Neoplasm* partial-area taxonomy. The three experts found 76 concepts out of 190 (40.0% = 76/190) having errors. Table 5.28 shows the number of audited concepts (that appeared in the random sample), the number of erroneous concepts and the error rate for each partial-area node size. For example, the random sample contained 10 concepts from partial-area nodes with size=1, and among these the experts found five concepts (50.0%) exhibiting errors.

Table 5.28 Distribution of Erroneous Concepts According to Partial-area Node Size in the Partial-area Taxonomy

Partial-area node size	# of Concepts Audited	# of Erroneous Concepts	Error Rate (%)
1	10	5	50.0%
2	17	3	17.6%
3	10	5	50.0%
4	15	7	46.7%
5	13	3	23.1%
6	12	6	50.0%
7	21	8	38.1%
8	17	9	52.9%
9	19	10	52.6%
10	16	11	68.8%
11-19	-	-	-
≥ 20	40	9	22.5%
Total:	190	76	40.0%

According to the results in Table 5.28, the average error rate for the concepts from small partial-area nodes with up to 10 concepts is 44.7% (= 67/150), while the error rate for the concepts from large partial-area nodes, summarizing at least 20 concepts, is 22.5%. Table 5.29 summarizes the comparison of these two error rates. The error rate difference is statistically significant ($p < 0.05$ according to Fisher's exact test). Therefore, the results confirm Hypothesis 5.9, namely that concepts represented by small partial-area nodes are more likely to have errors than concepts in large partial-area nodes.

Table 5.29 The 2x2 Contingency Table for Small Partial-areas and Large Partial-areas

	# Erroneous Concepts	**# Concepts w/o Errors**	**Error Rate**
small partial-area nodes (size < 11)	67	83	44.7%
large partial-area nodes (size ≥ 20)	9	31	22.5%

With one exception, all errors reported by the three domain experts are errors of omission. Table 5.30 summarizes the distribution of erroneous concepts according to different error types. The most common type of error is the omission of roles, which occurs for 60 concepts from small partial-area nodes (40.0% = 60/150) and five concepts from large partial-area nodes (12.5% = 5/40). The second most common type of error is "missing parent" with 13 (8.67% = 13/150) concepts having this error among small partial-area nodes and three (7.5% = 3/40) among large partial-area nodes.

Table 5.31 lists some examples of concepts having modeling errors in their IS-A relationships. For example, the concept *Benign Epithelial Neoplasm* from a partial-area node summarizing only one concept is missing an IS-A relationship to *Benign Neoplasm*. Another concept *Benign Iris Neoplasm* from a small partial-area node was found missing

a child *Iris Nevus*. The experts also found one concept, *Combined Carcinoid and Adenocarcinoma*, from a large partial-area node having an incorrect IS-A relationship.

As mentioned in Section 2.1.3, the *Disease, Disorder or Finding* hierarchy has 29 role types. For the 150 concepts from small partial-area nodes, the three domain experts found 60 concepts missing 12 role types. Table 5.32 shows the distribution of these 60 erroneous concepts by role types and gives an example of one erroneous concept for each role type. For example, there are 10 concepts from small partial-area nodes missing the role *Disease Excludes Abnormal Cell*. *Acantholytic Squamous Cell Skin Carcinoma* is one of the 10 erroneous concepts, and it is missing this role with the target *Malignant*

Table 5.30 Comparison of Error Distribution by Types between Concepts from Small and Large Partial-area Nodes

Error Type	# Erroneous Concepts from Small Partial-area nodes	# Erroneous Concepts from Large Partial-area nodes	Total
Missing role	60	5	65
Missing parent	13	3	16
Missing child	1	1	2
Incorrect parent	0	1	1

Table 5.31 Examples of Erroneous Hierarchical Relationships

Partial-area node size	Error Type	Example Concept with Error	Reported Error
small	Missing parent	*Benign Epithelial Neoplasm*	Missing parent *Benign Neoplasm*
small	Missing child	*Benign Iris Neoplasm*	Missing child *Iris Nevus*
large	Missing parent	*Reproductive Endocrine Neoplasm*	Missing parent *Endocrine Neoplasm*
large	Missing child	*Intraventricular Brain Neoplasm*	Missing child *Glioblastoma* and *Pilocytic Astrocytoma*
large	Incorrect parent	*Combined Carcinoid and Adenocarcinoma*	Change the parent from *Carcinoma* to *Adenocarcinoma*

Basaloid Cell. Another issue discovered was that 32 concepts are missing the *Disease Has Finding* role.

The error correction propagation method (Section 5.2.1.3) was applicable to a number of concepts in the sample, leading to the unification of several partial-area nodes from several area nodes into one larger partial-area node. This can be observed, for

Table 5.32 The Number of Concepts from Small Partial-area Nodes Missing Roles for Each Role Type

Role Type	# Erroneous Concepts	Example Concept Missing Role Type	Target of Missing Role
Disease Has Finding	32	*Benign Cerebellar Neoplasm*	*Benign Cellular Infiltrate*
Disease Excludes Abnormal Cell	10	*Acantholytic Squamous Cell Skin Carcinoma*	*Malignant Basaloid Cell*
Disease Excludes Finding	8	*Amelanotic Melanoma*	*Favorable Clinical Outcome*
Disease Has Primary Anatomic Site	7	*High Grade Vaginal Intraepithelial Neoplasia*	*Vagina*
Disease Has Normal Tissue Origin	6	*Anal Canal Neuroendocrine Tumor*	*Neuroendocrine Tissue*
Disease Has Normal Cell Origin	5	*Granulosa Cell Tumor*	*Granulosa Cell*
Disease Excludes Primary Anatomic Site	4	*Acute Erythroid Leukemia*	*Lymphatic System*
Disease Has Associated Anatomic Site	3	*Hemolymphangioma*	*Lymphatic Vessel*
Disease Excludes Normal Tissue Origin	2	*Acute Erythroid Leukemia*	*Lymphoid Tissue*
Disease Has Associated Disease	2	*Adenocarcinoma in Multiple Adenomatous Polyps*	*Polyposis*
Disease Has Abnormal Cell	1	*Clear Cell Adenoma*	*Neoplastic Clear Cell*
Disease Mapped To Gene	1	*Borderline Ovarian Clear Cell Adenofibroma*	*ARID1A Gene*

example, by looking at the roots *Benign Muscle Neoplasm (11)*, *Benign Brain Neoplasm (15)* and *Benign Female Reproductive System Neoplasm (11).*

The consensus erroneous concepts (76) were submitted to the NCIt editorial team. The NCIt team confirmed 17 erroneous concepts (22.4%), of which only one concept is from a large partial-area (with size 53) and the other 16 concepts are from small partial-areas. The NCIt team did not review any other concepts from the 190 concept sample and thus their review cannot be considered an alternative QA study. Table 5.33 lists five example concepts with errors that were confirmed and corrected by the NCIt team. Table 5.34 shows another four concepts that were reported as having errors by the domain experts that were not corrected by the NCIt team. The third column in Table 5.34 reports the NCIt team's reasons for not correcting the concepts, while the fourth column presents the external domain experts' consensus counter arguments, explaining why nevertheless these errors should be considered legitimate. Table 5.35 shows the distribution of concepts into small, medium and large partial-areas.

