Spring 5-31-2015

Designing novel abstraction networks for ontology summarization and quality assurance

Christopher Ochs

New Jersey Institute of Technology

ABSTRACT

DESIGNING NOVEL ABSTRACTION NETWORKS FOR ONTOLOGY SUMMARIZATION AND QUALITY ASSURANCE

by
Christopher Ochs

Biomedical ontologies are complex knowledge representation systems. Biomedical ontologies support interdisciplinary research, interoperability of medical systems, and Electronic Healthcare Record (EHR) encoding. Ontologies represent knowledge using concepts (entities) linked by relationships. Ontologies may contain hundreds of thousands of concepts and millions of relationships. For users, the size and complexity of ontologies make it difficult to comprehend “the big picture” of an ontology’s content. For ontology editors, size and complexity make it difficult to uncover errors and inconsistencies. Errors in an ontology will ultimately affect applications that utilize the ontology.

In prior studies *abstraction networks* (AbNs) were developed to provide a compact summary of an ontology’s content and structure. AbNs have been shown to successfully support ontology summarization and quality assurance (QA), e.g., for SNOMED CT and NCIt. Despite the success of these previous studies, several major, unaddressed issues affect the applicability and usability of AbNs. This thesis is broken into five major parts, each addressing one issue.

The first part of this dissertation addresses the scalability of AbN-based QA techniques to large SNOMED CT hierarchies. Previous studies focused on relatively small hierarchies. The QA techniques developed for these small hierarchies do not scale to large hierarchies, e.g., *Procedure* and *Clinical finding*. A new type of AbN, called a *subtaxonomy*, is introduced to address this problem. Subtaxonomies summarize a subset of an ontology's content. Several types of subtaxonomies and subtaxonomy-based QA studies are discussed.

The second part of this dissertation addresses the need for summarization and QA methods for the twelve SNOMED CT hierarchies with no lateral relationships. Previously developed SNOMED CT AbN derivation methodologies, which require lateral relationships, cannot be applied to these hierarchies. The Tribal Abstraction Network (TAN) is a new type of AbN derived using only hierarchical relationships. A TAN-based QA methodology is introduced and the results of a QA review of the *Observable entity* hierarchy are reported.

The third part focuses on the development of generic AbN derivation methods that are applicable to groups of structurally similar ontologies, e.g., those developed in the Web Ontology Language (OWL) format. Previously, AbN derivation techniques were applicable to only a single ontology at a time. AbNs that are applicable to many OWL ontologies are introduced, a preliminary study on OWL AbN granularity is reported on, and the results of several QA studies are presented.

The fourth part describes Diff Abstraction Networks, which summarize and visualize the structural differences between two ontology releases. Diff Area Taxonomy and Diff Partial-area Taxonomy derivation methodologies are introduced and Diff Partial-area taxonomies are derived for three OWL ontologies. The Diff Abstraction Network approach is compared to the traditional ontology diff approach.

Lastly, tools for deriving and visualizing AbNs are described. The Biomedical Layout Utility Framework is introduced to support the automatic creation, visualization, and exploration of abstraction networks for SNOMED CT and OWL ontologies.

DESIGNING NOVEL ABSTRACTION NETWORKS FOR ONTOLOGY SUMMARIZATION AND QUALITY ASSURANCE

by
Christopher Ochs

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

Department of Computer Science

May 2015

.

APPROVAL PAGE

DESIGNING NOVEL ABSTRACTION NETWORKS FOR ONTOLOGY SUMMARIZATION AND QUALITY ASSURANCE

Christopher Ochs

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

Dr. James Geller, Committee Member Date
Professor and Chair of Computer Science, NJIT

Dr. Michael Halper, Committee Member Date
Professor of Information Systems and IT Program Director, NJIT

Dr. Huanying Gu, Committee Member Date
Associate Professor of Computer Science, NYIT

Dr. Gai Elhanan, Committee Member Date
CMIO, Halfpenny Technologies

Dr. James Cimino, Committee Member Date
Director, Informatics Institute, University of Alabama at Birmingham

Dr. Marek Rusinkiewicz, Committee Member Date
Professor and Dean, College of Computing Sciences, NJIT

BIOGRAPHICAL SKETCH

Author:	Christopher Ochs
Degree:	Doctor of Philosophy
Date:	May 2015

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science
  New Jersey Institute of Technology, Newark, NJ, 2015

- Master of Science in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2011

- Bachelor of Science in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2010

- Associates of Science in Computer Science, Physics, and Mathematics,
  Brookdale Community College, Lincroft, NJ, 2008

Major: Computer Science

Publications

Published Journal Papers

Halper, M., Gu, H., Perl, Y., Ochs, C. Abstraction Networks for Terminologies: Supporting Management of "Big Knowledge". Artificial Intelligence in Medicine. 2015. In press.

Ochs, C., Geller, J., Perl, Y., et al. A Tribal Abstraction Network for SNOMED CT Hierarchies without Attribute Relationships. Journal of the American Medical Informatics Association (J Am Med Inform Assoc). 2014. Epub: 20 October 2014.

Ochs, C., Geller, J., Perl, Y., et al. Scalable Quality Assurance for Large SNOMED CT Hierarchies Using Subject-based Subtaxonomies. J Am Med Inform Assoc. 2014. Epub: 21 October 2014.

Ochs, C., Geller, J., Perl, Y. A Relationship-centric Hybrid Interface for Browsing and Auditing the UMLS. Journal of Integrated Design and Process Science. 2011;15.4:3-25.

Submitted Journal Papers

Ochs, C., Perl, Y., Geller, J., et al. Summarizing and Visualizing Structural Changes during the Evolution of Biomedical Ontologies Using a Diff Abstraction Network. Journal of Biomedical Informatics (J Biomed Inform). 2015. Revision submitted for review.

Published Conference Papers

Ochs, C., Perl, Y., Halper, M., et al. Gene Ontology Summarization to Support Visualization and Quality Assurance. 7th International Conference on Bioinformatics and Computational Biology (BICoB). 2015. *Best paper finalist*.

Ochs, C., Perl, Y., Geller, J., et al. Scalability of Abstraction-Network-Based Quality Assurance to Large SNOMED Hierarchies. American Medical Informatics Association Annual Symposium Proceedings (AMIA Annu Symp Proc) 2013:1071-80.

He, Z., Ochs, C., Agrawal, A., et al. A Family-Based Framework for Supporting Quality Assurance of Biomedical Ontologies in BioPortal. Proc AMIA Annu Symp Proc. 2013:581-90.

Ochs, C., He, Z., Perl, Y., et al. Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology. Proceedings of the 4th International Conference on Biomedical Ontology (ICBO). 2013:84-9.

He, Z., Ochs, C., Soldatova, L., et al. Auditing Redundant Import in Reuse of a Top Level Ontology for the Drug Discovery Investigations Ontology. Vaccine and Drug Ontology Studies Workshop (VDOS). 2013.

Geller, J., Ochs, C., Perl, Y., et al. New Abstraction Networks and a New Visualization Tool in Support of Auditing the SNOMED CT Content. AMIA Annu Symp Proc. 2012:237-46.

Ochs, C., Agrawal, A., Perl, Y., et al. Deriving an Abstraction Network to Support Quality Assurance in OCRe. AMIA Annu Symp Proc. 2012:681-9.

Ochs, C., Tian, T., Geller, J., Chun, S.A. Google Knows Who is Famous Today - Building an Ontology From Search Engine Knowledge and DBpedia. Fifth IEEE International Conference on Semantic Computing (ICSC). 2011: 320–7

Published Peer-reviewed Abstracts

Ochs, C. BLUSNO Tool for SNOMED CT Visualization and QA Support. IHTSDO Implementation Showcase. 2013.

Ochs, C., Perl, Y., Geller, J. BLUSNO: A System for Orientation, Visualization, and Quality Assurance of SNOMED CT Using Abstraction Networks. Proceedings of the International Conference on Biomedical Ontology. 2013:128-9.

Posters

Geller, J., Ochs, C., He, Z., Perl, Y. A structural meta-ontology for the BioPortal ontologies. Bio-Ontologies. 2013.

Presentations

Using Computer Science to Solve Problems. Invited talk at Brookdale Community College Computer Science Club, Lincroft, NJ. March 24, 2015.

Gene Ontology Summarization to Support Visualization and Quality Assurance. BICoB 2015. Honolulu, HI. March 10, 2015.

A Brief Guide to a Career in Computer Science. Invited talk at Brookdale Community College Computer Science Club, Lincroft, NJ. September 23, 2014.

Using Ontologies to Disambiguate Web Search Queries. Invited talk at Manhattan College, Bronx, NY, NY. February 24, 2014.

Scalability of Abstraction-Network-Based Quality Assurance to Large SNOMED Hierarchies. AMIA Annual Symposium, Washington, D.C., November 17, 2013.

BLUSNO Tool for SNOMED CT Visualization and QA Support. 2013 IHTSDO Implementation Showcase, Washington, D.C., October 10, 2013.

Choosing the Granularity of Abstraction Networks for Orientation and Quality Assurance of the Sleep Domain Ontology. 4th International Conference on Biomedical Ontology (ICBO 2013), July 9, 2013.

Deriving an Abstraction Network to Support Quality Assurance in OCRe. AMIA Annual Symposium, Chicago, IL, November 6, 2012.

BLUSNO - Innovative Tools for Visualizing & Auditing SNOMED-CT. Co-presenter with James Geller. Invited talk at Department of Biomedical Informatics, Columbia University, New York City, NY, October 27, 2011.

Google Knows Who is Famous Today - Building an Ontology From Search Engine Knowledge and DBpedia. IEEE International Conference on Semantic Computing (ICSC 2011), Stanford University, Palo Alto, CA, September 20, 2011.

I dedicate this dissertation to Karina Aliaga

ACKNOWLEDGMENT

I thank Drs. Yehoshua Perl and James Geller for their guidance. Their expertise and continuous feedback made the work in this dissertation possible. I thank all of my committee members for their time and effort. I would also like to thank Dr. Michael Halper for his assistance on many of the research projects described in this dissertation.

I thank Dr. Ankur Agrawal, Dr. Zhe He, David Daudelin, and the many other collaborators for all of their contributions to the work in this thesis. I also thank the many student developers who contributed their time and skill to the software projects that made this thesis possible.

Lastly, I thank my parents, my girlfriend, and all of my friends for their support and encouragement over the last four years.

TABLE OF CONTENTS

Chapter **Page**

1 INTRODUCTION……........................……………….………………………… 1

1.1 Motivation ………………………………………………………………... 1

1.2 Dissertation Overview ……………………………………………………. 4

2 BACKGROUND……………………………….…………………….………... 5

2.1 Biomedical Ontologies ...…………………………………………………. 5

2.1.1 SNOMED CT ………………………………………………………... 6

2.1.2 Web Ontology Language (OWL) ..……………………………….……….. 9

2.1.3 NCBO BioPortal ……………………………………………………...… 11

2.2 Abstraction Networks ……………………………………………….……. 13

2.2.1 Previously Developed Abstraction Networks ……………………….....… 16

2.2.2 Abstraction Networks for Quality Assurance ...………………………… 26

2.3 Additional Related Work ………………………………………………… 28

2.3.1 Ontology Summarization ……………………………………………….. 28

2.3.2 Ontology Quality Assurance ……………………………………………. 30

2.3.3 Ontology Software ……………………………………………………… 33

2.3.4 Ontology Visualization …………………………………………………. 34

2.3.5 Ontology Diff …………………………………………………………… 35

3 DESIGNING NOVEL ABSTRACTION NETWORKS…………………………… 38

3.1 Subtaxonomies for Large SNOMED CT Hierarchies ...……………………….. 38

3.1.1 Relationship-constrained Partial-area Subtaxonomy ...…………...…….. 40

TABLE OF CONTENTS (Continued)

Chapter **Page**

3.1.2 Root-constrained Partial-area Subtaxonomy ……………………………… 52

3.1.3 Subject Subtaxonomy ……………………………………………………... 54

3.1.4 Focus Subtaxonomy ……………………………………………………….. 74

3.2 Tribal Abstraction Network ……………………………………………………… 76

3.2.1 Derivation …………………………………………………………………. 77

3.2.2 In Support of Quality Assurance ……………………………………….. 84

3.2.3 Limitations ………………………………………………………………… 99

3.2.4 Additional TAN Applications …………………………………………... 100

3.2.5 Discussion …………………………………………………………………. 102

3.3 Abstraction Networks for OWL Ontologies …………………………………... 103

3.3.1 Domain-defined Derivation Methodology ……………………………... 105

3.3.2 Restriction-defined Derivation Methodology …………………………... 112

3.3.3 Granularity of OWL Abstraction Networks …………………………… 115

3.3.4 In Support of Quality Assurance ……………………………………….. 119

3.4 Diff Abstraction Networks ……………………………………………………….. 137

3.4.1 Derivation ………………………….. 140

3.4.2 DPAT-based Quality Assurance Methodology ………………………… 151

3.4.3 Application of Diff Partial-area Taxonomies …………………………... 152

3.4.4 Comparison to Traditional Ontology Diff Output ……………………… 163

3.4.5 Discussion …………………………………………………………………. 166

3.5 Abstraction Network Tools …...……………………………………………….. 169

TABLE OF CONTENTS
(Continued)

Chapter **Page**

3.5.1 Biomedical Layout Utility for SNOMED CT (BLUSNO) …………….. 171

3.5.2 Biomedical Layout Utility for OWL (BLUOWL) ……………………… 183

4 FUTURE WORK ……………………………………………………...…....... 191

4.1 Subtaxonomies for Large SNOMED CT Hierarchies …………………………. 191

4.2 Tribal Abstraction Network …………………………………………………… 194

4.3 Abstraction Networks for OWL Ontologies …………………………………... 198

4.4 Diff Abstraction Networks …………………………………………………….. 201

4.5 Abstraction Network Tools ……………………………………………………. 203

5 CONCLUSIONS …………………………………………………………...… 205

REFERENCES ……………………………………………………………….... 209

LIST OF TABLES

Table **Page**

3.1 Taxonomy Metrics for Seven of SNOMED CT's Hierarchies (Jan 2013 release) 39

3.2 Four Examples of Inconsistencies Identified by the Domain Expert in her Quality Assurance Review of the Relationship Subtaxonomy ...…………..…… 50

3.3 Summary of Errors Found in the Audit of the Sample Partial-areas of the Relationship Subtaxonomy ………………………………………………….... 51

3.4 Summary of Erroneous Concepts Broken Down into Small and Large Partial-areas ………………………………………………………………………... 51

3.5 Subject Subtaxonomy Metrics for the Ten Leading Causes of Death in the US ... 60

3.6 Errors Found in *Bleeding* Subject Subtaxonomy by Error Type ………………... 66

3.7 Auditing Results for Overlapping Concepts and Non-overlapping Concepts in Small and Large Disjoint Partial-areas for Several Boundary Points Between "Small" and "Large" …………………………………………………………….. 67

3.8 Auditing Results Broken Down by Disjoint Partial-area Taxonomy Level …….. 69

3.9 Five Examples of Errors Reported by the Auditors ……………………………... 69

3.10 Recommended Order of Auditing in the *Bleeding* Subject Subtaxonomy ……… 73

3.11 A Breakdown by Hierarchy of Active SNOMED CT Concepts with Multiple Parents …………………………………………………………………………… 77

3.12 Metrics for the Three Levels of the *Observable entity* Hierarchy's Band and Cluster Tribal Abstraction Networks …………………………………………… 87

3.13 The Distribution of Concepts, Errors, and Error Rates Among the Six Bins …… 91

3.14 An Analysis of Bins 1, 2 ("Large Clusters") Broken Down by Level ………….. 92

3.15 An Analysis of Bins 3-6 ("small clusters") Broken Down by Level …………… 92

3.16 A Sample of Five Errors Taken from the Auditing Results ……………………. 92

3.17 Lexical Containment Audit Results ……………………………………………... 95

3.18 Three Examples of Errors Found in the Lexical Containment Pairs ……………. 95

LIST OF TABLES
(Continued)

Table **Page**

3.19 Auditing Results for Number of Parents Study …………………………………. 96

3.20 Auditing Results for Fully Specified Name Length Study ……………………… 97

3.21 Precision, Sensitivity, and Specificity for the TAN and Lexical Containment Methodologies ……………………………………………………………………… 98

3.22 Structure of GO’s Biological process Taxonomy by Levels ……………………. 129

3.23 Overlapping Terms in {*part of*} by Overlap Degree ……………………………. 130

3.24 Quality Assurance Review Results According to Degree of Overlap ………….. 135

3.25 Three Examples of Overlapping Term Errors …………………………………... 135

3.26 Summary of the BLU Framework’s Subcomponents …………………………… 170

LIST OF FIGURES

Figure **Page**

2.1 An example of SNOMED CT's structure for the concept *Bacterial pneumonia*. Concepts are shown as labeled boxes, IS-A relationships are shown as thin blue arrows directed upwards. Dashed blue arrows represent a sequence of IS-A relationships. Attribute relationships are shown with thick labeled arrows ………………………………………………………………………………. 7

2.2 The concepts and relationships needed to define the concept *Hematoma of pinna* in SNOMED CT. Concepts are shown as labeled grey boxes, thick black edges represent IS-A relationships, while thin labeled edges represent attribute relationships ………………………………………………………………………... 8

2.3 The BioPortal user interface. The Ontology of Clinical Research (OCRe) was selected from the list of available ontologies ………………………………………. 12

2.4 The hierarchical relationships between the 4,503 concepts in the *Physical object* hierarchy of SNOMED CT ………………………………………………………… 14

2.5 The concept *Hematoma of pinna* as viewed using the CliniClue SNOMED CT concept browser ……………………………………………………………………. 15

2.6 The general process of deriving an abstraction network for an ontology or terminology ………………………………………………………………………… 16

2.7 (a) A sample of 17 concepts taken from the *Specimen* hierarchy. (b) The area taxonomy for the concepts in (a). (c) The partial-area taxonomy for the concepts in (a). Taxonomic elements have been labeled with red text and arrows 18

2.8 The complete area taxonomy for the *Specimen* hierarchy of SNOMED CT ……. 20

2.9 The partial-area taxonomy for the *Specimen* hierarchy. Child-of links between partial-areas are hidden due to space limitations ……………………………...... 21

2.10 The {*Identity, Substance*} area in SNOMED CT's *Specimen* hierarchy partitioned into inheritance regions ……………………………………….…..... 22

2.11 A hierarchy of overlapping concepts within the *Specimen* hierarchy's {*Substance*} area. Partial-area roots are singly colored. Overlapping roots are multicolored according to the partial-areas they overlap between ……………… 24

2.12 The disjoint partial-area taxonomy for the example of overlapping concepts in Figure 2.11 ……………………………………………………………………. 25

2.13 An example of an ontology diff taken from Protégé's "Compare Ontologies" tool, with the modified object property *duration* selected ………………………. 36

LIST OF FIGURES
(Continued)