To summarize this study, errors in the NCIt *Neoplasm* subhierarchy are not uniformly distributed. Uncommon modeling of concepts, which is reflected in small partial-areas in the partial-area taxonomy, resulted in a significantly larger percentage of erroneous concepts than in a control group of concepts from large partial-areas. The error rate for small partial-areas (44.7%) was twice as large as the error rate for large partial-areas (22.5%). Furthermore, group-based auditing, using groups constituted in the partial-area taxonomy, was demonstrated to support easy discovery of additional erroneous concepts in the same partial-areas of the partial-area taxonomy. By error correction propagation, additional errors at lower levels in the partial-area taxonomy were also

Table 5.33 Example Concepts with Errors Confirmed by NCIt Curators

Concept	Confirmed Error	Correction
Benign Buccal Mucosa Neoplasm	Missing roles: *Disease Has Finding* with targets *Benign Cellular Infiltrate* and *Indolent Clinical Course*	Add these two roles
Granulosa Cell Tumor	Missing role: *Disease Has Normal Cell Origin* with the target *Granulosa Cell*	Add the role
High Grade Vaginal Intraepithelial Neoplasia	Missing role: *Disease Has Primary Anatomic Site* with the target *Vagina*	Add the role
Human Papillomavirus-Related Malignant Neoplasm in AIDS Patient	Missing role: *Disease Has Associated Disease* with the target *Acquired Immunodeficiency Syndrome*	Add the role
Reproductive Endocrine Neoplasm	Missing parent: *Endocrine Neoplasm*	Add an IS-A link to *Endocrine Neoplasm*

Table 5.34 Example Concepts with Errors Not Corrected by NCIt Curators

Concept	Reported Error	Curator's Reaction	Counter Argument
High Grade Prostatic Intraepithelial Neoplasia	Missing role: *Disease Has Finding* with the target *High Grade Lesion*	This concept already has the role *Disease Is Grade* with the target *High Grade.*	The role *Disease Is Grade* pointing to *High Grade* only hints at the existence of the role *Disease Has Finding* pointing to *High Grade Lesion*. In an ontology, information should be explicit so it is usable for computers. Besides, some concepts in NCIt have both these roles.
Benign Skeletal Muscle Neoplasm	Missing role: *Disease Has Associated Anatomic Site* with the target *Skeletal Muscle Tissue*	This concept already has the role *Disease Has Normal Tissue Origin* with the target *Skeletal Muscle Tissue.*	Similarly as above, the suggested role is not implied by the existing one.
Lymphoplasmacyte-Rich Meningioma	Missing role: *Disease Has Primary Anatomic Site* with the target *Meninges*	This concept already has the role *Disease Has Normal Tissue Origin* with the target *Meninges*.	Same as previous. However, some concepts in NCIt, (e.g., *Benign Meningioma*), have both roles *Disease Has Primary Anatomic Site* and *Disease Has Normal Tissue Origin* pointing to *Meninges*.
Clear Cell Squamous Cell Skin Carcinoma	Missing parent: *Primary Malignant Neoplasm*	*Primary Malignant Neoplasm* is under "*Neoplasm by Special Category.*" The NCIt team had decided not to populate such terms with specific histopathologies and not to model/define such terms.	The NCIt curator agrees with the correction and does not implement it due to NCIt editorial policy.

Table 5.35 The Neoplasm Concept Distribution According to Partial-area Size

Partial-area size	# of Partial-areas	Total # of Concepts	% of Concepts
1-10	4762	6581	80.59%
11-19	44	642	7.86%
≥20	18	1014	12.42%
Total:	4824	8166	

found and corrected with minimal additional effort. On a more general level, this study concluded that Abstraction Networks were again successful in aiding the process of discovering "suspicious" concepts.

5.2.2 Quality Assurance of Concept Roles in NCIt Biological Process Hierarchy

Roles, an important component of NCIt modeling, are used to define concepts and are inherited down the hierarchies. The complete and accurate representation of biomedical knowledge for a concept through roles is important for the NCIt applications such as reasoning. Hence, it is necessary to conduct a QA study concentrated on missing role errors. One of the inherent concept groupings in an area taxonomy, called the *top area*, comprises all concepts without any roles at all. This is a natural place to search for concept with missing role errors. Besides, factors such as disproportional size or disproportional growth over time of the top area could be indicators in determining whether QA efforts are warranted. Moreover, the hierarchical depth of the top area can be another factor to be considered.

As mentioned in Section 2.1.3, 513 concepts (44.8% of the complete *Biological Process* hierarchy) are in the top area (see Figure 5.6 on Page 101). For comparison, in the year 2004 [5], only 47 concepts out of 589 concepts (8%) were in the top area. That is, while the *Biological Process* hierarchy grew about two-fold, the top area grew about

eleven-fold. When there is such disproportional growth of the top area, it can be interpreted as an anomaly alerting QA experts to the possibility of widespread missing-role errors.

For each NCIt hierarchy, there is a list of role types that can be used to define its concepts. The *Biological Process* (*BP*) hierarchy has seven possible associated roles whose full names and abbreviated names were given in Table 2.2. Figure 5.6 shows the complete area taxonomy of the *BP* hierarchy. This section presents a QA study on the concepts in the top area of the area taxonomy for the *BP* hierarchy.

5.2.2.1 Methods. Quality assurance (QA) efforts on large and complex biomedical ontologies need to be highly focused and aided by techniques that automatically identify sets of concepts that are suspicious or anomalous and warrant special attention. Such concepts are expected to be in error, with a high likelihood. One characterization of concepts shown to have a higher error rate identified by Abstraction Networks is uncommonly modeled concepts, which were described in Section 2.3.

In this study, a concept residing in the top area Ø was considered to be an uncommonly modeled concept, one of the major themes for concept sets expected to harbor errors at a high rate [4]. More specifically, when the top area Ø for a given hierarchy is disproportionately large, this can be taken to be anomalous. This means that a high percentage of concepts in the hierarchy suffer from a lack of roles and inherent under-definition. The *BP* hierarchy is an example where this phenomenon exists. To be sure, there are concepts that rightfully do not have any roles. But those are typically concepts capturing general subjects or categories, for which no roles are relevant due to their general nature, e.g., *Pathologic Process* and *Reproductive Process*. Typically, such

concepts reside close to the hierarchy's root as its children or grandchildren. However, in general, the number of such concepts is expected to be relatively small.

This study proposed that an anomalous top area warrants special attention in QA efforts in regard to the error of omission "missing role." The following Hypothesis 5.10 was formulated.

Hypothesis 5.10: If a large percentage of a hierarchy's concepts are in the top area of its area taxonomy, then the percentage of concepts in the top area that are lacking roles is statistically significantly higher than the percentage of such concepts in other areas.

A QA study was conducted to assess Hypothesis 5.10. In the study, the QA analysis of the *BP* hierarchy's top-area concepts and control concepts was carried out manually by the domain expert Dr. Yan Chen, who has medical training and extensive experience in ontology QA. A manual review by a domain expert is required since human understanding is needed for such judgements. However, the detection of sets of concepts with high likelihood of errors was performed algorithmically. The missing-role errors found in the analysis were submitted for secondary review and confirmation by a curator of the NCIt.

In some hierarchies of large ontologies, even the top area is very large, and a QA review of all its concepts is not practical. For example, in the February 2015 release of NCIt, the *Disease, Disorder, or Finding* hierarchy contains 25,360 concepts, and its top area has 14,347 concepts. Similarly, the top area in the *Clinical finding* hierarchy of SNOMED CT contains 7,000 concepts. In such a case, the challenge is to narrow down the QA effort to a more promising subset of the top area. This is where another version of

"complex modeling of concepts" was employed.

One way to measure the complexity of a concept is by the number of roles defined for it as in the two Sections 5.1.4 and 5.1.5. A concept with six roles is likely to be more complex than a concept with, say, one or two roles. Intuitively, it is more difficult to correctly model a complex concept than it is for a simpler concept, and thus there is a higher likelihood of introducing a modeling error for the former. However, this measure of complexity is not relevant to the top area, where concepts have no roles at all. Another way of characterizing the notion of "complex concept" in that context was needed.

The hierarchical distance of concepts from the root of the top area is one rational way to measure the complexity of concepts in the top area. For example, *DNA Major Groove Binding* has a path of seven IS-A links to the root concept, *Biological Process*, of the whole hierarchy (see Figure 5.10). The concepts along the path accumulate more complexity in their nature and definitions as they are getting farther away from the root.

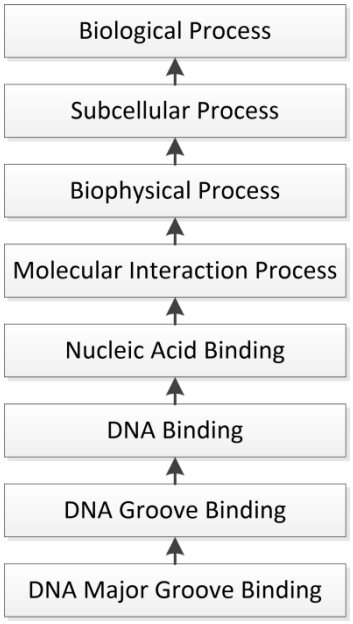

Figure 5.10 Path of seven IS-A links to the root.

In this light, the assumption is that the likelihood of a missing-role error increases with the additional refinement and complexity associated with the increasing distance from the root. In other words, one can expect a higher percentage of concepts with missing roles at the top area's levels with higher level numbers, where "level" is defined as the number of IS-A links in the path from the root to a given concept.

For example, in Figure 5.10, the level of *DNA Major Groove Binding* is seven. By definition, the root, *Biological Process*, resides on Level 0. (When a concept has multiple parents—and hence there are multiple paths to the root—its longest path defines its level. Topological sort [133] can be used to calculate the longest-path distance for all concepts in the top area in linear time.) Assuming a continuum of increased complexity along a path of concepts from the root, the levels of the hierarchy are divided into two halves, the upper and lower halves, with the expectation of more missing roles in the upper-half (by path length) of the hierarchy where concepts are more complex.

The above assumption has practical implications for QA in the case where the top area is too large to be reviewed in its entirety. In such a case, it is recommended that QA processing be concentrated on the levels with higher level numbers, since their concepts are generally more complex and are expected to have more missing roles. *These levels appear lower in the diagram.* In this regard, the following Hypothesis 5.11 was tested in the study. In this hypothesis, the phrase "upper-half levels" refers to levels with the numbers $\left\lfloor \frac{n}{2} \right\rfloor$, $\left\lfloor \frac{n}{2} \right\rfloor + 1$, …, n, assuming there are n levels in total in the top area. These are the levels farthest from the root. The "lower-half levels" are 0, 1, …, $\left\lfloor \frac{n}{2} \right\rfloor - 1$. These levels appear closer to the root in the diagram. For example, in Figure 5.10, where there

are eight levels, the lower-half levels are 0, 1, 2, and 3, and the upper-half levels are 4, 5, 6, and 7.

Hypothesis 5.11: Concepts in the upper-half levels of the top area have a higher likelihood of missing-role errors than concepts in the lower-half levels.

Since the *BP* hierarchy has an unusually large top area of 513 concepts (44.8% of the overall hierarchy), in the study, all concepts of its top area were reviewed for the specific error of "missing role." The number of erroneous concepts found in each level and their percentages were analyzed. As a control group, a random sample of 100 concepts was selected from all areas excluding the top area. However, since previous research with the NCIt *BP* hierarchy, mentioned in Section 2.3, by Min et al. [5] reported a higher likelihood of errors in small partial-areas, concepts from such groups were excluded so as not to bias this study.

Note that the QA analysis of concepts in the top area has potential implications beyond that area. Specifically, if a concept *C* from the top area is found to be missing a role *R* pointing to a target concept *D*, then all of *C*'s descendant concepts that do not have the role *R*, both inside and outside the top area, should also have the role *R*. Such a role R, if added, will point either to the same target *D* or to a descendant of *D*. If some such descendants do not have *R*, then they must be missing it. Hence, there is a potential propagation of the QA efforts for the top area into other areas of the hierarchy. Such propagation will save the QA analyst effort, since the additional concepts to be reviewed can be identified automatically by looking for missing roles.

5.2.2.2 Results. The overall results are summarized in Table 5.36, which shows the level distribution of concepts in the top area and the number of concepts found to be

Table 5.36 Missing-role Error Distribution by Level in the Top Area

Level	# Concepts	# Concepts Missing Roles	% of Concepts Missing Roles
0	1	0	0%
1	7	0	0%
2	69	15	21.7%
3	138	53	38.4%
4	125	58	46.4%
5	88	61	69.3%
6	44	32	72.7%
7	14	8	57.1%
8	23	5	21.7%
9	4	0	0%
Total:	513	232	45.2%

missing roles at the different levels. For example, at Level 5, consisting of 88 concepts, 61 concepts (69.3%) were missing roles. Out of the 513 concepts in the top area, 45.2% were missing roles. Furthermore, as the level number increases, the percentage of concepts missing roles at each level also increases. The exceptions to this are Levels 7, 8, and 9, probably due to the fact the numbers of concepts at these three levels are relatively small. It is not surprising that there are no concepts missing roles at Level 0 and Level 1, because these concepts are very general and roles are often introduced at more specific levels. For example, two general concepts at Level 1 are *Regulatory Process* and *Pathologic Process*.

Table 5.37 lists the number of concepts reported as having missing roles, for each different kind of role, and how many of them were confirmed. For example, 103 concepts were deemed to be missing the role *Location*, but only 84 of these were confirmed in the secondary review. As can be seen in the table, the largest numbers of missing roles in the

Table 5.37 Number of Concepts in the Top Area Reported Missing Roles for Each Role Kind

Role	# Concepts w/Missing Role	# Concepts Confirmed by Curator
Location	103	84
Initiator Chemical or Drug	1	0
Initiator BP	2	0
Resulting Anatomy	1	1
Resulting BP	3	1
Resulting Chemical or Drug	20	10
Part of Process	113	4
Total:	232	99

initial QA analysis appeared with respect to *Location* and *Part of Process*. But the curator's highest levels of agreements were for *Location* (82%) and *Resulting Chemical or Drug* (50%).

Table 5.38 shows five examples of concepts missing various kinds of roles confirmed by the curator of NCIt. For example, *Adrenal Hormone Activity Induction* is indeed missing the role *Location* that should be directed to the concept *Adrenal Gland*. On the other hand, Table 5.39 shows some examples of findings that were rejected by the curator, along with accompanying reasons. For example, initial QA analysis deemed the concept *Glucocorticoid Secretion Process* to be missing the *Resulting Chemical or Drug* role directed to *Glucocorticoid*. However, while it is true that in order for a product (e.g., a hormone) to be secreted, it first has to be produced, the set of processes (and enzymes) involved in production may not overlap with those involved in secretion. (Thyroid hormone is a good example of a product where production and secretion are two completely separate processes.)