Figure **Page**

3.1 An example of deriving a relationship area subtaxonomy for a theoretical hierarchy with four relationships R_1-R_4 with $R' = \{R_1, R_2, R_4\}$ …………………. 41

3.2 The relationship partial-area subtaxonomy for the *Specimen* hierarchy with *R'*={*Morphology, Substance, Identity*} ……………………………………….. 42

3.3 A small portion (69 areas) of the ten levels of the Procedure hierarchy's area taxonomy ……………………………………………………………………… 44

3.4 Area sub-taxonomy with respect to the three relationships using access device, procedure site – direct, and method …………………………………………... 45

3.5 Levels 0 and 1 and part of Level 2 of the partial-area sub-taxonomy with respect to the three selected relationships ……………………………………….. 47

3.6 The only third-level area of the sub-taxonomy with respect to the relationships *Method, Procedure site – direct,* and *Using access device* ……………………… 48

3.7 The process of deriving a root-constrained partial-area subtaxonomy. Partial-area P_2 was chosen as the root …………………………………………………... 53

3.8 An example of a root-constrained partial-area subtaxonomy for the *Specimen* hierarchy, with *Lesion sample* selected as the root partial-area. Child-of links are hidden for readability ……………………………………………………... 54

3.9 Top five (out of six) levels of the *Bleeding* subject subtaxonomy. A total of 932 bleeding-related concepts are summarized by 199 partial-areas in 42 areas. Over half (56%=522/932) of the concepts summarized by this subtaxonomy are in {*Associated morphology, Finding site*} ………………………………………. 58

3.10 The *Cancer* subject subtaxonomy. The Cancer subject subtaxonomy summarizes 3,531 concepts by 125 partial-areas in 19 areas. The 64 partial-areas that do not appear in the complete *Clinical finding* taxonomy are highlighted in yellow ……………………………………………………………. 59

3.11 An excerpt of 23 disjoint partial-areas from the disjoint partial-area subtaxonomy derived for the concepts in {*Associated morphology, Finding by site*} …………………………………………………………………………… 61

3.12 The error rates for small overlapping disjoint partial-areas for the various boundary points between small and large ……………………………………….. 68

3.13 The *Pneumonia* focus subtaxonomy …………………………………………….. 75

LIST OF FIGURES
(Continued)

Figure **Page**

3.14 An excerpt of 20 concepts from the *Observable entity* hierarchy with abbreviated tribal names in braces …………………………………………… 79

3.15 The concepts from Figure 3.14 grouped by common tribal sets ………………… 80

3.16 The band TAN derived from Figure 3.10. Each box represents a band. Child-of links are represented using arrows between bands ……………………………… 82

3.17 The cluster TAN derived from Figure 3.15. Child-of links are represented by arrows between clusters ………………………………………………………. 84

3.18 The Band Tribal Abstraction Network for the *Observable entity* hierarchy. Levels are organized into rows due to space limitations. Some child-of edges are hidden for readability ………………………………………………………. 88

3.19 The cluster tribal abstraction network for *Observable entity*. Child-of edges are hidden for readability. Each level is organized into several rows due to space limitations. Level 1 (not shown) is the same as in Figure 3.13 …………………. 89

3.20 A portion of OCRe's *Entity* hierarchy in an indented format with object properties introduced. The number after "+" indicates the number of inherited properties ………………………………………………………………… 108

3.21 The grouping of classes from Figure 3.20 into areas based on each class' set of object properties. Edges are *child-of* links between areas ……………………… 110

3.22 Partial-areas derived for each area in Figure 3.21. The targets of *child-of* links within partial-areas are indicated after "*CHILD OF*" …………………………… 110

3.23 Partial-area taxonomy for the subset of classes from OCRe's *Entity* hierarchy in Figure 3.20……………………………………………………………….…. 111

3.24 The restriction-defined partial-area taxonomy for the Sleep Domain Ontology's *Entity* hierarchy. Levels have been organized into rows and child-of edges are hidden for readability ………………………………………………………. 114

3.25 Domain-defined partial-area taxonomy for the SDO's *Entity* hierarchy ………... 116

3.26 The (domain or restriction)-defined partial-area taxonomy for the Sleep Domain Ontology's *Entity* hierarchy. The {*part of*} and {*has part*} areas are not shown .. 118

3.27 Complete partial-area taxonomy for OCRe's *Entity* hierarchy prior to the auditing efforts …………………………………………………………... 120

LIST OF FIGURES
(Continued)

Figure **Page**

3.28 Partial-area taxonomy for OCRe's *Entity* hierarchy, revised after an audit …….. 122

3.29 Excerpt of GO's *Biological process* partial-area taxonomy. Due to space limitations, certain areas are hidden and only a subset of partial-areas is shown for {*part of*}. Child-of's between partial-areas are hidden to enable readability. The number of terms in each partial-area is shown in parenthesis ……………… 128

3.30 A subhierarchy of 18 classes taken from OCRe Version 244, as shown in Protégé ……………………………………………………………………... 139

3.31 The subhierarchy from Figure 3.30 after several editing operations have been applied to the classes. This excerpt is from OCRe Version 258 ………………… 139

3.32 The excerpt of 18 classes from Figure 3.30 shown as a diagram, using bubbles to identify sets of classes with the same object properties ………………………. 142

3.33 The excerpt of classes, after corrections (from Release 258) corresponding to the excerpt of Figure 3.31, shown as a diagram ………………………………… 143

3.34 The visualization of the diff area taxonomy between the ontology excerpts in Figure 3.32 and Figure 3.33. The diff areas are organized into color coded levels according to the number of their object properties. The level numbers appear at the left edge of the figure ……………………………………………………... 146

3.35 The visualization of the DPAT between Figure 3.32 and Figure 3.33 ………….. 150

3.36 Color-coded text output for the DPAT between Figure 3.32 and Figure 3.33….. 150

3.37 The complete diff partial-area taxonomy for OCRe …………………………….. 154

3.38 The Sleep Domain Ontology's diff partial-area taxonomy. An excerpt of the child-of links between added/removed diff partial-areas is shown ……………… 156

3.39 The top portion of the ERO diff partial-area taxonomy, summarizing all classes with 0–11 object properties …………………………………………………… 160

3.40 The bottom portion of the ERO diff partial-area taxonomy. Most of the diff areas at lower levels are removed areas due to the reduction in ERO's complexity …………………………………………………………………….. 161

3.41 (a) The subcomponents of the BLUSNO and BLUOWL. (b) The structure of the major components of the BLU Framework ……………………………………... 169

LIST OF FIGURES
(Continued)

Figure **Page**

3.42 The BLUSNO abstraction network derivation user interface …………………… 172

3.43 BLUSNO's graphical interface with the *Specimen* hierarchy's partial-area taxonomy shown. The partial-area *Respiratory sample* (36) has been selected by clicking on it (yellow) and its parent (blue) and children (purple) are highlighted 174

3.44 (Left) The Partial-area Summary Dialog for the *Respiratory sample* partial-area. (Right) The singly rooted hierarchy of concepts in the *Respiratory sample* partial-area, with concept *Lower respiratory sample* (in yellow) selected ……… 175

3.45 The Area Summary Dialog for the {*Substance*} area …………………………… 178

3.46 The disjoint partial-area taxonomy for the area {*Substance*} in the Specimen hierarchy's partial-area taxonomy. The disjoint partial-area *Body fluid sample* (yellow) has been selected and its parents (blue) and children (purple) are highlighted ………………………………………………………………………. 178

3.47 Part of a TAN derived for the *Mass of body structure* partial-area in the *Clinical finding* partial-area taxonomy ………………………………………………… 180

3.48 The BLUSNO concept browser with *Pneumonia* as the focus concept ……….. 181

3.49 The BLUOWL taxonomy derivation user interface …………………………….. 183

3.50 A domain-defined partial-area taxonomy derived for OCRe's *Entity* hierarchy in BLUOWL ……………………………………………………………………….. 183

3.51 The diff partial-area taxonomy derivation user interface ………………………... 185

3.52 The top seven levels of the diff partial-area taxonomy for the eagle-I Research Resource Ontology (ERO), as shown in BLUOWL …………………………….. 186

3.53 The diff partial-area summary dialog for the removed partial-area *processual entity* ……………………...……………………...……………………...……….. 188

3.54 The diff partial-area explanation tab …………………………………………….. 186

3.55 A segment of the Gene Ontology's *Biological process* partial-area taxonomy, as viewed in the BLUOWL Protégé Plugin ………………………………………... 190

4.1 The concepts from Figure 3.14 grouped based on their common tribes. The two children of *Function* in the example, *Breast function* and *Digestive system function* are now defined as patriarchs instead of *Function* ……………………... 197

LIST OF FIGURES
(Continued)

Figure **Page**

4.2 (a) The first two levels of the structural meta-ontology. (b) The classes of Level 1 refined based on their usage. **(c)** The complete structural meta-ontology for 281 BioPortal ontologies with *F* = {*object property*, *data property*} …………... 199

LIST OF DEFINITIONS

Abstraction Network (“AbN”)
A compact summary of an ontology’s content and structure.

Area
A set of concepts/classes that have the exact same set of attribute relationship types. Summarizes structurally similar concepts/classes. Areas are disjoint.

Area root
A concept/class that has a different set of relationship types than all of its parents. I.e., all of its parents are in different areas. An area may have multiple roots.

Area taxonomy
An abstraction network where the nodes represent areas. Area nodes are connected by child-of links that summarize hierarchical relationships between concepts/classes of different areas. The areas in an area taxonomy are organized in color-coded levels based on their number of relationships.

Band Tribal Abstraction Network
An abstraction network that summarizes the hierarchically similar groups of concepts that belong to one or more of the same tribes.

Child-of link

area taxonomy
A link that summarizes the hierarchical relationships between concepts/classes in different areas.

band tribal abstraction network
A link that summarizes the hierarchical relationships between concepts in different bands.

cluster tribal abstraction network
A link that summarizes the hierarchical relationships between the cluster root and its parents, which are in clusters that are in different bands.

disjoint partial-area taxonomy
A link that summarizes the hierarchical relationships between the disjoint partial-area root and its parents, which may be in the same area (overlapping disjoint partial-area) or in a different area (non-overlapping disjoint partial-area).

Child-of link (continued)

partial-area taxonomy

A link that summarizes the hierarchical relationships between the partial-area root and its parents, which are in partial-areas that are in different areas.

Cluster Tribal Abstraction Network

A refinement of the Band Tribal Abstraction Network into subhierarchies of closely semantically related concepts.

Disjoint partial-area

The subhierarchy of concepts/classes in an area that are descendants of the same overlapping root (and no other overlapping roots). Each concept/class in a hierarchy belongs to exactly one disjoint partial-area.

Disjoint partial-area taxonomy

An abstraction network where the nodes represent disjoint partial-areas that are connected by child-of links. Disjoint partial-areas are labeled with their root's fully specified name and total number of concepts in their subhierarchy. Disjoint partial-areas shown using a single color (non-overlapping) or with multiple colors (overlapping), where each color represents which partial-area its concepts overlap between.

Disjoint partial-area subtaxonomy

A disjoint partial-area taxonomy derived for the concepts in a subject subtaxonomy. Concepts may overlap with concepts that are outside of the subtaxonomy. Thus, this external overlap information is considered to provide a complete picture of overlapping in a subject subtaxonomy.

Domain-defined partial-area taxonomy

A partial-area taxonomy derived for an OWL ontology using object properties with explicitly defined domains.

Expanded Tribal Abstraction Network

A tribal abstraction network created by using a subset of the grandchildren of a hierarchy root as tribal patriarchs.

Focus subtaxonomy

A partial-area taxonomy that summarizes of all of the ancestors and descendants of a chosen concept that represents a particular subject area.

Non-overlapping concept/class

A concept/class that is summarized by one partial-area in a partial-area taxonomy. Also referred to as *disjoint concept/class*.

Non-overlapping disjoint partial-area

A disjoint partial-area where all of its concepts/classes are non-overlapping concepts/classes.

Overlapping concept/class
A concept/class that is summarized by multiple partial-areas in a partial-area taxonomy.

Overlapping disjoint partial-area
A disjoint partial-area where all of its concepts/classes are overlapping concepts/classes.

Partial-area
A subhierarchy of concepts in an area that are descendants of the same area root concept. It summarizes a set of semantically similar concepts. Partial-areas are not necessarily disjoint. A concept can be a descendant of multiple area roots, and, thus, being summarized by multiple partial-areas.

Partial-area subtaxonomy
A subset of a partial-area taxonomy.

Partial-area taxonomy
An abstraction network where the nodes are partial-areas, connected by child-of links. The partial-areas in a partial-area taxonomy are shown as boxes within their respective areas, labeled with their root concept's fully specified name and the total number of concepts in the subhierarchy.

Recursive Tribal Abstraction Network
A tribal abstraction network derived for the concepts in a tribal abstraction network cluster.

Relationship subtaxonomy
Short for "relationship-constrained partial-area subtaxonomy." A subtaxonomy derived using a subset of relationships from a SNOMED CT hierarchy. Creates a subset of structurally similar concepts.

Restriction-defined partial-area taxonomy
A partial-area taxonomy derived for an OWL ontology with object properties that are used in class restrictions.

Root subtaxonomy
Short for "root-constrained partial-area subtaxonomy." A subtaxonomy that is derived by selecting a new root partial-area. Creates a subset of semantically similar concepts.

Subject subtaxonomy
A partial-area taxonomy rooted at a concept that represents a particular subject area.

Tribal Abstraction Network
An abstraction network based on the IS-A relationships in a SNOMED CT hierarchy.

CHAPTER 1

INTRODUCTION

1.1 Motivation

Biomedical ontologies and terminologies are knowledge structures used to represent the entities and relationships between entities within the domain of biomedicine. These knowledge structures support information encoding and interoperability in healthcare information systems, such as Electronic Healthcare Records (EHRs) [1-3]. Ontologies and terminologies also support interdisciplinary research [4, 5], information retrieval [4, 5], knowledge management [6-9], natural language processing (NLP) [10-12], and many other applications [4, 13].

Ontologies and terminologies represent knowledge using *concepts* and *relationships*. A concept represents a unique entity within a domain. A relationship represents a connection between exactly two concepts. Concepts are hierarchically organized using *subsumption relationships* (i.e., x ***IS-A*** y; x is a ***SUBCLASS*** of y) which form the backbone of an ontology. Subsumption relationships define the generalization and specialization of a concept [14]. For example, within the SNOMED CT terminology [15, 16], the concept *Infective pneumonia* has an IS-A relationship to the concept *Pneumonia* because infective pneumonia is a specialization of pneumonia (see Figure 2.1). Concepts can be further defined using labeled *lateral relationships*, which express non-hierarchical connections between two concepts. In SNOMED CT, the concept *Pneumonia* has a ***Finding site*** lateral relationship to the concept *Lung* because pneumonia occurs in the lungs.

In general, ontologies are more formally modeled than terminologies. Terminologies, such as SNOMED CT, may be structurally similar to ontologies. Stenzhorn et al. [17] discuss the differences between terminologies and ontologies in the context of clinical ontologies. Schulz et al. briefly discuss the differences between terminologies and ontologies [18]. For this dissertation, ontologies and terminologies in general will be referred to as just *ontologies*. When referring to a specific ontology or terminology the appropriate term will be used, e.g., SNOMED CT *terminology* and OWL *ontology*.

Ontologies are typically large and complex. They often contain hundreds of thousands of concepts and millions of relationships. The size and complexity of most ontologies makes it difficult to comprehend their content and structure. Comprehension is important for ontology integration, usability, and quality assurance. Errors and inconsistencies are unavoidable and difficult to detect in a large ontology. An error in an ontology may lead to errors in applications which use the ontology [19]. However, resources for ontology quality assurance are typically very limited. Therefore, it is necessary to develop methods that assist ontology quality assurance efforts.

Traditionally, ontologies are viewed through the lens of a concept browser. These browsers typically show information for one concept at a time. Concepts browsers provide information about a concept’s neighborhood: its parents, children, relationships, and attributes. Concept browsers are good at providing a micro-level view of an ontology’s content but they are not helpful for understanding the content and structure of an entire ontology, i.e., the macro-level “big picture.”

In previous work, *abstraction networks* were developed to summarize the content and structure of several different ontologies [20-25]. An abstraction network is a relatively compact collection of nodes and links derived from the underlying ontology. Each abstraction network node represents a subset of concepts which are determined to be "similar," where the definition of similar is based on the type of abstraction network being derived. Abstraction network links summarize the ontology's subsumption hierarchy. Abstraction networks provide ontology users and developers with a compact visualization of an ontology's content and structure. Additionally, abstraction networks have been shown to support ontology quality assurance by identifying groups of concepts that are more likely to contain errors and inconsistencies than other concepts [22, 26-29].

This thesis describes improvements to existing abstraction network derivation methodologies, derivation techniques for new types of abstraction networks, and new abstraction-network-based quality assurance methods. Additionally, tools for the automatic derivation and visualization of abstraction networks are introduced. Five important research areas are described:

1. The scalability of previously developed abstraction network quality assurance methodologies to large SNOMED CT hierarchies

2. A hierarchy-based abstraction network for SNOMED CT hierarchies that have no outgoing attribute relationships

3. Generalized abstraction network derivation techniques for groups of structurally similar ontologies

4. An abstraction-network-based summary of structural changes between two ontology releases called a Diff Abstraction Network

5. Software tools for automatically deriving, visualizing, and exploring abstraction networks

1.2 Dissertation Overview

Chapter 2 provides background information on biomedical ontologies, abstraction networks, abstraction-network-based quality assurance methodologies, and a literature review of related work. Chapter 3 describes the methods and results for the various research topics introduced at the end of Section 1.1. Section 3.1 introduces four methodologies for creating subsets of SNOMED CT abstraction networks called *subtaxonomies*. Section 3.2 describes a new type of abstraction network, called the Tribal Abstraction Network, for SNOMED CT hierarchies which have no lateral relationships. In Section 3.3 generalized abstraction network derivation methodologies for structurally similar Web Ontology Language ontologies are described. Section 3.4 describes Diff Abstraction Networks for summarizing and visualizing the structural changes between two ontology versions. Lastly, Section 3.5 describes software systems for creating and visualizing abstraction networks for SNOMED CT and OWL ontologies. Chapter 4 describes future work and Chapter 5 contains concluding remarks.

CHAPTER 2

BACKGROUND

2.1 Biomedical Ontologies

Ontologies are formally modeled knowledge structures which cover the concepts, individuals, relationships, attributes, axioms, rules, and terms of a particular domain. Ontologies are a type of controlled terminology. An ontology's concepts are usually organized into a *subsumption hierarchy* (e.g., IS-A or subclass relationships; *Infective pneumonia* **IS-A** *Pneumonia*). Further connections between pairs of related concepts are represented by lateral relationships, e.g., *Pneumonia* ***Finding site*** *Lung*. Knowledge about an individual concept, such as its unique identifier, name, definition, and synonyms, are often treated as *attributes* of the concept. Many ontologies are developed using Description Logic (DL) [30], which enables the formal definition of concepts. The subsumption hierarchy serves as a skeleton of an ontology and supports the inheritance of properties, such as relationships and attributes, by a concept from its parent concepts. Several examples of ontologies will be provided throughout this dissertation.