Table 5.38 Examples of Concepts Confirmed to Have Missing Roles in the Top Area for Different Roles

Role	Example Confirmed Concept Missing Role	Target of Missing Role
Location	*Adrenal Hormone Activity Induction*	*Adrenal Gland*
Resulting Anatomy	*Coagulation Process*	*Fibrin*
Resulting BP	*Evolution*	*Genetic Drift*
Resulting Chemical or Drug	*Histamine Production*	*Histamine*
Part of Process	*Postpartum Recovery*	*Postpartum Process*

Table 5.39 Rejected Examples of Concepts Missing Roles in the Top Area for Different Roles

Role	Reported Example of Concept Missing Role	Proposed Target of Missing Role	Reason for Rejection
Location	*RNA Processing*	*Nucleus*	Not always true
Resulting BP	*Antigen Binding*	*Immune Response Process*	Not always true
Resulting Chemical or Drug	*Glucocorticoid Secretion Process*	*Glucocorticoid*	Secretion processes do not produce chemicals
Part of Process	*Defecation*	*Gastrointestinal Process*	*Gastrointestinal Process* is the parent of *Defecation*

The examples of Table 5.39 demonstrate the subtleties of the modeling issues involved and that it is possible that different experts differ in their opinions. The last example in Table 5.39 demonstrates two legitimate modeling options. *Defecation* IS-A *Gastrointestinal Process* and can also be viewed as *Part of Process* linked to *Gastrointestinal Process*. The curator, however, followed rules established in the overall modeling of the *BP* hierarchy.

Out of the 100 control concepts gleaned from non-top areas, 13 concepts were found by the domain expert to be missing roles. Table 5.40 shows the contingency table

Table 5.40 The 2x2 Contingency Table for the Concepts with Errors in the Top Area and Non-top Areas

	# Erroneous Concepts	# Concepts w/o Errors
Non-top areas	13	87
Top area	232	281

for the control (concepts in non-top areas) and study concepts (those in the top area). The Fisher's exact test two-tailed [118] *p*-value is less than 0.0001, which means the result has statistical significance. In other words, the concepts in the top area are significantly more likely to have missing roles than concepts in non-top areas. Thus, Hypothesis 5.10 is confirmed. Out of the 13 erroneous control concepts, the secondary review of the NCIt curator confirmed 10 concepts (76.9% = 10/13). Table 5.41 is the corresponding contingency table for erroneous concepts in the top and non-top areas confirmed by the NCIt curator. The two-tailed *p*-value by Fisher's exact test is 0.0311, which also confirmed Hypothesis 5.10 only considering the confirmed erroneous concepts.

Table 5.41 The 2x2 Contingency Table for Erroneous Concepts in the Top Area and Non-top Areas Confirmed by the NCIt Curator

	# Erroneous Concepts	# Concepts w/o Errors
Non-top areas	10	90
Top area	99	414

Table 5.42 summarizes the comparison between concepts missing roles at the lower-half levels (Levels 0–4) and those missing roles at the upper-half levels (Levels 5–9). There are 340 concepts in the lower-half levels, which is nearly twice of the 173 concepts in the upper-half levels. However, the percentage of concepts missing roles in the upper-half levels (61.3%) is higher than that in the lower-half levels (37.1%). The

two-tailed *p*-value is less than 0.0001 by Fisher's exact test. Thus, the results confirm Hypothesis 5.11 that concepts in the upper-half levels (Levels 5–9) of the top area have a significantly higher likelihood of missing roles than those in the lower-half levels.

Table 5.42 The 2x2 Contingency Table for Concept Errors between the Lower-half Levels and Upper-half Levels

Level Range	# Erroneous Concepts	# Concepts w/o Errors
0–4 (lower-half)	126	214
5–9 (upper-half)	106	67

Out of the 513 concepts in the top area, 354 concepts (69%) are leaves, i.e., do not have any descendants. Among the 159 non-leaf concepts, 68 concepts were found to be missing roles. Due to role inheritance, after correction, the descendants of those 68 concepts should now also have the same kinds of roles, with targets that are the same or more specific. Therefore, for those descendants outside the top area, it is necessary to check whether they have the roles that were missing in their ancestors in the top area. The results for the descendants of these 68 concepts are shown in Table 5.43. For five of them (Row 1), all their descendants are in non-top areas. For another 40 (Row 3), all their descendants reside with them in the top area. For the remaining 23 concepts (Row 2), some of their descendants are in the top area with them and others reside in areas below. The number of affected descendants reported (last column of Table 5.43) is the number of descendant concepts missing the same roles as their ancestors plus the number of descendants having the roles, but with targets different from their ancestors' (and not more specific than those). Figure 5.11 shows the new area taxonomy of the *Biological Process* hierarchy to illustrate the changes that occurred as a result of the QA analysis, including corrections in the non-top areas due to the propagation of the additional roles.

Table 5.43 Affected Descendants of the 68 Non-leaf Concepts Missing Roles in the Top Area

	# Concepts	Total # Descendants Outside Top Area	# Affected Descendants
All Descendants are in Non-Top Areas	5	15	5
Some Descendants are in the Top Area	23	102	50
All Descendants are in the Top Area	40	N/A	N/A
Total:	68	117	55

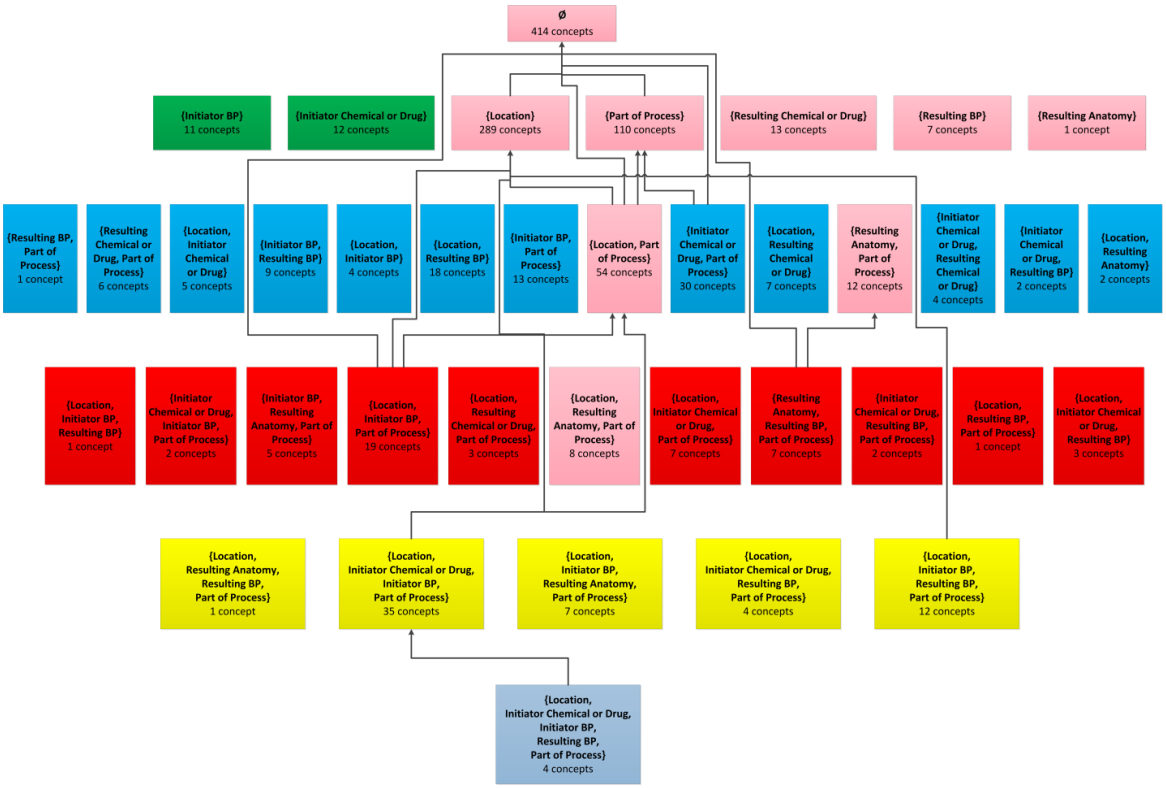

Figure 5.11 Revised area taxonomy for the *Biological Process* hierarchy incorporating the confirmed corrections. Pink highlights the areas that are different from the original in Figure 5.6.