"[Biomedical] researchers must aggregate and integrate information, and they need tools to enable knowledge discovery in this data-rich paradigm. [Ontologies] describe the structure of their complex domains and relate their data to shared representations of biomedical knowledge" [4]. Thus, modern biomedical research, which often relies on the interoperability of large sets of data, is more difficult without ontologies. In the field of biomedicine, ontologies have become important for constructing intelligent decision-support systems, simulation systems, and information-retrieval systems [4, 5, 31, 32]. Ontologies are also becoming increasingly important for natural language processing systems [10-12] and the standardized encoding of Electronic Healthcare Record (EHR) data [1, 2]. Several ontologies and ontology frameworks that play a significant role in this thesis will now be explained in detail.

2.1.1 SNOMED CT

SNOMED CT (formerly the Systematized Nomenclature of Medicine – Clinical Terms, or SNOMED Clinical Terms) [15, 16] is a large medical terminology managed by the International Health Terminology Standards Development Organization (IHTSDO), a multinational organization with over 20 member nations [33]. SNOMED CT was created by merging the SNOMED Reference Terminology (SNOMED RT) with the Clinical Terms Version 3 (CTV3) terminology [16]. New versions of SNOMED CT are released in January and July of each year. Recent versions of SNOMED CT contain nearly 300,000 active medical concepts connected by almost 1.5 million relationships. SNOMED CT's concepts are organized hierarchically using IS-A relationships.

SNOMED CT plays an important role in EHRs by providing standardized encodings for healthcare data. By 2015, SNOMED CT is slated to be the standard terminology for encoding diagnoses and problem lists in EHRs in the United States [34]. SNOMED CT can also be used for natural language processing, data mining, and cross mapping between other terminologies, such as ICD-10 [35]. Subsets of concepts can be extracted from SNOMED CT to cover a particular domain. Examples of such subsets include the Clinical Observations Recording and Encoding (CORE) problem list [36], the Veterans Health Administration problem list, and the Kaiser Permanente problem list [37]. These problem lists are composed of sets of concepts that are deemed useful for the encoding of clinical information.

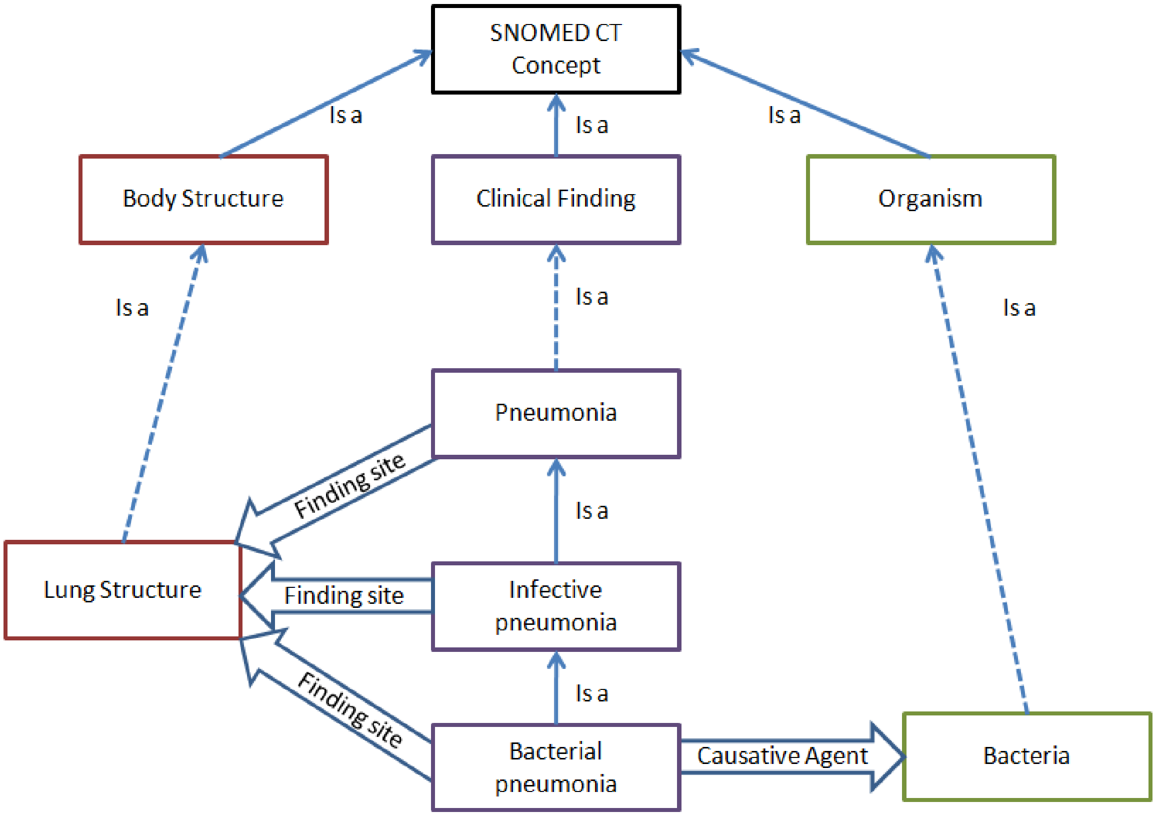


Figure 2.1 An example of SNOMED CT's structure for the concept Bacterial pneumonia. Concepts are shown as labeled boxes, IS-A relationships are shown as thin blue arrows directed upwards. Dashed blue arrows represent a sequence of IS-A relationships. Attribute relationships are shown with thick labeled arrows.
Source: [38]

SNOMED CT is created using a DL language named EL [39]. EL includes a subset of the functionality available in complete DL. SNOMED CT is publicly distributed in two forms: the inferred view and the stated view. SNOMED CT is structured as a directed acyclic graph (DAG); concepts may have more than one parent concept. Concepts are organized into 19 mostly disjoint top-level hierarchies that cover topics such as medical procedures, clinical findings, and anatomy. SNOMED CT calls lateral relationships by the somewhat misleading name "attribute relationships" (just relationships for short).

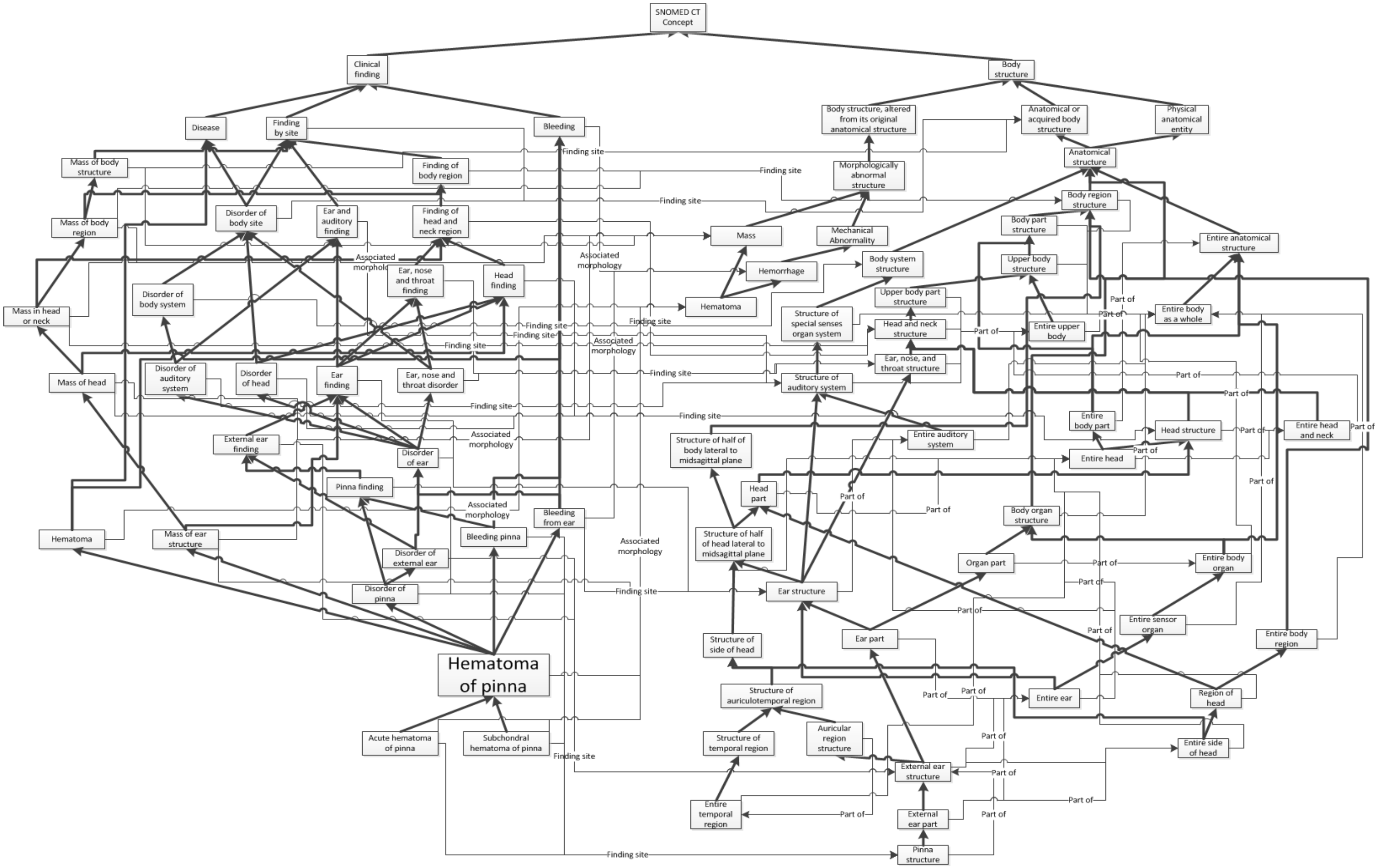


Figure 2.2 The concepts and relationships needed to define the concept Hematoma of pinna in SNOMED CT. Concepts are shown as labeled grey boxes, thick black edges represent IS-A relationships, while thin labeled edges represent attribute relationships.

Lateral relationships are used to further define concepts by creating nonhierarchical associations to other concepts in SNOMED CT, e.g., *Bacterial pneumonia* is the source of a *finding site* relationship with a target concept *Lung structure* (Figure 2.1).

Figure 2.1 provides an example of SNOMED CT's structure using a box and arrow diagram where each concept is a labeled box and arrows are used to express relationships between concepts. Figure 2.1 shows the relationships used to define the concept *Bacterial pneumonia*. This figure shows that *Bacterial pneumonia* IS-A *Infective pneumonia* which is caused by a type of *Bacteria* (expressed using the *causative agent* relationship). Additionally, *Bacterial pneumonia* has a *finding site* of *Lung structure*. The *finding site* attribute relationship is inherited from *Pneumonia,* the grandparent of *Bacterial pneumonia*. Figure 2.2 shows a more complicated example, with the concepts and relationships needed to define the concept *Hematoma of pinna* (i.e., *Bleeding pinna*).

SNOMED CT concepts are pre-coordinated; a single concept identifier is used to represent a single clinical idea. However, SNOMED CT also enables the use of post-coordination to represent a concept by combining two or more concepts, e.g., by creating a single expression consisting of several concepts related to each other by attributes. A SNOMED CT concept, and in general every DL concept, is either *fully-defined* or *primitive*. In primitive concepts, the definition is underspecified, meaning automated detection of subconcepts is not possible [30, 38].

2.1.2 Web Ontology Language (OWL)

The Web Ontology Language (OWL) [40], developed by the World Wide Web Consortium (W3C), is a standardized framework and family of languages for creating ontologies. Many well-known biomedical ontologies have been developed using OWL, including the NCI Thesaurus (NCIt) [41], the Gene Ontology (GO) [42], and the Ontology of Biomedical Investigations (OBI) [43]. Ontologies developed in OWL will be referred to as *OWL ontologies*.

OWL is based on DL and provides formal methods for defining ontological elements, such as concepts (called classes) and their relationships (called properties). There are several OWL sublanguages, including OWL Lite, OWL DL, and OWL Full. Each sublanguage offers different levels of DL expressiveness, with OWL Full being the most expressive. Additionally, there are several different OWL syntaxes, including OWL XML syntax and Manchester syntax [44].

OWL ontologies are composed of *classes*, which represent sets of DL concepts. Classes are organized as a subsumption hierarchy using *subclass of* relationships (also known as *superclass* relationships). The example below, expressed in Manchester syntax, is the definition of the class *Multiple Sleep Latency Test* in the Sleep Domain Ontology (SDO) [45]. It states that *Multiple Sleep Latency Test* is a subclass of *Polysomnography*. Within OWL, annotations, such as labels and comments, can be provided for individual classes and properties. For example, the class *Multiple Sleep Latency Test* is annotated with a text definition (rdfs:comment) and a label (rdfs:label).

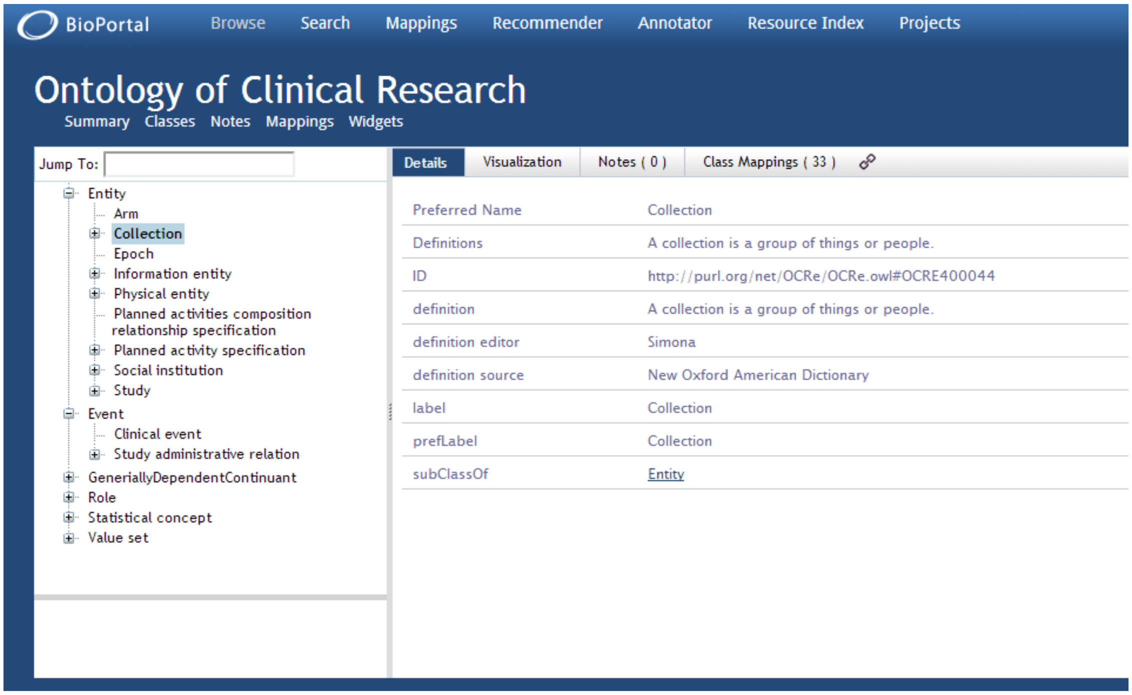
```
Class: SDO:MultipleSleepLatencyTest
Annotations:
    rdfs:comment "A validated, objective measure of the ability or tendency
            to  fall  asleep  under  standardized  conditions.   [Sleep  Medicine
            Essentials - Teofilo L. Lee-Chiong]",
    rdfs:label "Multiple Sleep Latency Test (MSLT)"
 SubClassOf:
    SDO:Polysomnography,
    <http://purl.org/cpr/hasOutput> some SDO:SleepOnsetLatency
```

Classes can be further defined by using properties. A property is a directed binary relation between two or more classes (*object properties*) or between classes and literal values (*data properties*). Both types of properties can be explicitly assigned domains and ranges (e.g., source and target classes of the binary relation), which serve as global restrictions on a property's use. Alternatively, properties can be used as restrictions on class definitions, serving as a local restriction on their use. In the above example, the object property *has output* with the uniform resource identifier (URI) *http://purl.org/cpr/hasOutput* is used as a restriction on the class *Multiple Sleep Latency Test*.

OWL ontologies can import and extend other ontologies. A common ontology design pattern is to use a top level ontology, such as the Basic Formal Ontology (BFO) [46], and add classes and properties specific to a domain. Top domain ontologies, such as the Ontology for General Medical Sciences (OGMS) [47] and BioTop[48], extend top level ontologies and introduce general domain knowledge. Many ontologies, such as the Sleep Domain Ontology (SDO) [45] and Vital Sign Ontology (VSO) [49], import top domain ontologies and extend them with their own specific knowledge, e.g., sleep medicine and vital signs, respectively. This approach of importing ontologies enables interoperability and reuse of ontologies.

2.1.3 NCBO BioPortal

The National Center for Biomedical Ontology (NCBO) BioPortal is a large repository of ontologies that are focused on the domains of medicine and biology [50]. BioPortal is currently one of the largest ontology repositories, containing over 300 ontologies.



Preferred Name	Collection
Definitions	A collection is a group of things or people.
ID	http://purl.org/net/OCRe/OCRe.owl#OCRE400044
definition	A collection is a group of things or people.
definition editor	Simona
definition source	New Oxford American Dictionary
label	Collection
prefLabel	Collection
subClassOf	Entity

Figure 2.3 The BioPortal user interface. The Ontology of Clinical Research (OCRe) [51] was selected from the list of available ontologies.

The ontologies in BioPortal are made available in various formats such as Web Ontology Language (OWL) [40], Open Biological and Biomedical Ontologies (OBO) [52], and Resource Description Framework (RDF) [53]. Ontologies available on BioPortal will be referred to as *BioPortal ontologies*. BioPortal provides an interface for browsing, searching, and visualizing ontologies hosted in its repository. Figure 2.3 shows the BioPortal interface when the Ontology of Clinical Research (OCRe) [51] is selected from a list of hosted ontologies. BioPortal provides public APIs for retrieving ontologies and information about the concepts and relationships within an ontology.

The BioPortal ontologies cover a wide variety of topics in the field of biomedicine, including infectious diseases, drugs, and anatomy. BioPortal is also an important resource for understanding trends and preferences in ontology development. By analyzing the BioPortal ontologies, one can obtain an understanding of common design techniques and knowledge modeling choices used by the BioPortal community to create ontologies. For example, one can analyze which ontologies utilize a certain top level ontology, or compare how two ontologies model the same concept. Mortensen et al. [54] analyzed the use of ontology design patterns in BioPortal ontologies and He et al. [55] analyzed the structure of a sample of BioPortal ontologies.

2.2 Abstraction Networks

Some biomedical ontologies are very large and complex knowledge structures. Size and complexity prevent users of an ontology from seeing the “big picture” of its content. Seeing the big picture of an ontology is important for browsing and searching for content, integration into applications, extending content, reusing content, and cross mapping to make associations with other ontologies. Additionally, seeing the big picture is important for quality assurance of ontology content.

Visualizing an entire ontology using a box and arrow diagram, such as Figure 2.1, allows a user to see the big picture for many concepts at once. Figure 2.2 shows all of the concepts, and most of the relationships, needed to define the SNOMED CT concept *Hematoma of pinna*. However, as more concepts and relationships are added to such a figure, it becomes overwhelming and its usefulness is lost. Figure 2.2 is considered to be on the boundary of being too overwhelming to be useful. Figure 2.4 shows the hierarchical relationships between the 4,503 concepts in the *Physical object* hierarchy of SNOMED CT and is overwhelming to the point of being useless. Figure 2.4 would be even less comprehensible if incoming relationships to target concepts within the hierarchy were also included.

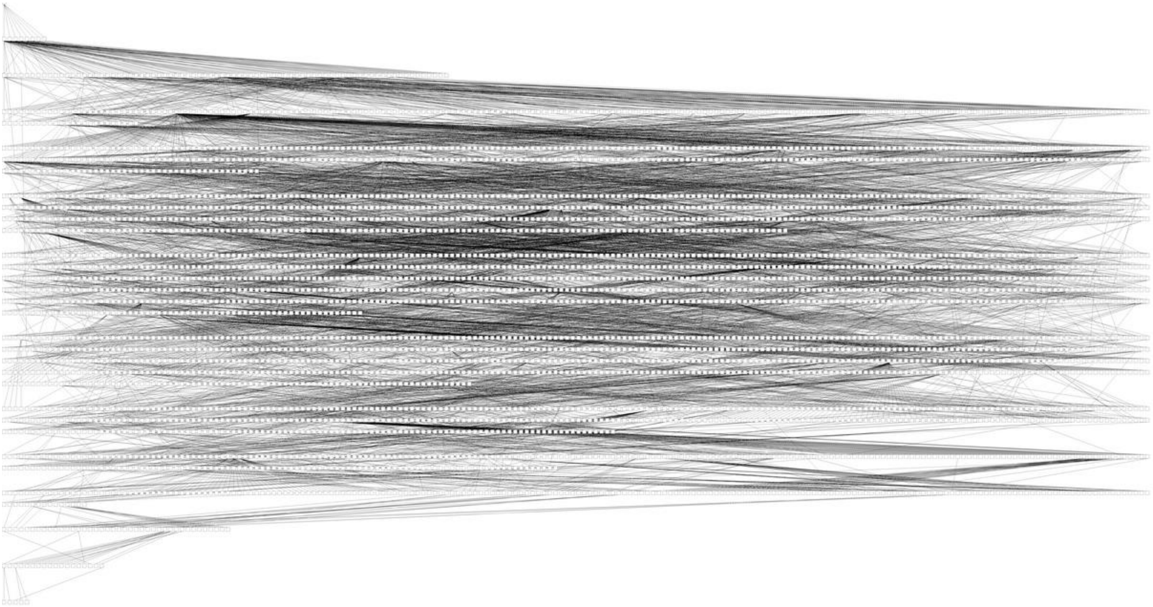
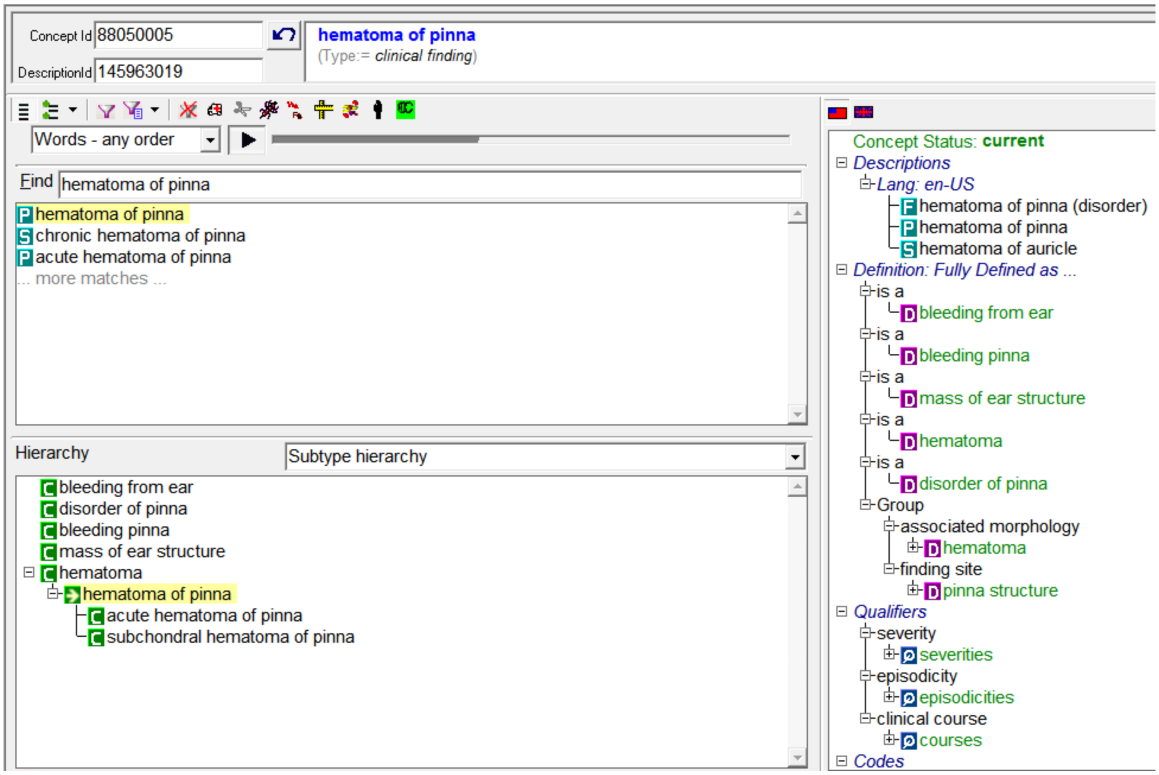


Figure 2.4 The hierarchical relationships between the 4,503 concepts in the *Physical object* hierarchy of SNOMED CT.

Due to the difficulty of visualizing an ontology, content is traditionally viewed using a concept browser. Concept browsers show information for a small number of concepts at a time (often only one concept, called the *focus concept*). Examples of concept browsers include Protégé [56], CliniClue [57], and the UTS browsers for the UMLS and SNOMED CT [58]. Figure 2.5 shows the SNOMED CT concept *Hematoma of pinna* as displayed in the CliniClue browser. Most concept browsers only show one concept and its immediate neighborhood: its parents, children, relationships, and synonyms. This view of an ontology is very limited. A user cannot obtain the big picture of an ontology's content and structure.

One way of obtaining the big picture of an ontology is through summarization. In previous research [20-25], various types of a*bstraction networks* have been developed to summarize the content and structure of various ontological and terminological systems.



Concept Id 88050005

DescriptionId 145963019

Words - any order

Find hematoma of pinna

Hierarchy Subtype hierarchy

Figure 2.5 The concept *Hematoma of pinna* as viewed using the CliniClue SNOMED CT concept browser.

Abstraction networks consist of *nodes* which summarize a set of “similar” concepts, where the definition of similar is dependent on the type of abstraction network being created. Figure 2.6 illustrates the general process of deriving an abstraction network from an ontology. On the left, a hierarchy of concepts is represented using small, filled colored ovals. IS-A relationships are represented as black lines. Groups of similar concepts are illustrated using large, colored ellipses. The abstraction network created from the groupings of similar concepts is shown on the right side of Figure 2.6. Each group appears as one rectangular box, referred to as a node.

Abstraction networks support usability, comprehensibility, visualization, and quality assurance by producing a compact view of an ontology.

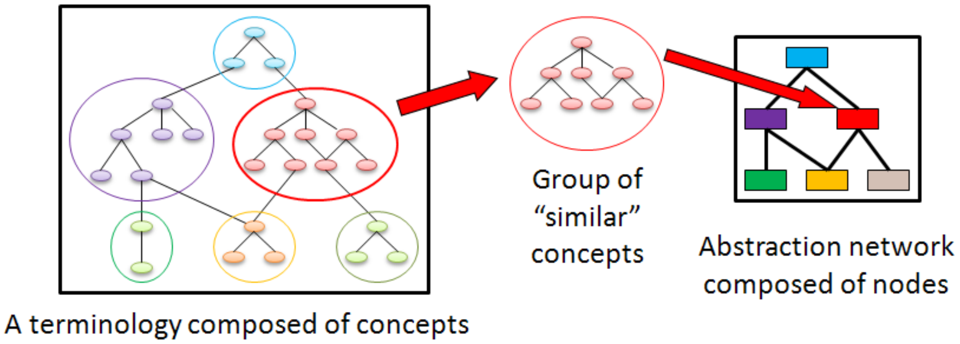


Figure 2.6 The general process of deriving an abstraction network for an ontology or terminology.

Abstraction networks are designed to be significantly reduced in size and complexity when compared to the underlying ontology. Using an abstraction network, one can view large portions of an ontology to obtain the "big picture" of an ontology's content and structure.

2.2.1 Previously Developed Abstraction Networks

Different types of abstraction networks have been developed to summarize several different ontologies and terminologies. The abstraction network paradigm has been applied as the Refined Semantic Network (RSN) [59] to the Unified Medical Language System (UMLS) [60] and as the Schema [61] for the Medical Entities Dictionary (MED) [62]. The area and partial-area taxonomy abstraction networks were developed [22] for the National Cancer Institute thesaurus (NCIt) [41] and in [24] for SNOMED CT [16] hierarchies with attribute relationships (7 out of 19). The disjoint partial-area taxonomy abstraction network for SNOMED CT [23] further refines a partial-area taxonomy into disjoint groups called disjoint partial-areas. Due to the importance of the SNOMED CT area, partial-area, and disjoint partial-area taxonomies for this dissertation, their derivation methodologies will now be explained in detail.

2.2.1.1 Area and Partial-area Taxonomies for SNOMED CT. The area taxonomy and partial-area taxonomy are abstraction networks for SNOMED CT that summarize structurally and semantically similar concepts into groups called *areas* and *partial-areas*, respectively [24]. These taxonomies were developed as part of an ongoing effort to summarize SNOMED CT and enable the quality assurance of its content. An *area* summarizes a set of concepts that all share the exact same set of outgoing relationships. An *area taxonomy* is an abstraction network where the areas are nodes. Diagrammatically, an area is a box labeled with the common relationship names. In text, the relationship names are placed in braces to form the area name. Concept information aside from the relationships and number of concepts is abstracted away.

To demonstrate this, consider Figure 2.7 (a) with 17 concepts (labeled with their fully specified names) from the *Specimen* hierarchy. The thin arrows, which are directed upwards, are IS-A relationships between concepts. Concepts with the same outgoing attribute relationships are grouped together in a common dashed colored bubble. For example, the concepts *Swab* and *Biopsy sample* have a single relationship named *Procedure. Specimen, Living sample, Genetic sample, Parasite sample,* and *Polar body sample* have no relationships and are thus grouped in the ∅ (empty set) bubble.

Figure 2.7 (b) shows the area taxonomy for the concepts in Figure 2.7 (a). *Swab*, *Biopsy sample*, and *Swab of inanimate object* are now represented solely by the *Procedure* area [55] with three concepts. Similarly, *Upper respiratory swab sample, Cough swab, Swab from larynx, Swab from abdomen,* and *Swab from appendix* are represented by the area {*Procedure, Topography*}. Areas are organized into color-coded levels based on the number of relationships.

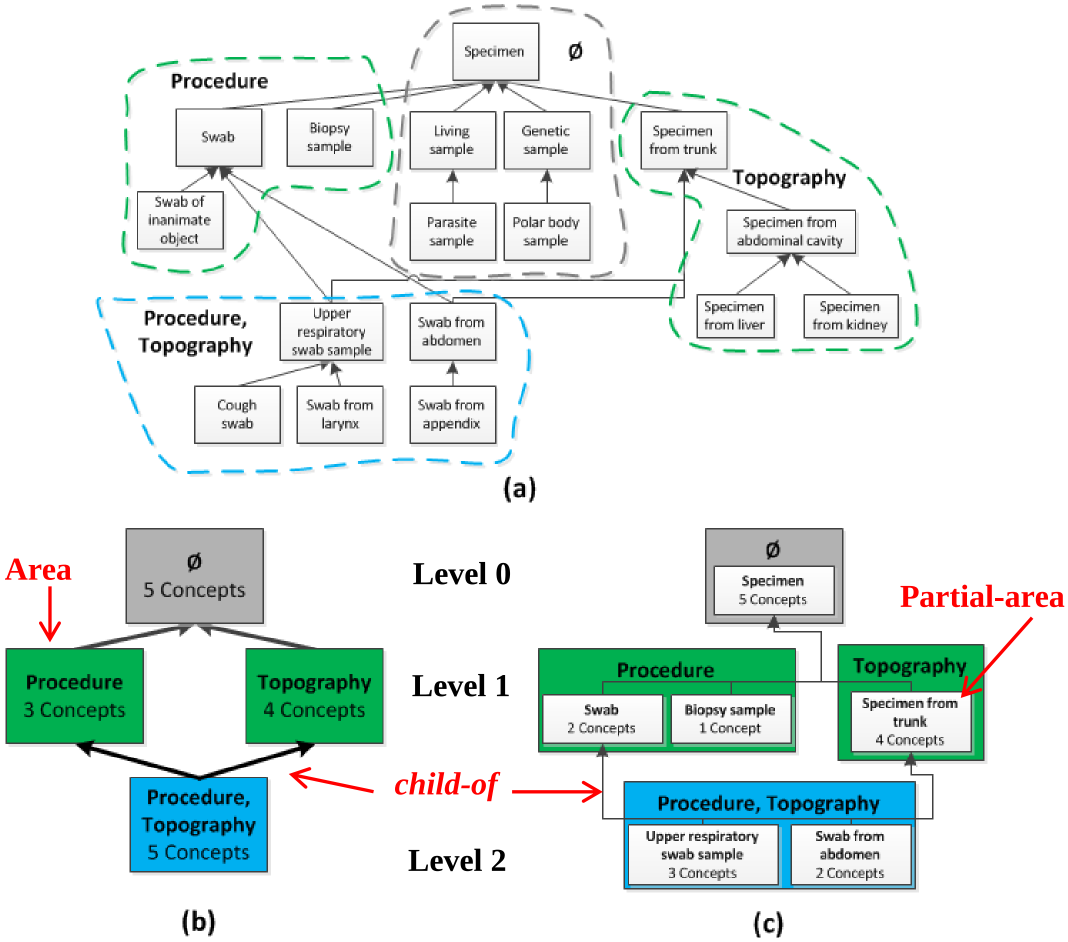


Figure 2.7 **(a)** A sample of 17 concepts taken from the *Specimen* hierarchy. **(b)** The area taxonomy for the concepts in (a). **(c)** The partial-area taxonomy for the concepts in (a). Taxonomic elements have been labeled with red text and arrows.

Figure 2.8 shows the complete area taxonomy for the *Specimen* hierarchy. It consists of 22 areas organized into five levels. At the top are concepts with no relationships.

In every area there will be one or more concepts that do not have a parent concept within the area. Such concepts are called *roots*. An IS-A from a root to its parent in another area yields a hierarchical connection between the respective areas called *child-of*. In Figure 2.7 (b), *child-of*'s are represented as bold lines. For example, {*Procedure, Topography*} is *child-of* {*Procedure*} and {*Topography*}. IS-As between concepts within an area are abstracted away, just as the concepts they are connecting. In Figure 2.8 *child-*

of relationships between areas are colored according to the child areas' level to enable readability. Every concept is in exactly one area, i.e., all areas are disjoint.

The *partial-area taxonomy* refines the area taxonomy with the inclusion of *partial-areas*, each consisting of a root and all of its descendants in its area. Thus, the number of partial-areas in an area is equal to the number of roots. Figure 2.7 (c) shows an example of the derivation of partial-areas. Each partial-area appears as a white box inside its area. Each partial-area is labeled with its root's fully specified name and the number of concepts grouped into the partial-area, e.g., the partial-area *Specimen* contains five concepts. For a more compact visualization, the number of concepts in a partial-area may be shown in parenthesis next to the root's name (e.g., *Specimen* (29) in Figure 2.9). All other information about the concepts in the partial-area is abstracted away.

The concept *Swab*, a root of {*Procedure*}, and its child, *Swab of inanimate object*, are grouped into the partial-area *Swab*, the white box in {*Procedure*}. Partial-areas are also linked by *child-of*'s derived from the underlying IS-A relationships. Specifically, a partial-area *A* is a *child-of* another partial-area *B* if *A*'s root has a parent concept in *B*. In Figure 2.7(c), the partial-area *Swab* is *child-of* the partial-area *Specimen.* Figure 2.9 shows the complete partial-area taxonomy for the *Specimen* hierarchy. It consists of 419 partial-areas. Level 2 and Level 3 been organized into rows due to space limitations. All of the *child-of* links have been hidden for readability purposes.