In conclusion, this study introduced a QA methodology targeted at missing-role errors. The foundation of the approach was an Abstraction Network called "area taxonomy." An anomalous feature of the area taxonomy, when present, was used as an

indicator in guiding the QA analysts in their search for missing-role errors. The methodology was demonstrated with an application to the NCIt's *Biological Process* hierarchy. A statistically significant number of missing-role errors was discovered by an external reviewer and confirmed by a curator of the NCIt. Overall, the methodology can be seen as a useful addition to the arsenal of tools available to QA personnel.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In conclusion, this dissertation conducted several studies based on Abstraction Networks with the goal of trying to solve the Big Knowledge to Use (BK2U) problem, which is the implied problem of the Big Data to Knowledge (BD2K) challenge. The studies in this dissertation can be summarized under the following two main subjects:

1. Advanced Abstraction Networks for Big Knowledge summarization and visualization: the weighted aggregate partial-area taxonomy and the Ingredient Abstraction Network (IAbN).
2. Applications of the Big Knowledge summarization and visualization techniques: the identification of major topics in ontologies, the multi-layer multi-granularity visualization scheme for ontology comprehension, the discovery of Drug-Drug Interactions and the family-based quality assurance of large biomedical ontologies.

The weighted aggregate partial-area taxonomy was developed based on the partial-area taxonomy, which provides a more compact summary of ontologies to get a better big picture of the content in ontologies compared with the partial-area taxonomy. The weighted aggregate partial-area taxonomy was applied to identify and visualize major topics in SNOMED CT's *Specimen* hierarchy with the guidance of a list of gold standard topics provided by a domain expert. A new multi-layer multi-granularity visualization approach based on the weighted aggregate partial-area taxonomy was developed for the comprehension of Big Knowledge in ontologies.

The weighted aggregate partial-area taxonomy will be applied to enhance an existing ontology, for example, the Ophthalmology-related components in SNOMED CT. According to Ophthalmologists, the current components in SNOMED CT are not suitable for coding in an EHR, thus, they are not used in clinical practice as desired. In order to improve this situation, first, a weighted aggregate partial-area taxonomy will be created for each of the subhierarchies related to Ophthalmology, which will provide major subjects in the subhierarchy, serving as a "big picture." The weighted aggregate partial-area taxonomy will be expanded to derive a second level weighted aggregate partial-area taxonomy, showing more details than the first level. These first two levels will be modified to make each of the subhierarchies more suitable to express clinical terms used in practice. After that, each node in the second level partial-area taxonomy will be further expanded into a subhierarchy. This process will result in a more practical subhierarchy that will be used to code ophthalmological concepts in an EHR.

Due to NDF-RT's structure, the previously developed Abstraction Networks are not suitable to summarize most of NDF-RT's hierarchies. Hence, in this dissertation, the Ingredient Abstraction Network (IAbN) was designed to summarize the NDF-RT's *Chemical Ingredients* hierarchy. In fact, the idea of the IAbN could be applied to the other six hierarchies in NDF-RT, except for the *Pharmaceutical Preparations* hierarchy. In this dissertation, the IAbN was applied to discover missing Drug-Drug interactions (DDIs) from First Databank's DDI knowledge base.

In March 2018, NDF-RT was replaced by the Medication Reference Terminology (MED-RT), which refers to clinical drug concepts, chemical ingredient concepts, and assorted other concepts in external terminologies such as RxNorm, MeSH, and

SNOMED CT, and maintains the relationships between concepts. In future, a virtual *Chemical Ingredients* hierarchy will be built according to the relationships between clinical drug concepts and chemical ingredient concepts. A new Abstraction Network called *Virtual Ingredient Abstraction Network* (VIAbN) will be derived to summarize and visualize the virtual *Chemical Ingredients* hierarchy. The VIAbN could be applied to discover missing Adverse Drug Reactions in First Databank's knowledge base.

Chapter 5 presented eight Abstraction Network-based quality assurance (QA) studies on different hierarchies in an ontology or on different ontologies. The two characterizations of concepts with higher error rates, complex concepts and uncommonly modeled concepts, were successfully utilized to guide the QA studies. Such studies are needed in order to demonstrate the validity of the family-based quality assurance approach. A quality assurance technique is required to be successfully demonstrated for six out of six BioPortal ontologies in the same family to claim that it is applicable to the whole family.

The three studies in Chapter 5 on overlapping concepts within partial-area taxonomies for the NCIt's *Neoplasm* subhierarchy and *Gene* hierarchy and the SNOMED CT's *Infectious disease* hierarchy, combined with the previous three studies by the SABOC team on another three ontologies in the same family, made the overlapping concept QA technique applicable to the whole family with 76 ontologies in BioPortal. Furthermore, new QA techniques were introduced in this dissertation, i.e., utilizing the partial-area sub-taxonomy to look for additional overlapping concepts to increase the impact of the overlapping concept QA technology, utilizing group auditing and error propagation methods to save QA efforts. In future, the additional overlapping concept

technique will be investigated in other large hierarchies/ontologies with a limited number of overlapping concepts in their partial-area taxonomies.

The two studies on the NCIt's *Biological Process* hierarchy and the ChEBI ontology focused on concepts with many lateral relationship types in area taxonomies, which is another type of complex concepts, that is, laterally complex concepts. This QA technique was demonstrated successfully on the above two ontologies. In order to show that this technique is applicable to a whole family, in future, four additional studies on another four ontologies in this family will be conducted.

Furthermore, in this dissertation, the new Abstraction Network IAbN was applied to support the quality assurance of the *Chemical Ingredients* hierarchy in NDF-RT. The two new hypotheses for this topic focused on a new characterization of complex concepts, namely chemical ingredients in ingredient groups with multiple parent ingredient groups. In addition, the chemical ingredients with more parent ingredient groups were demonstrated as more complex, thus, more likely to have errors.

For uncommonly model concepts, the study on the NCIt's *Neoplasm* subhierarchy investigated concepts in small partial-areas within its partial-area taxonomy. This is the third study showing that the small partial-area QA technique is successful. The previous two studies by the SABOC team demonstrated this technique successfully on the NCIt's *Biological Process* hierarchy and the SNOMED CT's *Procedure* hierarchy. In future, three more QA studies on the ChEBI ontology, the SNOMED CT's *Specimen* hierarchy, and the NCIt's *Gene* hierarchy belonging to the same family, will be performed to meet the requirement of "six out of six" under the family-based QA framework.