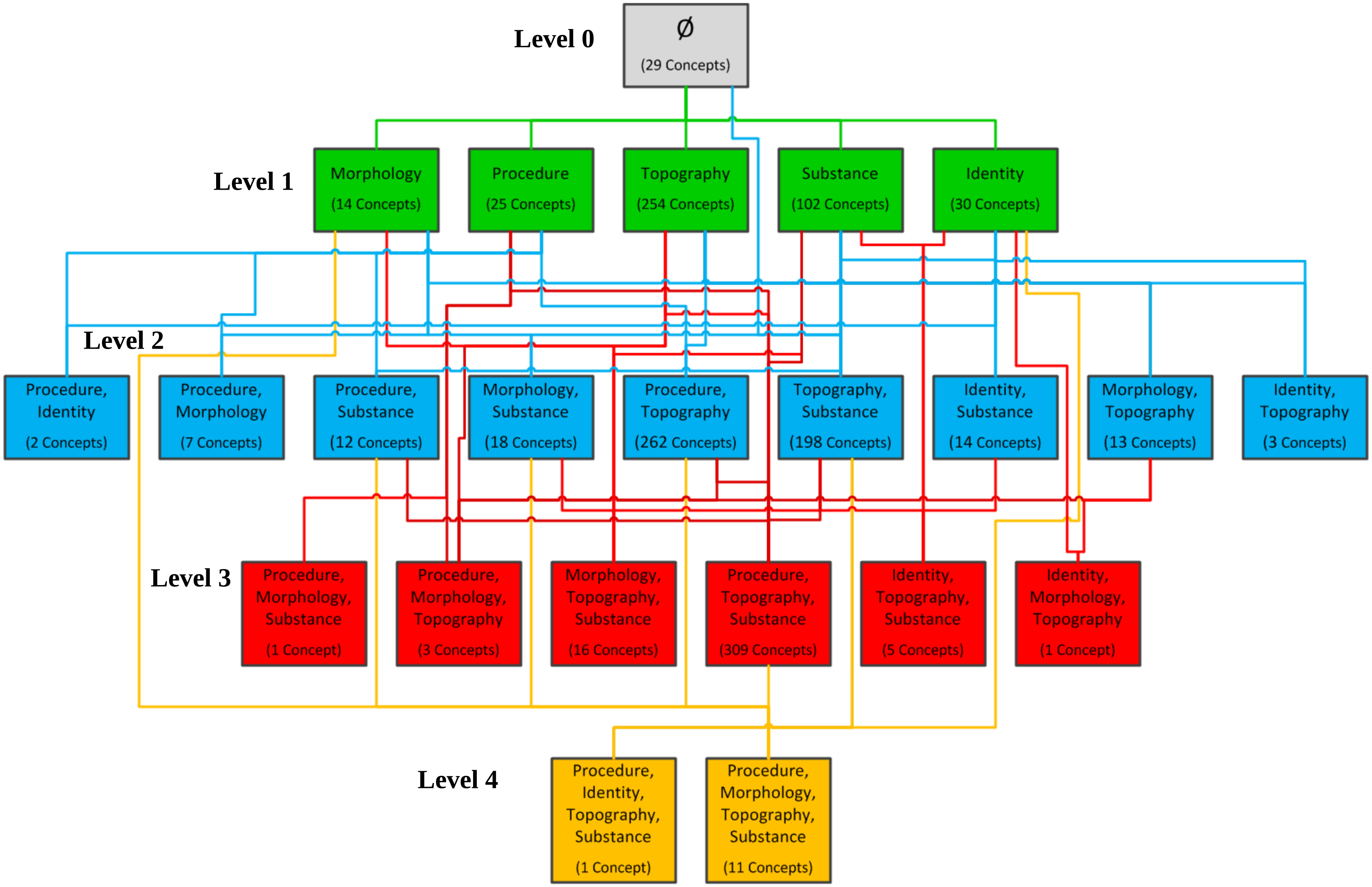


Figure 2.8 The complete area taxonomy for the *Specimen* hierarchy of SNOMED CT.

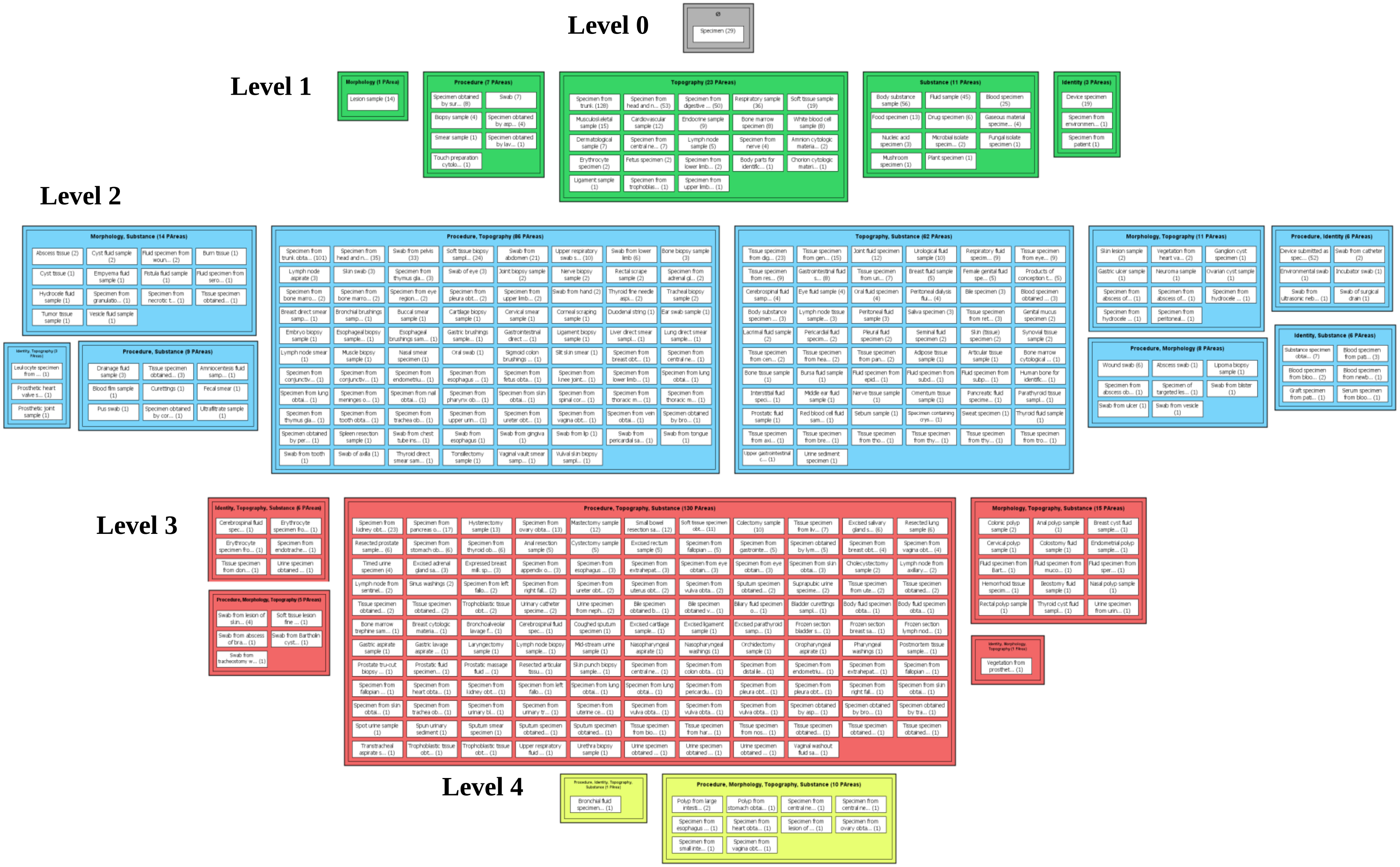

Level 0
∅
Specimen (29)
Level 1
Morphology (1 PArea)
Lesion sample (14)
Procedure (7 PAreas)
Topography (23 PAreas)
Substance (11 PAreas)
Identity (3 PAreas)
Level 2
Morphology, Substance (14 PAreas)
Procedure, Topography (86 PAreas)
Topography, Substance (62 PAreas)
Morphology, Topography (11 PAreas)
Procedure, Identity (6 PAreas)
Identity, Topography (3 PAreas)
Procedure, Substance (9 PAreas)
Procedure, Morphology (8 PAreas)
Identity, Substance (6 PAreas)
Level 3
Identity, Topography, Substance (6 PAreas)
Procedure, Morphology, Topography (5 PAreas)
Procedure, Topography, Substance (130 PAreas)
Morphology, Topography, Substance (15 PAreas)
Identity, Morphology, Topography (1 PArea)
Level 4
Procedure, Identity, Topography, Substance (1 PArea)
Procedure, Morphology, Topography, Substance (10 PAreas)

Figure 2.9 The partial-area taxonomy for the *Specimen* hierarchy. *Child-of* links between partial-areas are hidden due to space limitations.

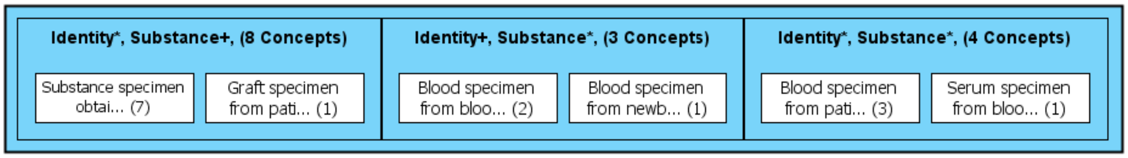

Identity*, Substance+, (8 Concepts)
Substance specimen obtai... (7)
Graft specimen from pati... (1)
Identity+, Substance*, (3 Concepts)
Blood specimen from bloo... (2)
Blood specimen from newb... (1)
Identity*, Substance*, (4 Concepts)
Blood specimen from pati... (3)
Serum specimen from bloo... (1)

Figure 2.10 The {*Identity, Substance*} area in SNOMED CT's *Specimen* hierarchy partitioned into inheritance regions.

Even though all concepts in an area have the same relationships (by definition), not all root concepts in an area obtain their relationship set in the same way. Some concepts inherit their relationships from a parent concept while others introduce a new type of relationship into the hierarchy. Areas can be partitioned into separate relationship obtainment pattern regions (called simply *regions*) [24]. Each region is distinguished by the pattern in which its relationships are introduced and/or inherited. Each region is named using the set of relationships for the associated area, but next to each relationship a '+' is appended to indicate if it is introduced at this concept or a '*' is appended to indicated if it is inherited from a parent of this concept. Graphically, all regions of a single area are drawn with a black outline within the same area box.

Figure 2.10 shows the three regions of the {*Identity, Substance*} area in SNOMED CT's *Specimen* hierarchy. The root concepts *Blood specimen from patient* and *Serum specimen from blood* inherit both relationships, *Identity* and *Substance*, while the root concepts *Blood specimen from blood donor* and *Blood specimen from newborn* inherit the *Substance* relationship and introduce the *Identity* relationship. Theoretically {*Identity, Substance*} may have up to four different regions: inherit the first relationships and inherit the second, inherit the first and introduce the second, introduce the first and inherit the second, and introduce both the first and the second relationships. However, in practice, many of the possible regions do not exist.

Partial-areas are not necessarily disjoint (remember that areas are disjoint, i.e., no concept can be in two areas). A given concept in a hierarchy may be grouped into more than one partial-area. When this occurs, the partial-areas that contain such concepts are called *overlapping partial-areas*. This situation occurs when a concept is a descendant of two or more roots in its area. These concepts are called *overlapping concepts*. Overlapping concepts elaborate the semantics of multiple roots within an area. In a partial-area taxonomy, overlapping concepts are counted as belonging to all of their root's respective partial-areas. For example, in the {*Substance*} area at Level 1 (green) of Figure 2.9, summing the number of concepts contained in each partial-area results in 157 concepts, which is a number larger than the number of unique concepts in the area, shown in Figure 2.8 (102 concepts). In {*Substance*} there are a total of 39 overlapping concepts, several of which overlap between three partial-areas.

The *disjoint partial-area taxonomy* abstraction network was developed to provide a complete and accurate view of the IS-A hierarchy within an area [23]. Based on the IS-A relationships between concepts within an area, a disjoint partial-area taxonomy partitions an area into disjoint, singly-rooted groups called disjoint partial-areas. Disjoint partial-areas are defined using concepts which are identified as *overlapping roots*.

Figure 2.11 provides an example of a hierarchy of overlapping concepts from {*Substance*} in SNOMED CT's *Specimen* hierarchy. Color is used to indicate which partial-areas an overlapping root overlaps between. For example, the overlapping root *Body fluid sample* overlaps between the partial-areas *Fluid sample* and *Body substance sample*. The single-colored concepts are "normal" partial-area roots. A concept is defined as a *base overlapping root* if all of its parents are non-overlapping concepts.

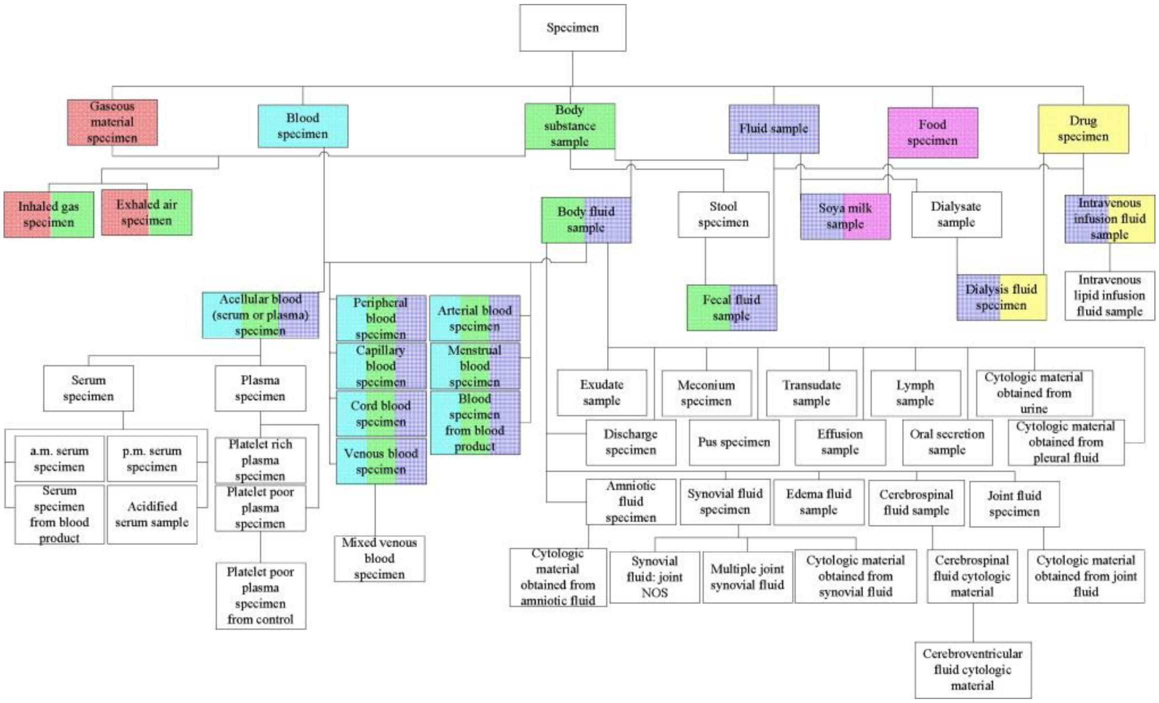

Specimen
Gaseous material specimen
Blood specimen
Body substance sample
Fluid sample
Food specimen
Drug specimen
Inhaled gas specimen
Exhaled air specimen
Body fluid sample
Stool specimen
Soya milk sample
Dialysate sample
Intravenous infusion fluid sample
Fecal fluid sample
Dialysis fluid specimen
Intravenous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

Figure 2.11 A hierarchy of overlapping concepts within the *Specimen* hierarchy's {*Substance*} area. Partial-area roots are singly colored. Overlapping roots are multicolored according to the partial-areas they overlap between.
Source: [24]

In Figure 2.11 *Inhaled gas specimen*, *Exhaled air specimen*, *Body fluid sample*, *Fecal fluid sample*, *Soya milk sample*, *Dialysis fluid specimen*, and *Intravenous fluid sample* are base overlapping roots. A concept L is an *overlapping root* if either it is a base overlapping root or there exist two concepts C1 and C2 (C1 ≠ C2) such that L is a descendant concept of both C1 and C2 and either C1 is an overlapping root and C2 is a partial-area root (or vice-versa) or both C1 and C2 are overlapping roots [23]. The blue-green-purple concepts in Figure 2.11 (e.g., *Arterial blood specimen*) are examples of non-base overlapping roots, as they all share a common ancestor (*Body fluid sample*) which is an overlapping root.

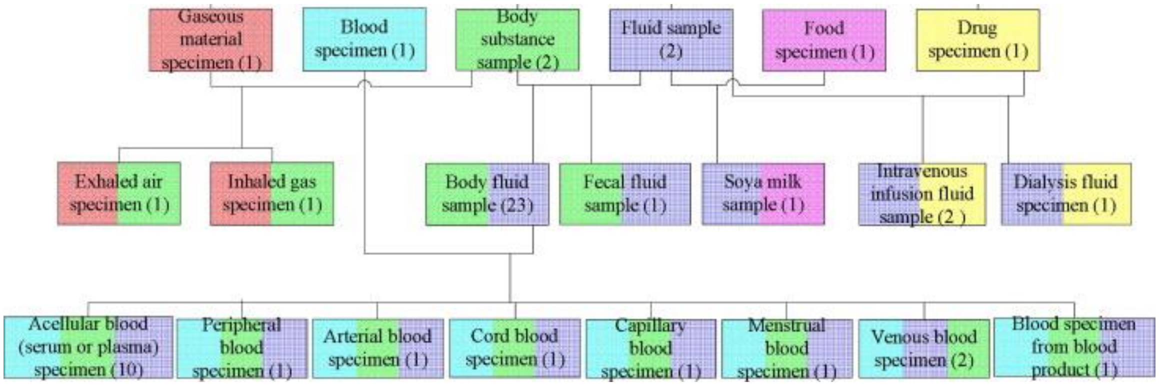

Gaseous material specimen (1)
Blood specimen (1)
Body substance sample (2)
Fluid sample (2)
Food specimen (1)
Drug specimen (1)
Exhaled air specimen (1)
Inhaled gas specimen (1)
Body fluid sample (23)
Fecal fluid sample (1)
Soya milk sample (1)
Intravenous infusion fluid sample (2)
Dialysis fluid specimen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

Figure 2.12 The disjoint partial-area taxonomy for the example of overlapping concepts in Figure 2.11.
Source: [24]

A disjoint partial-area consists of an overlapping root and all of its descendants that are not descendants of another overlapping root. That is, an overlapping concept *c* will belong to a disjoint partial-area *d* if there is a path from *c* to the root of *d* and no other overlapping roots are on the path from *c* to the root of *d*. For example, in Figure 2.11, the concept *Mixed venous blood specimen* belongs to the disjoint partial-area *Venous blood specimen* (2), and not the disjoint partial-area *Body fluid sample* (23), because the overlapping root concept *Venous blood specimen* is on the path from the concept *Mixed venous blood specimen* to the overlapping root concept *Body fluid sample*.

By definition, every concept in an area belongs to exactly one disjoint partial-area. Disjoint partial-areas are named after their overlapping root and are labeled with the total number of concepts summarized by the disjoint partial-area. For example, the 22 uncolored descendants of *Body fluid sample* in the bottom right of Figure 2.11will all belong to a disjoint partial-area named *Body fluid sample* (23) (see Figure 2.12). Disjoint partial-areas that summarize overlapping concepts are called *overlapping disjoint partial-areas*, while disjoint partial-areas that summarize non-overlapping concepts are referred to as *non-overlapping disjoint partial-areas*.

Disjoint partial-areas are formed into a *disjoint partial-area taxonomy* that summarizes the overlapping portions of an area. Like partial-areas, disjoint partial-areas are linked together by *child-of* edges based on the underlying IS-A hierarchy. Figure 2.12 shows the disjoint partial-area taxonomy derived from the overlapping concepts in Figure 2.11. It consists of 15 overlapping disjoint partial-areas and six non-overlapping disjoint partial-areas (shown at the top). Disjoint partial-areas are organized into rows based on how many partial-areas the concepts overlap between. For example, the disjoint partial-area *Body fluid sample* (23) is at Level 2, because its concepts overlap between two partial-areas. Non-overlapping disjoint partial-areas are assigned a single color and overlapping disjoint partial-areas are colored according to which partial-areas their concepts overlap between.

2.2.2 Abstraction Networks for Quality Assurance

Quality assurance is an important part of an ontology's lifecycle [22]. However, quality assurance for large ontologies is time consuming and manpower intensive. Resources for ontology quality assurance are typically very limited. Comprehensive reviews of an ontology's content are impractical due to the size of most ontologies. However, as they are compact summarizations of ontologies, abstraction networks can be used to support quality assurance efforts. While abstraction networks do not automatically identify errors, reviewing the nodes of an abstraction network can lead to the identification of errors and inconsistencies in the underlying ontology.

For example, reviewing the "Schema" abstraction network for the Medical Entities Dictionary (MED) [62] helped identify errors in its modeling [61]. Several types of errors were uncovered when reviewing the nodes of the Refined Semantic Network

(RSN) [59, 63] for the UMLS [60]. Ochs et al. [25] and He et al. [55] showed that abstraction networks can support the quality assurance of OWL ontologies, such as the Ontology of Clinical Research (OCRe) [51] and the Cancer Chemoprevention Ontology (CanCo) [64]. He et al. [55] introduced a family-based quality assurance approach using abstraction networks for structurally similar OWL ontologies.

For various abstraction networks, certain nodes have been identified as being more likely to contain erroneous concepts than other nodes. Analysis of the RSN found that concepts that belong to a kind of small node (i.e., a node that summarizes few concepts) called an intersection semantic type are more likely to contain errors than concepts which belong to large nodes [65]. Similarly, small partial-area nodes in the NCI thesaurus partial-area taxonomies were identified as being more likely to contain erroneous concepts when compared to concepts in large partial-area nodes [22].

Extensive research has been conducted on the use of SNOMED CT [15] partial-area taxonomy abstraction networks [24] for quality assurance. Wang et al. [24] identified several groups of concepts that are more likely to contain errors, including areas with a few small partial-areas, small partial-areas with many relationships, and small partial-areas in strict inheritance regions. Halper et al. [26] identified three groups of concepts in a partial-area taxonomy which have higher error rates: concepts in strict inheritance regions, concepts in mixed regions, and concepts in small partial-areas. Ochs et al. [28] found that concepts in small partial-are more likely to contain errors in a subset of the large *Procedure* hierarchy (see Section 3.1). Wei and Bodenreider [66] showed that partial-area taxonomies can identify errors that cannot be uncovered using Description Logic classifiers, such as HermiT [67] and Pellet [68].

The disjoint partial-area taxonomy [23] for SNOMED CT was also shown to support quality assurance. Wang et al. [29] showed that concepts that overlap between two or more partial-areas (*overlapping concepts*) are more likely to be erroneous than concepts in only one partial-area. Ochs et al. [27] used the Tribal Abstraction Network (described in detail in Section 3.2) to review the *Observable entity* hierarchy of SNOMED CT. It was found that concepts in large nodes were more likely to be erroneous than concepts in small nodes. Lastly, Wei et al. [69] applied the Converse Abstraction Network (CAN) to SNOMED CT's *Physical object* hierarchy, uncovering errors in its content.

2.3 Additional Related Work

2.3.1 Ontology Summarization

Abstraction networks produce a structural summary and compact visualization of an ontology. Ontology summarization is similar to the processes of ontology modularization and partitioning [70]. The goal of ontology summarization is to assist users in understanding the content of an ontology. Summaries of ontologies are important for supporting content development, ontology reuse, and ontology usability [70]. Several ontology summarization methods have been described in the context of the Semantic Web. Li et al. [70] provide a survey of ontology summarization methods and discussed the need for ontology summarization. Several ontology summarization methods were evaluated.

Peroni et al. [71] utilized measurements of relationship density and coverage to extract key concepts from an ontology. Their approach produces a text-based list of *n* concepts which summarize an ontology's content. Unlike abstraction networks, their methodology only summarizes ontology content; it does not summarize structure. Dzbor et al. created the OntoSumViz plug-in [72] for the NeOn Toolkit [73], enabling visualization of the summaries created by the methodology described by Peroni et al. [71].

Zhang et al. [74] utilized an RDF sentence graph approach to create text-based summaries of ontologies. Their method creates “RDF sentences” from groups of related RDF statements. Sentences are organized into a graph and several centrality measurements (e.g., PageRank [75]) are applied to determine which sentences are, relatively, the most important. The most important RDF sentences are used to summarize the ontology. The summaries produced using this method are customizable; users can specify how many RDF sentences are included in the summary. Unlike Peroni et al., the methodology introduced by Zhang et al. summarizes an ontology’s structure via the RDF sentences.

Queiroz-Sousa et al. [76] describe a generic approach to ontology summarization and a method for creating personalized ontology summaries based on concept relevance. Their personalized ontology summary methodology utilizes parameters and centrality measures to identify key concepts within an ontology. The key concepts are then extracted to form an ontology summary. Additionally, Queiroz-Sousa et al. developed a tool called OWLSumBRP which enables derivation and visualization of ontology summaries.

Like the methods of Peroni et al., Zhang et al., and Queiroz-Sousa et al., many abstraction networks (e.g., the various taxonomies [22, 24, 25]) use the underlying graph structure of an ontology in the summarization process. However, Peroni et al., Zhang et al., and Queiroz-Sousa et al. each use centrality measures to extract individual key concepts. Abstraction networks, on the other hand, typically use structural information (e.g., relationship sets) to define nodes that summarize a set of similar concepts.

2.3.2 Ontology Quality Assurance

Abstraction networks do not automatically identify and fix erroneous concepts. In previous abstraction network studies, domain experts manually reviewed individual concepts for errors [22, 26-29]. Manual concept review is a time consuming process, even when limited to a small sample. Zhu et al. [77] provide a list of ontology quality factors which can be used to assess ontology content. Two examples of such quality factors are consistency, e.g., representing terms in a consistent manner, and soundness meaning the accuracy of the knowledge in the ontology. Zhu et al. also provide a comprehensive survey of manual, semi-automatic, and automatic ontology quality assurance methodologies.

The quality assurance lifecycle of National Cancer Institute thesaurus (NCIt) is discussed by de Coronado et al. [78]. NCIt quality assurance reviews are conducted during the development and release phases of the ontology's release cycle. Additional quality assurance reviews are conducted periodically to address errors reported by the users of the ontology. The SNOMED CT User Guide [38] gives a high level explanation of the quality assurance methods utilized by the IHTSDO for SNOMED CT.

Verspoor et al. [79] developed an automatic lexical transformation method to cluster together lexically similar Gene Ontology (GO) concepts. Clusters were analyzed to find redundant terms. They found that GO's content is generally high in quality, but still found 67 redundant terms in their study.

Due to its size, complexity, and importance, SNOMED CT is a common target for ontology quality assurance studies. Agrawal et al. [80-82] utilized a combination of lexical and structural techniques to identify inconsistently modeled concepts in SNOMED CT. Lexically similar concepts that had different relationship structures were found to have a relatively high error rate [82]. Concepts with long fully specified names and many parents were also found to contain relatively many errors [81].

Ceusters et al. [83] describe an ontology-based technique that utilized an external ontology, LinkBase [84], to uncover errors in SNOMED CT. Semantic, structural, and ontological techniques are offered by Rector [19, 85] and by Schulz [18, 86, 87] for quality assurance of the SNOMED CT terminology. Rector et al. [19] identified seven major types of errors in SNOMED CT that were caused by problems in Description Logic modeling or in the concept classification process. Their approach consisted of reviewing hierarchies of SNOMED CT concepts. The review would start from a given concept and then all of its ancestors and descendants were reviewed. Most of the identifier errors were in the parents (or some higher ancestor). When an issue was uncovered during this manual review process it was further analyzed to determine the cause of the error.

Rector et al. [85] used lexical and semantic techniques to analyze the correctness of post coordination with qualifier values that involve “acute” and “chronic,” e.g., *Acute disease* and *Chronic disease*. They utilized the pre-coordinated terms of SNOMED CT which start with “chronic” or “acute” as a sample set. A large number of misclassifications (44%) were identified among this sample.

Schulz et al. [86] utilized an ontology-based method to analyze the correctness of relationship groups in SNOMED CT. Schulz et al. [18] also analyzed SNOMED CT's "health" from an ontological perspective and a logical perspective. They identified eight major problems, such as taxonomic dystrophy (problems with the hierarchy, e.g., relatively too many concepts with multiple parents) and relationship idiosyncrasies (poorly defined relationships).

Mortensen et al. [88] described a crowdsourcing [89] methodology to verify the correctness of axioms in an ontology. In their method, axioms (e.g., relationships) from an ontology were transformed into English sentences. Members of the crowd then determined if a given sentence is correct. Many members of the crowd were given the same sentence and statistical analysis was used to determine if a given axiom was incorrect, according to the number of crowd members who said the sentence was false. Using this approach they were able to replicate Rector et al.'s [19] results with 85% accuracy. The goal of their research is to enable large scale ontology quality assurance using the "knowledge of the crowd." In [90] Mortensen et al. found that the crowd can perform nearly as well as a panel of domain experts.

Quality assurance techniques can be combined with abstraction networks to enable more efficient ontology quality assurance. For example, applying a given quality assurance technique, say a lexical method or crowdsourcing method, to concepts in small partial-areas is expected to uncover more errors than applying the same technique to a random sample of concepts.

2.3.3 Ontology Software

It would not be possible to create, manage, and browse ontologies without well developed software tools. Ontology software can be broken into two major categories: development tools and browsers. Ontology development tools allow a user to create and modify ontologies. Browsers allow a user to search and explore ontology content (e.g., concepts, relationships).

One of the most popular and widely used development tools is Protégé [56], which was developed by Musen et al. in the late 1980s and is maintained by Stanford University. The Protégé community currently has over 240,000 members. Protégé enables the development of OWL ontologies. In Protégé a user can create and edit any aspect of an ontology, including classes, relationships, and attributes. Protégé is extendable via plugins that add functionality to the tool. Some examples of plugins include Description Logic reasoners, like HermiT [67], and OWLDiff [91], for comparing two ontologies.

WebProtégé [92] is an open source, web-based collaborative ontology development tool backed by Protégé. Swoop [93] is a similar web-based development tool. OBO-Edit [94] was created to support the development of OBO Foundry [52] ontologies. The IHTSDO Workbench [95] is a collaborative development tool used to create and manage SNOMED CT's content.

Many ontology users are only interested in viewing an ontology's content. These users do not need the editing functionality provided by development tools. *Ontology browsers* provide a simple user interface for navigating and viewing ontology content.

Ontology development tools can function as browsers. For example, Protégé is typically used to browse OWL ontology content.

Bodenreider et al. [96] provide a survey of features available in over a dozen SNOMED CT browsers. Two examples of SNOMED CT browsers are CliniClue [57] and Snow Owl [97]. The National Library of Medicine's UMLS Terminological Services (UTS) web site offers web-based browsers for SNOMED CT and the UMLS [58].

BioPortal [50] includes a web-based browser for the ontologies hosted in their repository. The Neighborhood Auditing Tool (NAT) [98] is a hybrid text-diagram browser for the UMLS. The Relationship, Audit Set, and Concept NAT (RAC-NAT) [99] extended the NAT by adding a relationship-centric browser, among other features. Many browsing tools are available for the Gene Ontology (GO) [42], e.g., AmiGO and QuickGO [100]. A partial list of publicly available GO browsers has been published [101].

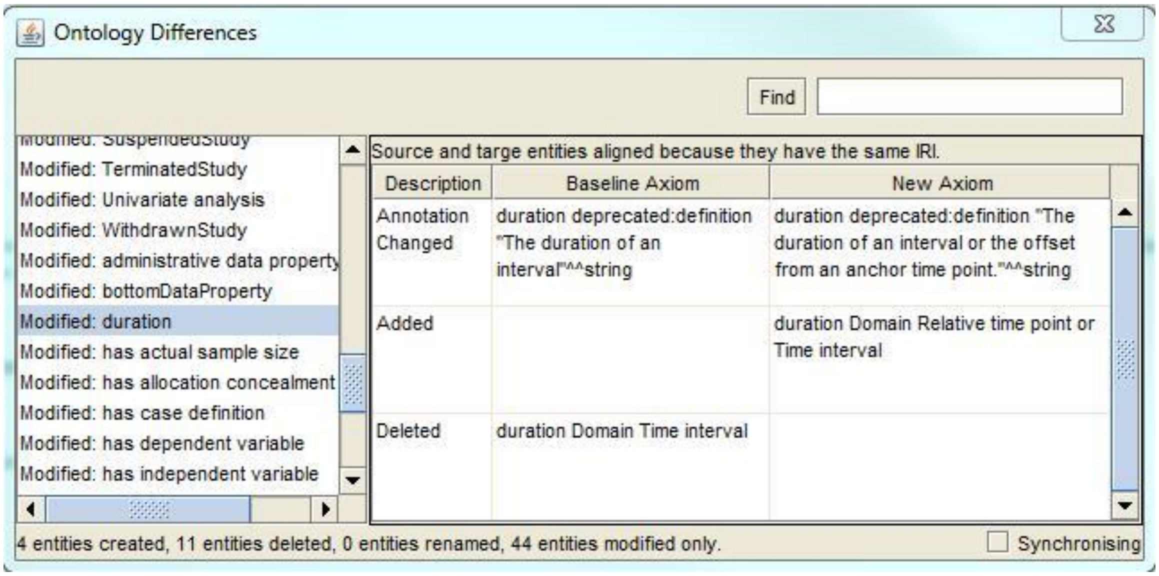
2.3.4 Ontology Visualization

One of the important features of an abstraction network is its ability to provide a compact visualization of an ontology's content. The problem of ontology visualization can be considered a subproblem of graph visualization in general. Katifori et al. [102] provide a comprehensive survey of ontology visualization techniques across several dozen ontology tools. Katifori et al. [103] performed a comparative analysis on four different ontology visualization schemes available in Protégé. Storey et al. [104] performed a similar study on two ontology visualization methods available in Protégé. Lanzenberger et al. [105] surveyed various ontology visualization techniques. Fu et al. [106] studied the usability of two ontology visualization techniques.

Most ontology visualization schemes fall into one of two major schemes: indented hierarchies and Node-link (box and arrow) diagrams. An indented hierarchy shows an ontology's content similar to a file system browser, e.g., Windows Explorer. Child concepts are shown under their parent(s) and indented to the right. Other ancestors and descendants can be viewed by expanding individual concepts in the hierarchy. Most development tools and browsers display an ontology using an indented hierarchy. Node-link diagrams show the ontology as a graph of labeled nodes, which represent concepts, and labeled edges, which represent relationships between concepts. Various tools exist for viewing an ontology as a node-link diagram. Some examples includes GraphViz [107], Jambalaya [108], and OntoSphere [109]. FlexViz [110] is a web-based ontology visualization tool used in BioPortal. BioMixer [111] is a web-based collaborative ontology visualization tool.

2.3.5 Ontology Diff

A "diff" is a comparison method that identifies the differences between two versions of a file. Difference detection is important for tracking content evolution and version control. Hunt and McIlroy [112] developed the *diff* utility for detecting differences between text files. However, the textual diff approach generally does not work well for identifying structural changes between ontology versions. The OWL [40] and OBO [52] formats do not define an ordering of ontological elements, thus, the same ontology can be defined using two or more different textual representations. Noy et al. [113] discuss the importance of detecting changes during ontology evolution.



Modified: SuspendedStudy
Modified: TerminatedStudy
Modified: Univariate analysis
Modified: WithdrawnStudy
Modified: administrative data property
Modified: bottomDataProperty
Modified: duration
Modified: has actual sample size
Modified: has allocation concealment
Modified: has case definition
Modified: has dependent variable
Modified: has independent variable

Source and targe entities aligned because they have the same IRI.

Description	Baseline Axiom	New Axiom
Annotation Changed	duration deprecated:definition "The duration of an interval"^^string	duration deprecated:definition "The duration of an interval or the offset from an anchor time point."^^string
Added		duration Domain Relative time point or Time interval
Deleted	duration Domain Time interval	

4 entities created, 11 entities deleted, 0 entities renamed, 44 entities modified only.

Figure 2.13 An example of an ontology diff taken from Protégé's "Compare Ontologies" tool, with the modified object property *duration* selected.

To overcome this problem, various *structural diff* approaches have been developed. Instead of identifying the textual changes in OWL files, a structural diff identifies individual axiom changes between two ontology versions. Noy and Musen [114] developed PromptDiff, a fixed point algorithm that uses heuristic matchers to compare the axioms of two ontologies. Kremen et al. [91] developed OWLDiff, an open source application for comparing OWL ontologies. Jiménez-Ruiz et al. [115] describe a structural diff approach in support of collaborative ontology development. Goncalves et al. [116] discuss Ecco, a diff tool that uses structural and semantic techniques. Redmond and Noy [117] discuss the OWL Difference Engine, an open source tool for comparing OWL ontologies.

Figure 2.13 provides an example of a structural diff created using Protégé's "Compare Ontologies" tool, which is based on the OWL Difference Engine. Entities (e.g., classes or object properties) that have been added, removed or modified are shown on the left. Clicking on an entity shows which axioms were changed. In the example of

Figure 2.13, on the right, the domain of the object property *duration* in the Ontology of Clinical Research (OCRe) changed from *Time interval* to *Relative time point* **or** *Time interval*. Additionally, an annotation associated with the object property was also changed.

CHAPTER 3

DESIGNING NOVEL ABSTRACTION NETWORKS

Several studies have been completed to address the five abstraction network research problems introduced at the end of Section 1.1. Geller et al. [118] and Ochs et al. [28, 119] have introduced several methods for creating *subtaxonomies* to enable quality assurance of large SNOMED CT hierarchies. Ochs et al. [27] introduced the *Tribal Abstraction Network* (TAN) to summarize the content of SNOMED CT hierarchies which have no attribute relationships. Several different abstraction network derivation methodologies for OWL and OBO ontologies are described by Ochs et al. [25, 120]. Diff Abstraction Networks, which summarize and visualize structural differences between two ontology releases, were introduced by Ochs et al. [121]. Finally, the Biomedical Layout Utility for SNOMED CT (BLUSNO), a software tool for deriving and visualizing SNOMED CT abstraction networks, was introduced by Geller et al. [118]. The results of these studies, and additional results, will now be presented in detail.

3.1 Subtaxonomies for Large SNOMED CT Hierarchies

The amount of knowledge represented in different SNOMED CT hierarchies varies greatly. For example, in the January 2013 release of SNOMED CT, the *Procedure* and *Clinical finding* hierarchies contain 52,284 and 98,544 concepts, respectively. This is in contrast to the *Specimen* and *Event* hierarchies, which have only 1,329 and 3,661 concepts, respectively. The number of concepts in a hierarchy affects the applicability of the previously developed partial-area-taxonomy-based quality assurance methodologies. Additionally, the large number of attribute relationship types defined for certain hierarchies results in a large number of areas, causing a growth in partial-area taxonomy size.