There were 44.8% of concepts in the NCIt's *Biological Process* hierarchy having no lateral relationships, which was indicating that these concepts may miss lateral relationships. In the study on all these concepts, 45.2% were found missing lateral relationships. This study also confirmed the hypothesis that concepts in the area without any relationship are more likely to miss relationships than concepts in other areas within the area taxonomy. The complexity measure for the former type of concepts was explored as well, which will guide ontology curators to focus on auditing more complex concepts without relationships to achieve a better error yield, when it is impossible to perform quality assurance on all concepts without any relationship in the hierarchy/ontology. In future, this phenomenon will be investigated on other ontologies with a large number of concepts without lateral relationships, such as the NCIt's *Neoplasm* subhierarchy and the SNOMED CT's *Clinical finding* hierarchy.

REFERENCES

[1] Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. Journal of the American Medical Informatics Association : JAMIA. 2014;21(6):957-958.

[2] Perl Y, Geller J, Halper M, Ochs C, Zheng L, Kapusnik-Uner J. Introducing the Big Knowledge to Use (BK2U) challenge. Annals of the New York Academy of Sciences. 2017;1387(1):12-24.

[3] de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Studies in health technology and informatics. 2004;107(Pt 1):33-37.

[4] Halper M, Gu H, Perl Y, Ochs C. Abstraction networks for terminologies: Supporting management of "big knowledge". Artificial Intelligence in Medicine. 2015;64(1):1-16.

[5] Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. Journal of the American Medical Informatics Association : JAMIA. 2006;13(6):676-690.

[6] Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. Journal of biomedical informatics. 2007;40(5):561-581.

[7] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proceedings of the AMIA Symposium. 2001:662-666.

[8] Ochs C, He Z, Zheng L, Geller J, Perl Y, Hripcsak G, et al. Utilizing a structural meta-ontology for family-based quality assurance of the BioPortal ontologies. Journal of biomedical informatics. 2016;61:63-76.

[9] Zheng L, Perl Y, Elhanan G, Ochs C, Geller J, Halper M. Summarizing an ontology: a "Big Knowledge" coverage approach. Studies in health technology and informatics. 2017;245:978-982.

[10] Zheng L, Ochs C, Geller J, Liu H, Perl Y, De Coronado S. Multi-layer Big Knowledge visualization scheme for comprehending neoplasm ontology content. 2017 IEEE International Conference on Big Knowledge (ICBK). 2017:127-134.

[11] Ochs C, Zheng L, Gu H, Perl Y, Geller J, Kapusnik-Uner J, et al. Drug-drug interaction discovery using Abstraction Networks for "National Drug File - Reference Terminology" Chemical Ingredients. AMIA Annual Symposium Proceedings. 2015;2015:973-982.

[12] Zheng L, Yumak H, Chen L, Ochs C, Geller J, Kapusnik-Uner J, et al. Quality assurance of chemical ingredient classification for the National Drug File - Reference Terminology. Journal of biomedical informatics. 2017;73:30-42.

[13] Zheng L, Chen Y, Elhanan G, Perl Y, Geller J, Ochs C. Complex overlapping concepts: an effective auditing methodology for families of similarly structured BioPortal ontologies. Journal of biomedical informatics. 2018;83:135-149.

[14] Min H, Zheng L, Perl Y, Halper M, De Coronado S, Ochs C. Relating complexity and error rates of ontology concepts. more complex NCIt concepts have more errors. Methods of Information in Medicine. 2017;56(3):200-208.

[15] Zheng L, Min H, Chen Y, Xu J, Geller J, Perl Y. Auditing National Cancer Institute thesaurus neoplasm concepts in groups of high error concentration. Applied Ontology. 2017;12(2):113-130.

[16] Zheng L, Min H, Perl Y, Geller J. Discovering additional complex NCIt gene concepts with high error rate. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017:653-657.

[17] Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Briefings in bioinformatics. 2008;9(1):75-90.

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

[19] Smith B, Scheuermann RH. Ontologies for clinical and translational research: Introduction. Journal of biomedical informatics. 2011;44(1):3-7.

[20] Schulz S, Jansen L. Formal ontologies in biomedical knowledge representation. Yearbook of medical informatics. 2013;8:132-146.

[21] Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in bioinformatics. 2015;16(6):1069-1080.

[22] Shen F, Lee Y. Knowledge Discovery from Biomedical Ontologies in Cross Domains. PloS one. 2016;11(8):e0160005.

[23] SNOMED CT. Available from: https://www.nlm.nih.gov/healthit/snomedct/index.html. Accessed July 16, 2018.

[24] Millar J. The Need for a Global Language - SNOMED CT Introduction. Studies in health technology and informatics. 2016;225:683-685.

[25] SNOMED CT Basics. Available from: https://confluence.ihtsdotools.org/display/DOCSTART/4.+SNOMED+CT+Basics. Accessed July 16, 2018.

[26] National Drug File (NDF). Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/VANDF/. Accessed July 16, 2018.

[27] National Drug File – Reference Terminology (NDF-RT) Documentation February 2015 Version. Available from: https://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT%20Documentation.pdf. Accessed July 16, 2018.

[28] Medical Subject Headings (MeSH). Available from: http://www.ncbi.nlm.nih.gov/mesh. Accessed July 16, 2018.

[29] Chute CG, Carter JS, Tuttle MS, Haber M, Brown SH. Integrating pharmacokinetics knowledge into a drug ontology: as an extension to support pharmacogenomics. AMIA Annual Symposium proceedings. 2003:170-174.

[30] PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed/. Accessed July 16, 2018.

[31] Federal Medication Terminologies. Available from: https://www.cancer.gov/research/resources/terminology/fmt. Accessed July 16, 2018.

[32] National Drug File - Reference Terminology (NDF-RT). Available from: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/. Accessed July 16, 2018.

[33] National Drug File - Reference Terminology on BioPortal. Available from: http://bioportal.bioontology.org/ontologies/NDFRT. Accessed July 16, 2018.

[34] Carter JS, Brown SH, Erlbaum MS, Gregg W, Elkin PL, Speroff T, et al. Initializing the VA medication reference terminology using UMLS metathesaurus co-occurrences. Proceedings of the AMIA Symposium. 2002:116-120.

[35] Rosenbloom ST, Awad J, Speroff T, Elkin PL, Rothman R, Spickard A, 3rd, et al. Adequacy of representation of the National Drug File Reference Terminology Physiologic Effects reference hierarchy for commonly prescribed medications. AMIA Annual Symposium Proceedings. 2003:569-578.

[36] Carter JS, Brown SH, Bauer BA, Elkin PL, Erlbaum MS, Froehling DA, et al. Categorical information in pharmaceutical terminologies. AMIA Annual Symposium Proceedings. 2006:116-120.

[37] Zhu Q, Freimuth RR, Pathak J, Chute CG. PharmGKB Drug Data Normalization with NDF-RT. AMIA Joint Summits on Translational Science proceedings. 2013;2013:180.

[38] Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clinical pharmacology and therapeutics. 2012;92(4):414-417.

[39] Pathak J, Weiss LC, Durski MJ, Zhu Q, Freimuth RR, Chute CG. Integrating VA's NDF-RT drug terminology with PharmGKB: preliminary results. Pacific Symposium on Biocomputing. 2012:400-409.

[40] Pathak J, Murphy SP, Willaert BN, Kremers HM, Yawn BP, Rocca WA, et al. Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. AMIA Annual Symposium Proceedings. 2011;2011:1089-1098.

[41] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: Prescription for Electronic Drug Information Exchange. IT Professional. 2005;7(5):17-23.

[42] de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. The NCI Thesaurus quality assurance life cycle. Journal of biomedical informatics. 2009;42(3):530-539.

[43] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. Journal of biomedical informatics. 2005;38(2):114-129.

[44] Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics. 2007;40(1):30-43.

[45] Fragoso G, de Coronado S, Haber M, Hartel F, Wright L. Overview and utilization of the NCI thesaurus. Comparative and functional genomics. 2004;5(8):648-654.

[46] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic acids research. 2016;44(D1):D1214-1219.

[47] Hill DP, Adams N, Bada M, Batchelor C, Berardini TZ, Dietze H, et al. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics. 2013;14:513.

[48] Gao YF, Chen L, Huang GH, Zhang T, Feng KY, Li HP, et al. Prediction of drugs target groups based on ChEBI ontology. BioMed Research International. 2013;2013:132724.

[49] Consortium U. UniProt: a hub for protein information. Nucleic acids research. 2015;43(Database issue):D204-212.

[50] Lamurias A, Ferreira JD, Couto FM. Improving chemical entity recognition through h-index based semantic similarity. Journal of Cheminformatics. 2015;7(Suppl 1 Text mining for chemistry and the CHEMDNER track):S13.

[51] ChEBI user manual. Available from: http://www.ebi.ac.uk/chebi/userManualForward.do. Accessed September 12, 2017.

[52] Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. Nucleic acids research. 2013;41(Database issue):D456-463.

[53] ChEBI issue tracking system. Available from: https://github.com/ebi-chebi/ChEBI/issues. Accessed July 16, 2018.

[54] Structural Analysis of Biomedical Ontologies Center (SABOC). Available from: http://saboc.njit.edu. Accessed July 16, 2018.

[55] Ochs C, Perl Y, Halper M, Geller J, Lomax J. Quality assurance of the gene ontology using abstraction networks. Journal of bioinformatics and computational biology. 2016;14(3):1642001.

[56] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000;25(1):25-29.

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

[58] Ochs C, Geller J, Perl Y, Chen Y, Agrawal A, Case JT, et al. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. Journal of the American Medical Informatics Association : JAMIA. 2015;22(3):628-639.

[59] Ochs C, Agrawal A, Perl Y, Halper M, Tu SW, Carini S, et al. Deriving an abstraction network to support quality assurance in OCRe. AMIA Annual Symposium Proceedings. 2012;2012:681-689.

[60] He Z, Ochs C, Agrawal A, Perl Y, Zeginis D, Tarabanis K, et al. A family-based framework for supporting quality assurance of biomedical ontologies in BioPortal. AMIA Annual Symposium Proceedings. 2013;2013:581-590.

[61] Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. Journal of biomedical informatics. 2014;52:78-91.

[62] Zeginis D, Hasnain, A., Loutas, N., et al. A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web. 2014;5(2):127-142.

[63] Ochs C, He, Z., Perl, Y., et al. Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology. 2013 International Conference on Biomedical Ontology (ICBO). 2013:84-89.

[64] Arabandi S, Ogbuji C, Redline S, Chervin R, Boero J, Benca R, et al. Developing a Sleep Domain Ontology. Summit on Clinical Research Informatics. 2010.

[65] He Z, Ochs C, Soldatova L, Perl Y, Arabandi S, Geller J. Auditing redundant import in reuse of a top level ontology for the Drug Discovery Investigations Ontology. 2013 Vaccine and Drug Ontology Studies (VDOS 2013) international workshop.

[66] Qi D, King RD, Hopkins AL, Bickerton GR, Soldatova LN. An ontology for description of drug discovery investigations. Journal of integrative bioinformatics. 2010;7(3).

[67] Ochs C, Geller J, Perl Y, Musen MA. A unified software framework for deriving, visualizing, and exploring abstraction networks for ontologies. Journal of biomedical informatics. 2016;62:90-105.

[68] Cui L. COHeRE: Cross-Ontology Hierarchical Relation Examination for Ontology Quality Assurance. AMIA Annual Symposium Proceedings. 2015;2015:456-465.

[69] Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naive vs. formal. AMIA Annual Symposium Proceedings. 2005:550-554.

[70] Gu H, Chen Y, He Z, Halper M, Chen L. Quality Assurance of UMLS Semantic Type Assignments Using SNOMED CT Hierarchies. Methods of information in medicine. 2016;55(2):158-165.

[71] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. Journal of the American Medical Informatics Association : JAMIA. 1998;5(1):41-51.

[72] Mougin F, Grabar N. Auditing the multiply-related concepts within the UMLS. Journal of the American Medical Informatics Association : JAMIA. 2014;21(e2):e185-193.

[73] Xing G, Zhang GQ, Cui L. FEDRR: fast, exhaustive detection of redundant hierarchical relations for quality improvement of large biomedical ontologies. BioData mining. 2016;9:31.

[74] Bodenreider O. Identifying Missing Hierarchical Relations in SNOMED CT from Logical Definitions Based on the Lexical Features of Concept Names. 2016 International Conference on Biomedical Ontology (ICBO).

[75] Dentler K, Cornet R. Intra-axiom redundancies in SNOMED CT. Artificial intelligence in medicine. 2015;65(1):29-34.

[76] Agrawal A, Elhanan G. Contrasting lexical similarity and formal definitions in SNOMED CT: consistency and implications. Journal of biomedical informatics. 2014;47:192-198.

[77] Jiang G, Chute CG. Auditing the semantic completeness of SNOMED CT using formal concept analysis. Journal of the American Medical Informatics Association : JAMIA. 2009;16(1):89-102.

[78] Mougin F. Identifying redundant and missing relations in the gene ontology. Studies in health technology and informatics. 2015;210:195-199.

[79] Verspoor K, Dvorkin D, Cohen KB, Hunter L. Ontology quality assurance through analysis of term transformations. Bioinformatics. 2009;25(12):i77-84.

[80] Ceusters W. Applying evolutionary terminology auditing to the Gene Ontology. Journal of biomedical informatics. 2009;42(3):518-529.

[81] Kohler J, Munn K, Ruegg A, Skusa A, Smith B. Quality control for terms and definitions in ontologies and taxonomies. BMC Bioinformatics. 2006;7:212.

[82] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods of information in medicine. 2005;44(4):498-507.

[83] Cohen B, Oren M, Min H, Perl Y, Halper M. Automated comparative auditing of NCIT genomic roles using NCBI. Journal of biomedical informatics. 2008;41(6):904-913.

[84] Min H, Cohen B, Halper M, Oren M, Perl Y. Detecting role errors in the gene hierarchy of the NCI Thesaurus. Cancer informatics. 2008;6:293-313.

[85] Mougin F, Bodenreider O. Auditing the NCI thesaurus with semantic web technologies. AMIA Annual Symposium Proceedings. 2008:500-504.

[86] Wei D, Bodenreider O. Using the abstraction network in complement to description logics for quality assurance in biomedical terminologies - a case study in SNOMED CT. Studies in health technology and informatics. 2010;160(Pt 2):1070-1074.

[87] Rogers JE. Quality assurance of medical ontologies. Methods of information in medicine. 2006;45(3):267-274.

[88] Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of biomedical informatics. 2009;42(3):413-425.

[89] Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. Journal of biomedical informatics. 2009;42(3):407-411.

[90] Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, et al. Analysis of error concentrations in SNOMED. AMIA Annual Symposium Proceedings. 2007:314-318.