Specimen has only five different relationship types (e.g., *Topography* and *Morphology*), whereas *Procedure* has 28 (e.g., *Method* and *Procedure site*). With an order of magnitude increase in hierarchy size and number of relationship types, partial-area taxonomies tend to lose their compactness and, hence, their effectiveness from a summarization and quality assurance standpoint; e.g., *Procedure*'s partial-area taxonomy has over 10,000 partial-areas.

Table 3.1 Taxonomy Metrics for Seven of SNOMED CT's Hierarchies (January 2013 release)

Hierarchy	**# of Concepts**	**# of Relationships**	**# of Areas**	**# of Partial-areas**
Body Structure	31,117	1	2	23
Clinical Finding	99,440	14	357	10,614
Event	3,662	4	7	31
Pharmaceutical / Biologic Product	17,135	2	4	8,546
Procedure	53,147	28	739	10,828
Situation	3,350	6	9	865
Specimen	1,422	5	22	419

Source: [28]

Table 3.1 shows the number of concepts, relationships, areas, and partial-areas for the seven SNOMED CT hierarchies with attribute relationships. The number of areas in a taxonomy is dependent on (a) the number of concepts in the hierarchy, (b) the number of relationship types defined for the hierarchy, and (c) the combinations of relationships appearing at actual concepts. The partial-area taxonomy of the *Specimen* hierarchy (with five relationships and a total of 1,329 concepts) has only 22 areas and 409 partial-areas.

In the case of *Procedure*, with 52,284 concepts and 28 types of relationships, the partial-area taxonomy has 735 areas and 10,621 partial-areas.

To address this, methodologies for creating s*ubtaxonomies*, compact subsets of taxonomies, to partition taxonomies for large SNOMED CT hierarchies into more manageable subsets, are necessary. This subtaxonomy approach offers scalability of the previously developed taxonomy-based quality assurance regimen to large hierarchies, to which it was previously inapplicable. Several kinds of subtaxonomies will now be discussed.

3.1.1 Relationship-constrained Partial-area Subtaxonomy

Relationship-constrained area subtaxonomies and *relationship-constrained partial-area subtaxonomies* (*relationship subtaxonomies* for short, when there is no ambiguity), first introduced by Geller et al. [118], are defined as taxonomies generated using a subset of the outgoing attribute relationships (relationships for short) in a SNOMED CT hierarchy. This subtaxonomy methodology is based on the underlying relationship structure of concepts in SNOMED CT. Previously, area taxonomies were generated using the set of all relationships that exist within a given hierarchy. The relationship subtaxonomy methods allow a terminology auditor to generate areas with a chosen subset of relationships.

3.1.1.1 Derivation. To create relationship subtaxonomies, a subset of a hierarchy's defined relationships, *R'*, is chosen to derive a relationship subtaxonomy's areas. For example, assume a user is creating a partial-area taxonomy for a hierarchy of 10 relationships, R_1, R_2, …, R_{10}.

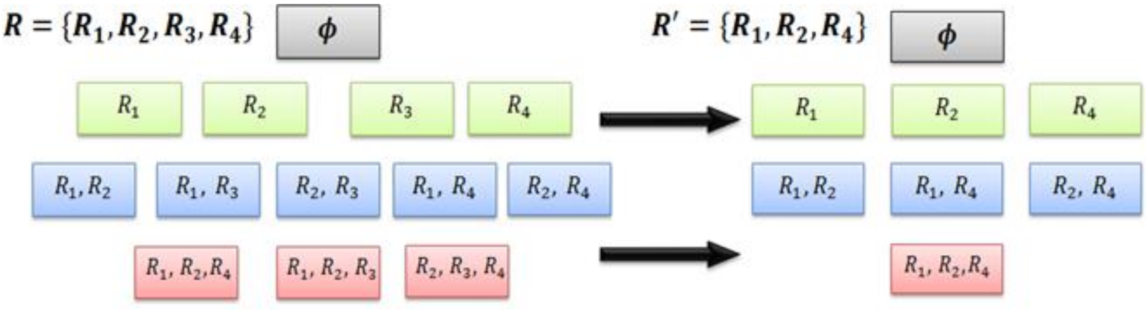


Figure 3.1 An example of deriving a relationship area subtaxonomy for a theoretical hierarchy with four relationships R_1-R_4 with R' = {R_1, R_2, R_4}.
Source: [118]

The relationship area subtaxonomy with respect to R'= {R_1, R_4, R_6, R_8} may only include areas {R_1, R_4, R_6, R_8}, {R_1, R_4, R_6}, {R_1, R_4, R_8}, {R_1, R_6, R_8}, {R_4, R_6, R_8}, {R_1, R_4}, etc. That is, only areas involving subsets of {R_1, R_4, R_6, R_8} (including the empty set, denoted by ∅) are allowed. As such, there are a maximum of $\binom{10}{4}$ (= 210) areas in the relationship area subtaxonomy. Figure 3.1 provides a visual example and illustrates the general process of deriving a relationship area subtaxonomy with four relationships R_1-R_4 and R'={R_1, R_2, R_4}. Note that only combinations that exist for concepts in the hierarchy are considered; many combinations of relationships may not exist in a hierarchy. For example, on the left side of Figure 3.1 there are no areas {R_1, R_3, R_4}, {R_1, R_2, R_3, R_4}.

Once a subset R' of relationships has been chosen, the definition of the relationship area subtaxonomy follows that of the complete area taxonomy but is restricted to the areas whose relationships are all members of R'. Because ∅ is a subset of any R', the area ∅ appears in every relationship subtaxonomy.

The definition of the relationship partial-area subtaxonomy with respect to R' also follows the definition of the "normal" partial-area taxonomy for the specific hierarchy, but is limited to the areas of the relationship area subtaxonomy for R'.

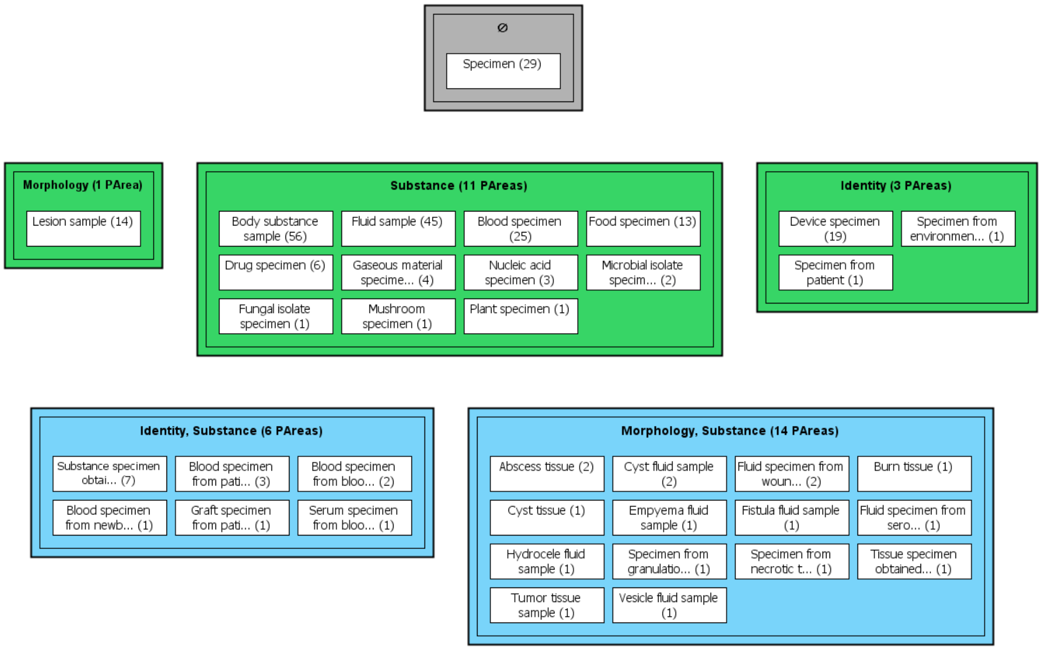

∅
Specimen (29)
Morphology (1 PArea)
Lesion sample (14)
Substance (11 PAreas)
Body substance sample (56)
Fluid sample (45)
Blood specimen (25)
Food specimen (13)
Drug specimen (6)
Gaseous material specime... (4)
Nucleic acid specimen (3)
Microbial isolate specim... (2)
Fungal isolate specimen (1)
Mushroom specimen (1)
Plant specimen (1)
Identity (3 PAreas)
Device specimen (19)
Specimen from environmen... (1)
Specimen from patient (1)
Identity, Substance (6 PAreas)
Substance specimen obtai... (7)
Blood specimen from pati... (3)
Blood specimen from bloo... (2)
Blood specimen from newb... (1)
Graft specimen from pati... (1)
Serum specimen from bloo... (1)
Morphology, Substance (14 PAreas)
Abscess tissue (2)
Cyst fluid sample (2)
Fluid specimen from woun... (2)
Burn tissue (1)
Cyst tissue (1)
Empyema fluid sample (1)
Fistula fluid sample (1)
Fluid specimen from sero... (1)
Hydrocele fluid sample (1)
Specimen from granulatio... (1)
Specimen from necrotic t... (1)
Tissue specimen obtained... (1)
Tumor tissue sample (1)
Vesicle fluid sample (1)

Figure 3.2 The relationship partial-area subtaxonomy for the *Specimen* hierarchy with *R'*={*Morphology, Substance, Identity*}.

Figure 3.2 shows the relationship partial-area subtaxonomy for the *Specimen* hierarchy with respect to the set of relationship types *R'*={*Morphology, Substance, Identity*}. This is in contrast to the complete *Specimen* partial-area taxonomy shown in Figure 2.9. The comparison of Figures 2.9 and 3.2 shows the significant reduction in size and complexity of the diagram that can be achieved by limiting the number of relationship types used to define the taxonomy.

While there is, by definition, only one possible area taxonomy and partial-area taxonomy for a hierarchy when using all of the hierarchy's relationship types, there are many possible subtaxonomies. Each relationship subtaxonomy is dependent on the selection *R'*. Therefore, one can select different subsets of relationship types to focus on portions of a hierarchy most relevant to a specific need. Since certain combinations of relationship types are not meaningful, and thus do not appear in any concepts, selecting an *R'* with such a combination of relationship types will result in a relationship subtaxonomy which contains only the root area ∅.

3.1.1.2 In Support of Quality Assurance. Quality assurance for large and highly complex hierarchies, such as *Procedure* and *Clinical finding*, is very difficult. The previously developed taxonomy-based quality assurance methodology described by Halper et al. [26], which entails auditing all concepts which belong to small partial-areas, is not possible for these large hierarchies. Using the definition of small described below (partial-areas with three or fewer concepts) there are 9,359 (9,236/10,621=88.1%) small partial-areas in the complete *Procedure* partial-area taxonomy, encompassing 11,239 concepts (11,239/52,284=21.5% of the hierarchy). This is still far too much information to process effectively. In other words, previously developed taxonomy-based quality assurance strategies do not scale to large hierarchies.

Relationship subtaxonomies allow an auditor to focus on a manageable subset of concepts. Additionally, subtaxonomies support the partitioning of collections of a large hierarchy's concepts into groups such that some groups comprise concepts expected to have a higher likelihood of errors and inconsistencies—thus, further helping to focus the efforts and increase the effectiveness of quality assurance personnel. As in Min et al. and Halper et al. [22, 26], this process entails separating concepts into two groups: those that belong to "small" partial-areas and those that belong to "large" partial-areas. Based the findings of Halper et al. [26], the following hypothesis emerges:

> *Hypothesis*: In a relationship subtaxonomy of a large SNOMED CT hierarchy, small partial-areas have higher error concentrations than large partial-areas.

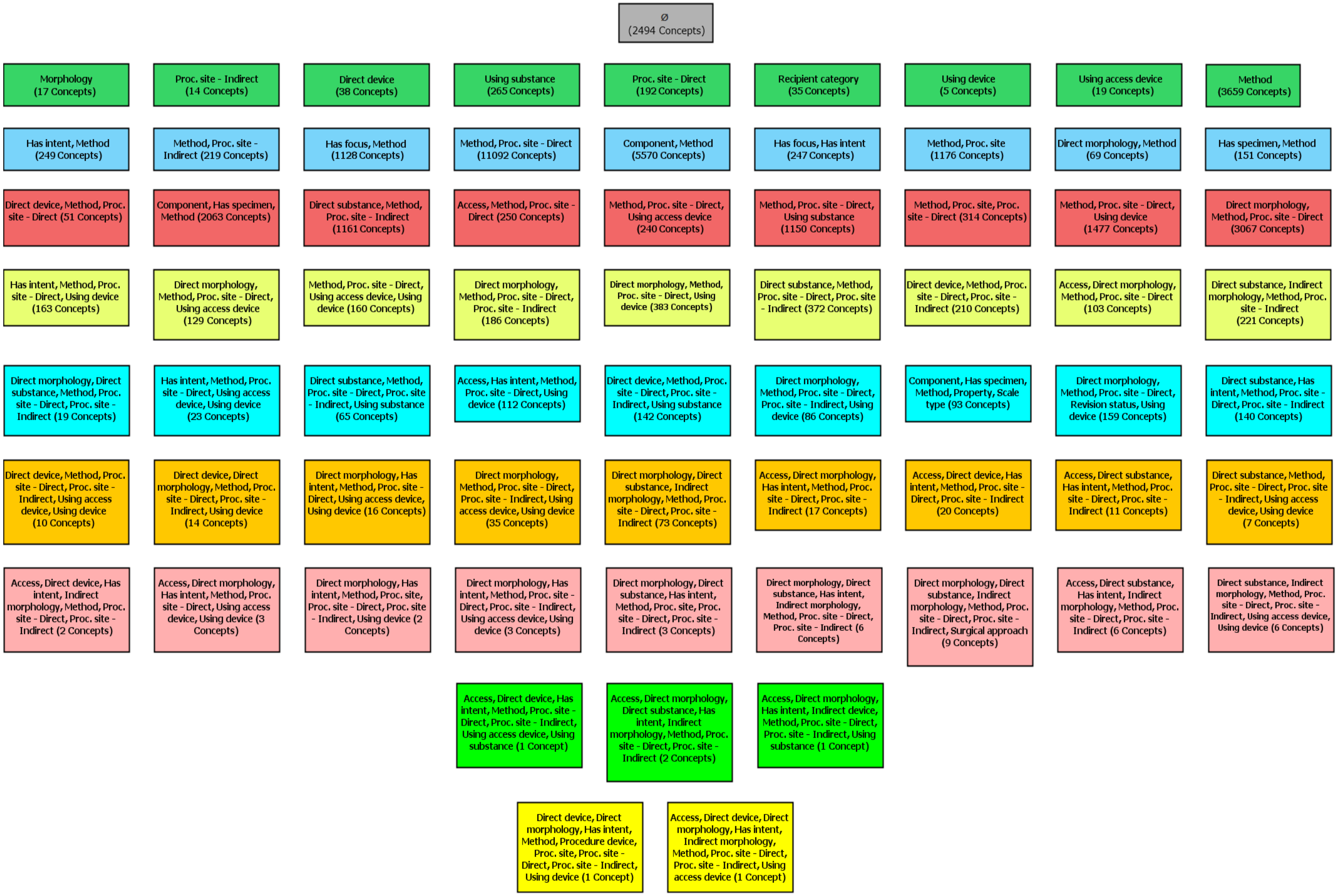

ø (2494 Concepts)
Morphology (17 Concepts)
Proc. site - Indirect (14 Concepts)
Direct device (38 Concepts)
Using substance (265 Concepts)
Proc. site - Direct (192 Concepts)
Recipient category (35 Concepts)
Using device (5 Concepts)
Using access device (19 Concepts)
Method (3659 Concepts)
Has intent, Method (249 Concepts)
Method, Proc. site - Indirect (219 Concepts)
Has focus, Method (1128 Concepts)
Method, Proc. site - Direct (11092 Concepts)
Component, Method (5570 Concepts)
Has focus, Has intent (247 Concepts)
Method, Proc. site (1176 Concepts)
Direct morphology, Method (69 Concepts)
Has specimen, Method (151 Concepts)
Direct device, Method, Proc. site - Direct (51 Concepts)
Component, Has specimen, Method (2063 Concepts)
Direct substance, Method, Proc. site - Indirect (1161 Concepts)
Access, Method, Proc. site - Direct (250 Concepts)
Method, Proc. site - Direct, Using access device (240 Concepts)
Method, Proc. site - Direct, Using substance (1150 Concepts)
Method, Proc. site, Proc. site - Direct (314 Concepts)
Method, Proc. site - Direct, Using device (1477 Concepts)
Direct morphology, Method, Proc. site - Direct (3067 Concepts)
Has intent, Method, Proc. site - Direct, Using device (163 Concepts)
Direct morphology, Method, Proc. site - Direct, Using access device (129 Concepts)
Method, Proc. site - Direct, Using access device, Using device (160 Concepts)
Direct morphology, Method, Proc. site - Direct, Proc. site - Indirect (186 Concepts)
Direct morphology, Method, Proc. site - Direct, Using device (383 Concepts)
Direct substance, Method, Proc. site - Direct, Proc. site - Indirect (372 Concepts)
Direct device, Method, Proc. site - Direct, Proc. site - Indirect (210 Concepts)
Access, Direct morphology, Method, Proc. site - Direct (103 Concepts)
Direct substance, Indirect morphology, Method, Proc. site - Indirect (221 Concepts)
Direct morphology, Direct substance, Method, Proc. site - Direct, Proc. site - Indirect (19 Concepts)
Has intent, Method, Proc. site - Direct, Using access device, Using device (23 Concepts)
Direct substance, Method, Proc. site - Direct, Proc. site - Indirect, Using substance (65 Concepts)
Access, Has intent, Method, Proc. site - Direct, Using device (112 Concepts)
Direct device, Method, Proc. site - Direct, Proc. site - Indirect, Using substance (142 Concepts)
Direct morphology, Method, Proc. site - Direct, Proc. site - Indirect, Using device (86 Concepts)
Component, Has specimen, Method, Property, Scale type (93 Concepts)
Direct morphology, Method, Proc. site - Direct, Revision status, Using device (159 Concepts)
Direct substance, Has intent, Method, Proc. site - Direct, Proc. site - Indirect (140 Concepts)
Direct device, Method, Proc. site - Direct, Proc. site - Indirect, Using access device, Using device (10 Concepts)
Direct device, Direct morphology, Method, Proc. site - Direct, Proc. site - Indirect, Using device (14 Concepts)
Direct morphology, Has intent, Method, Proc. site - Direct, Using access device, Using device (16 Concepts)
Direct morphology, Method, Proc. site - Direct, Proc. site - Indirect, Using access device, Using device (35 Concepts)
Direct morphology, Direct substance, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect (73 Concepts)
Access, Direct morphology, Has intent, Method, Proc. site - Direct, Proc. site - Indirect (17 Concepts)
Access, Direct device, Has intent, Method, Proc. site - Direct, Proc. site - Indirect (20 Concepts)
Access, Direct substance, Has intent, Method, Proc. site - Direct, Proc. site - Indirect (11 Concepts)
Direct substance, Method, Proc. site - Direct, Proc. site - Indirect, Using access device, Using device (7 Concepts)
Access, Direct device, Has intent, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect (2 Concepts)
Access, Direct morphology, Has intent, Method, Proc. site - Direct, Using access device, Using device (3 Concepts)
Direct morphology, Has intent, Method, Proc. site, Proc. site - Direct, Proc. site - Indirect, Using device (2 Concepts)
Direct morphology, Has intent, Method, Proc. site - Direct, Proc. site - Indirect, Using access device, Using device (3 Concepts)
Direct morphology, Direct substance, Has intent, Method, Proc. site, Proc. site - Direct, Proc. site - Indirect (3 Concepts)
Direct morphology, Direct substance, Has intent, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect (6 Concepts)
Direct morphology, Direct substance, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect, Surgical approach (9 Concepts)
Access, Direct substance, Has intent, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect (6 Concepts)
Direct substance, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect, Using access device, Using device (6 Concepts)
Access, Direct device, Has intent, Method, Proc. site - Direct, Proc. site - Indirect, Using access device, Using substance (1 Concept)
Access, Direct morphology, Direct substance, Has intent, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect (2 Concepts)
Access, Direct morphology, Has intent, Indirect device, Method, Proc. site - Direct, Proc. site - Indirect, Using substance (1 Concept)
Direct device, Direct morphology, Has intent, Method, Procedure device, Proc. site, Proc. site - Direct, Proc. site - Indirect, Using device (1 Concept)
Access, Direct device, Direct morphology, Has intent, Indirect morphology, Method, Proc. site - Direct, Proc. site - Indirect, Using access device (1 Concept)