[91] Wang Y, Halper M, Wei D, Gu H, Perl Y, Xu J, et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. Journal of biomedical informatics. 2012;45(1):1-14.

[92] Ochs C, Geller J, Perl Y, Chen Y, Xu J, Min H, et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. Journal of the American Medical Informatics Association : JAMIA. 2015;22(3):507-518.

[93] Elhanan G, Ochs C, Mejino JLV, Jr., Liu H, Mungall CJ, Perl Y. From SNOMED CT to Uberon: Transferability of evaluation methodology between similarly structured ontologies. Artificial intelligence in medicine. 2017;79:9-14.

[94] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome biology. 2012;13(1):R5.

[95] Patel VL, Kaufman DR, Arocha J. Conceptual change in the biomedical and health sciences domain. Advances in instructional psychology. 5 ed: Lawrence Erlbaum Associates; 2000. p. 329-392.

[96] Ochs C, Perl Y, Geller J, Musen M. Using aggregate taxonomies to summarize SNOMED CT evolution. 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2015:1008 - 1015.

[97] Blumenthal DK, Garrison JC. Chapter 3: Pharmacodynamics: Molecular Mechanisms of Drug Action. Goodman & Gilman's: The Pharmacological Basis of Therapeutics. 12 ed: The McGraw-Hill Companies, Inc; 2011.

[98] Manning CD, Raghavan P, Schütze H. Evaluation in information retrieval. Introduction to Information Retrieval: Cambridge University Press; 2008. p. 151-175.

[99] Katifori A, Halatsis C, Lepouras G, Vassilakis C, Giannopoulou E. Ontology visualization methods—a survey. ACM Computing Surveys. 2007;39(4).

[100] Ghoniem M, Fekete J-D, Castagliola P. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. Information Visualization archive. 2005;4(2):114 - 135.

[101] Ware C, Mitchell P. Visualizing graphs in three dimensions. ACM Transactions on Applied Perception (TAP). 2008;5(1).

[102] Models of working memory. Mechanisms of active maintenance and executive control. Cambridge University Press; 1999.

[103] Miller GA. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychological review. 1956;63(2):81-97.

[104] Cowan N. The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? Current directions in psychological science. 2010;19(1):51-57.

[105] Drug interactions of clinical importance. In: Davies DM, Ferner RE, Glanville HD, editors. Davies's textbook of adverse drug reactions. 5 ed. London: Chapman & Hall Medical; 1998. p. 888-912.

[106] Kuhlmann J, Muck W. Clinical-pharmacological strategies to assess drug interaction potential during drug development. Drug safety. 2001;24(10):715-725.

[107] Grymonpre RE, Mitenko PA, Sitar DS, Aoki FY, Montgomery PR. Drug-associated hospital admissions in older medical patients. Journal of the American Geriatrics Society. 1988;36(12):1092-1098.

[108] Jankel CA, Speedie SM. Detecting drug interactions: a review of the literature. DICP: the annals of pharmacotherapy. 1990;24(10):982-989.

[109] Scripture CD, Figg WD. Drug interactions in cancer therapy. Nature reviews Cancer. 2006;6(7):546-558.

[110] Zhan C, Correa-de-Araujo R, Bierman AS, Sangl J, Miller MR, Wickizer SW, et al. Suboptimal prescribing in elderly outpatients: potentially harmful drug-drug and drug-disease combinations. Journal of the American Geriatrics Society. 2005;53(2):262-267.

[111] Malone DC, Hutchins DS, Haupert H, Hansten P, Duncan B, Van Bergen RC, et al. Assessment of potential drug-drug interactions with a prescription claims database. American journal of health-system pharmacy. 2005;62(19):1983-1991.

[112] Janchawee B, Owatranporn T, Mahatthanatrakul W, Chongsuvivatwong V. Clinical drug interactions in outpatients of a university hospital in Thailand. Journal of clinical pharmacy and therapeutics. 2005;30(6):583-590.

[113] Janchawee B, Wongpoowarak W, Owatranporn T, Chongsuvivatwong V. Pharmacoepidemiologic study of potential drug interactions in outpatients of a university hospital in Thailand. Journal of clinical pharmacy and therapeutics. 2005;30(1):13-20.

[114] Aparasu R, Baer R, Aparasu A. Clinically important potential drug-drug interactions in outpatient settings. Research in social & administrative pharmacy. 2007;3(4):426-437.

[115] Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, et al. VA National Drug File Reference Terminology: a cross-institutional content coverage study. Studies in health technology and informatics. 2004;107(Pt 1):477-481.

[116] FDB Drug-Disease Contraindications Module. Available from: http://www.fdbhealth.com/fdb-medknowledge-clinical-modules/drug-disease-contraindications-module/. Accessed July 16, 2018.

[117] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic acids research. 2011;39(Web Server issue):W541-545.

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

[119] NCI Thesaurus property definitions. Available from: https://evs.nci.nih.gov/ftp1/ThesaurusSemantics/Properties.pdf. Accessed July 16, 2018.

[120] Ochs C, Perl Y, Geller J, Halper M, Gu H, Chen Y, et al. Scalability of abstraction-network-based quality assurance to large SNOMED hierarchies. AMIA Annual Symposium Proceedings. 2013;2013:1071-1080.

[121] Ochs C, Case JT, Perl Y. Tracking the Remodeling of SNOMED CT's Bacterial Infectious Diseases. AMIA Annual Symposium Proceedings. 2016;2016:974-983.

[122] MJ L, C B. Fast classification in Protégé: Snorocket as an OWL 2 EL reasoner. The 6th Australasian Ontology Workshop (IAOA'10): Conferences in Research and Practice in Information Technology. 2010. p. 45-49.

[123] Ceusters W, Bona JP. Analyzing SNOMED CT's Historical Data: Pitfalls and Possibilities. AMIA Annual Symposium Proceedings. 2016;2016:361-370.

[124] Zhang G-Q, Huang Y, Cui L. Can SNOMED CT Changes Be Used as a Surrogate Standard for Evaluating the Performance of Its Auditing Methods? AMIA Annual Symposium proceedings. 2017:1886-1895.

[125] Ochs C, Case JT, Perl Y. Analyzing structural changes in SNOMED CT's Bacterial infectious diseases using a visual semantic delta. Journal of biomedical informatics. 2017;67:101-116.

[126] Goodrich MT, Tamassia R. Divide-and-Conquer. Algorithm Design. 2 ed. New York: John Wiley & Sons, Inc; 2001. p. 263-273.

[127] Morrey CP, Geller J, Halper M, Perl Y. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. Journal of biomedical informatics. 2009;42(3):468-489.

[128] NCI Thesaurus. Available from: https://nciterms.nci.nih.gov/ncitbrowser/. Accessed July 16, 2018.

[129] Geller J, Ochs C, Perl Y, Xu J. New abstraction networks and a new visualization tool in support of auditing the SNOMED CT content. AMIA Annual Symposium Proceedings. 2012;2012:237-246.

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

[131] Dalkey NC, Helmer-Hirschberg O. An experimental application of the Delphi method to the use of experts. Management Science. 1963;9(3):458-467.

[132] He Z, Morrey CP, Perl Y, Elhanan G, Chen L, Chen Y, et al. Sculpting the UMLS Refined Semantic Network. Online journal of public health informatics. 2014;6(2):e181.

[133] Goodrich MT, Tamassia R. Data Structures and Algorithms in Java. Hoboken, New Jersey: Wiley Publishing; 2014.