Figure 3.3 A small portion (69 areas) of the ten levels of the Procedure hierarchy’s area taxonomy. *Source: [28]*

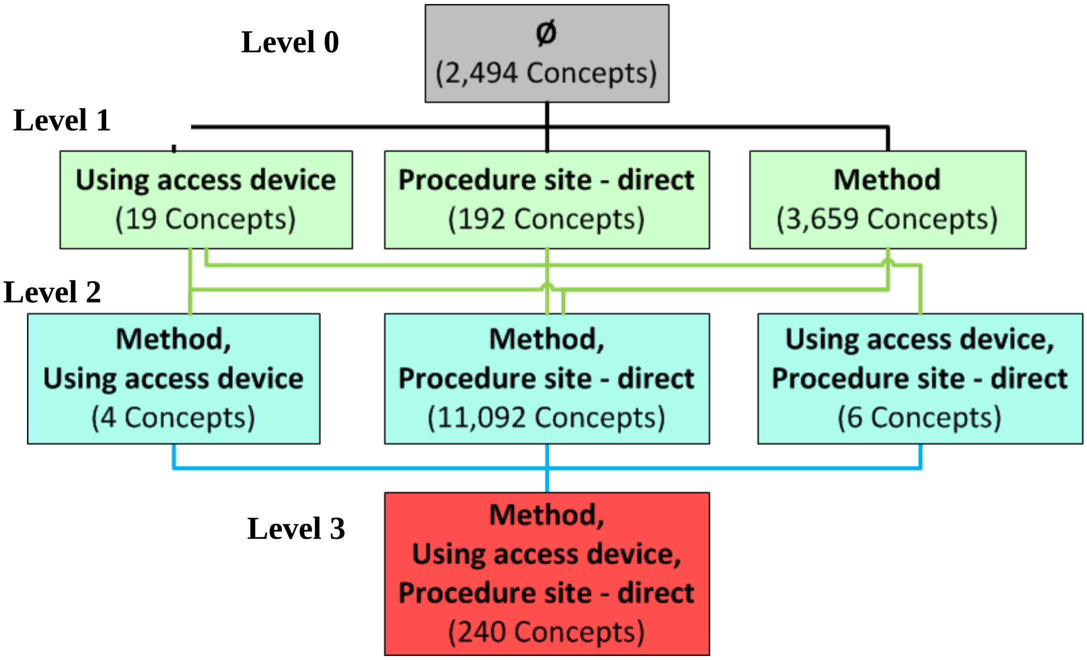


Figure 3.4 Area sub-taxonomy with respect to the three relationships using access device, procedure site – direct, and method.
Source: [28]

This hypothesis was investigated using a relationship subtaxonomy for the *Procedure* hierarchy. The complete *Procedure* taxonomy consists of over 10,000 partial-areas separated into 735 areas. Figure 3.3 shows a small portion (69 areas) of *Procedure*'s complete area taxonomy. At the scale of Figure 3.3 the entire area taxonomy would span 23 pages. Worse yet, the complete *Procedure* partial-area taxonomy at the scale of Figure 3.2 would be over 100 pages wide by four pages high.

A domain expert chose *R'*={*Method*, *Procedure site – direct*, *Using access device*}, resulting in a relationship subtaxonomy with eight areas (shown in Figure 3.4). When comparing Figure 3.3 to Figure 3.4, it is easy to see that the chosen relationship subtaxonomy is much more manageable than the complete taxonomy. Figures 3.5 and 3.6 show the associated relationship partial-area subtaxonomy, with the largest area {*Method*, *Procedure site – direct*} shown without partial-areas to save space in Figure 3.5.

In Level 1, the taxonomy has three areas, 104 partial-areas, and 3,870 concepts. The total number of concepts in the relationship subtaxonomy, namely, 17,706 (covering 34% of *Procedure*) is still overwhelming. Figure 3.4 shows the *child-of* links, between areas, using the same color as the parent area. The largest level is Level 2, containing the largest area {*Method, Procedure site – direct*} with 11,092 concepts. The fact that {*Method, Procedure site – direct*} is so large is not surprising; there are a very large number of methods for procedures and a large number of body sites. This large area is obtained when these multiplicities are combined. The second largest level is Level 1, mainly due to the area {*Method*}. This is followed by Level 0 with the area ∅.

Figures 3.5 and 3.6 show the relationship subtaxonomy for the selected *R'*. The partial-areas of this relationship subtaxonomy were separated into small and large according to their numbers of concepts, with the hypothesis being that errors appear in higher concentrations in the small partial-areas than they do in the larger ones. To test this hypothesis, a domain expert reviewed all of the concepts of two areas, {*Procedure site – direct*} (green) with 192 concepts and {*Method, Using access device, Procedure site – direct*} (red) with 240 concepts. One large partial-area *Neck excision* (118) from {*Method, Procedure site - direct*} (blue) was also audited. In total, 550 concepts from the relationship subtaxonomy were individually reviewed for errors and inconsistencies by the domain expert. The sample concepts were provided in alphabetical order and the auditor was blind to the methodology and hypothesis. The green and red areas that were selected have a few medium-sized partial-areas and many small ones.

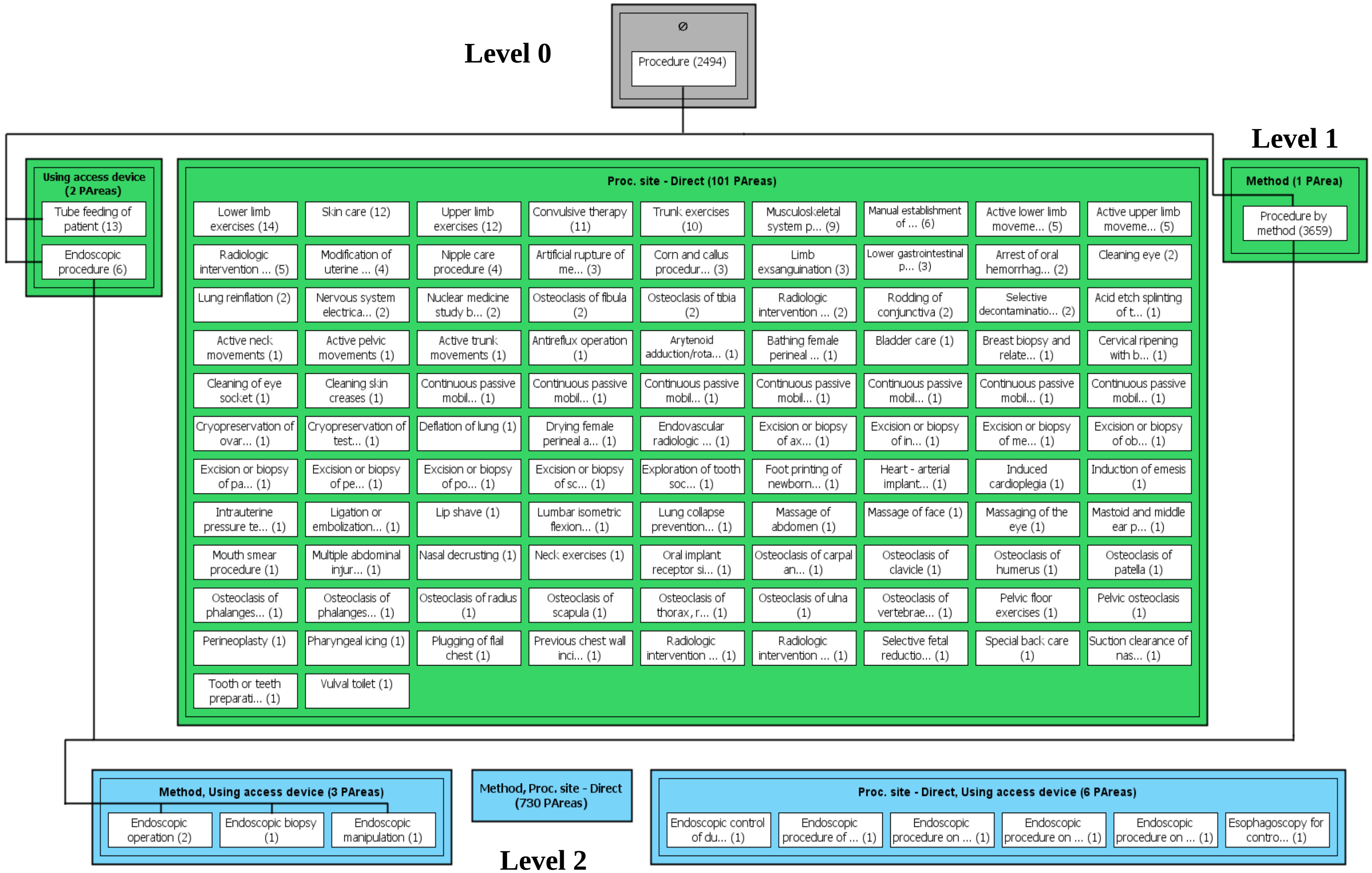

Level 0
∅
Procedure (2494)
Level 1
Using access device (2 PAreas)
Tube feeding of patient (13)
Endoscopic procedure (6)
Proc. site - Direct (101 PAreas)
Lower limb exercises (14)
Skin care (12)
Upper limb exercises (12)
Convulsive therapy (11)
Trunk exercises (10)
Musculoskeletal system p... (9)
Manual establishment of ... (6)
Active lower limb moveme... (5)
Active upper limb moveme... (5)
Radiologic intervention ... (5)
Modification of uterine ... (4)
Nipple care procedure (4)
Artificial rupture of me... (3)
Corn and callus procedur... (3)
Limb exsanguination (3)
Lower gastrointestinal p... (3)
Arrest of oral hemorrhag... (2)
Cleaning eye (2)
Lung reinflation (2)
Nervous system electrica... (2)
Nuclear medicine study b... (2)
Osteoclasis of fibula (2)
Osteoclasis of tibia (2)
Radiologic intervention ... (2)
Rodding of conjunctiva (2)
Selective decontaminatio... (2)
Acid etch splinting of t... (1)
Active neck movements (1)
Active pelvic movements (1)
Active trunk movements (1)
Antireflux operation (1)
Arytenoid adduction/rota... (1)
Bathing female perineal ... (1)
Bladder care (1)
Breast biopsy and relate... (1)
Cervical ripening with b... (1)
Cleaning of eye socket (1)
Cleaning skin creases (1)
Continuous passive mobil... (1)
Continuous passive mobil... (1)
Continuous passive mobil... (1)
Continuous passive mobil... (1)
Continuous passive mobil... (1)
Continuous passive mobil... (1)
Continuous passive mobil... (1)
Cryopreservation of ovar... (1)
Cryopreservation of test... (1)
Deflation of lung (1)
Drying female perineal a... (1)
Endovascular radiologic ... (1)
Excision or biopsy of ax... (1)
Excision or biopsy of in... (1)
Excision or biopsy of me... (1)
Excision or biopsy of ob... (1)
Excision or biopsy of pa... (1)
Excision or biopsy of pe... (1)
Excision or biopsy of po... (1)
Excision or biopsy of sc... (1)
Exploration of tooth soc... (1)
Foot printing of newborn... (1)
Heart - arterial implant... (1)
Induced cardioplegia (1)
Induction of emesis (1)
Intrauterine pressure te... (1)
Ligation or embolization... (1)
Lip shave (1)
Lumbar isometric flexion... (1)
Lung collapse prevention... (1)
Massage of abdomen (1)
Massage of face (1)
Massaging of the eye (1)
Mastoid and middle ear p... (1)
Mouth smear procedure (1)
Multiple abdominal injur... (1)
Nasal decrusting (1)
Neck exercises (1)
Oral implant receptor si... (1)
Osteoclasis of carpal an... (1)
Osteoclasis of clavicle (1)
Osteoclasis of humerus (1)
Osteoclasis of patella (1)
Osteoclasis of phalanges... (1)
Osteoclasis of phalanges... (1)
Osteoclasis of radius (1)
Osteoclasis of scapula (1)
Osteoclasis of thorax, r... (1)
Osteoclasis of ulna (1)
Osteoclasis of vertebrae... (1)
Pelvic floor exercises (1)
Pelvic osteoclasis (1)
Perineoplasty (1)
Pharyngeal icing (1)
Plugging of flail chest (1)
Previous chest wall inci... (1)
Radiologic intervention ... (1)
Radiologic intervention ... (1)
Selective fetal reductio... (1)
Special back care (1)
Suction clearance of nas... (1)
Tooth or teeth preparati... (1)
Vulval toilet (1)
Method (1 PArea)
Procedure by method (3659)
Method, Using access device (3 PAreas)
Endoscopic operation (2)
Endoscopic biopsy (1)
Endoscopic manipulation (1)
Method, Proc. site - Direct (730 PAreas)
Level 2
Proc. site - Direct, Using access device (6 PAreas)
Endoscopic control of du... (1)
Endoscopic procedure of ... (1)
Endoscopic procedure on ... (1)
Endoscopic procedure on ... (1)
Endoscopic procedure on ... (1)
Esophagoscopy for contro... (1)

Figure 3.5 Levels 0 and 1 and part of Level 2 of the partial-area sub-taxonomy with respect to the three selected relationships.
Source: [28]

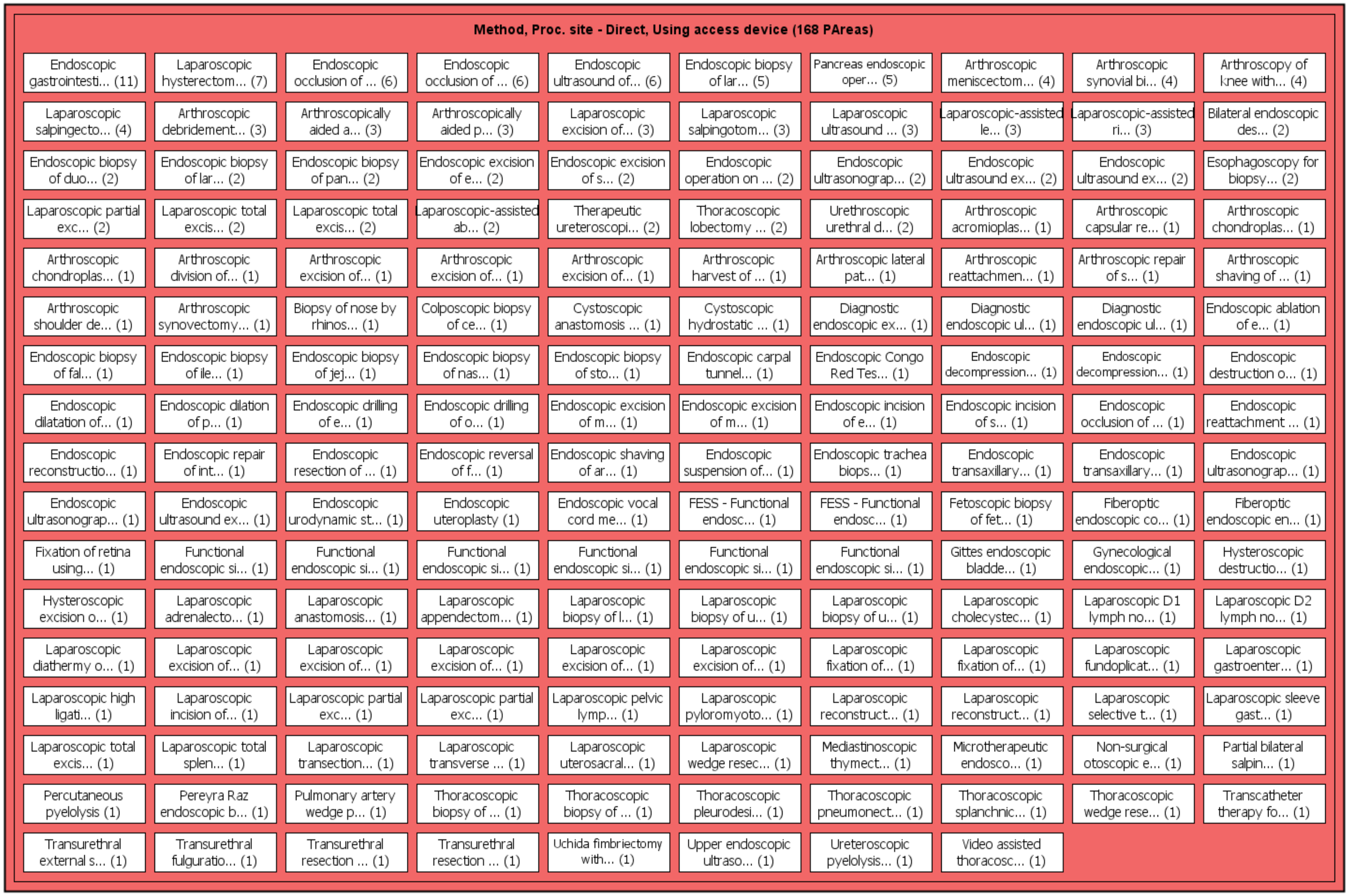

Method, Proc. site - Direct, Using access device (168 PAreas)
Endoscopic gastrointesti... (11)
Laparoscopic hysterectom... (7)
Endoscopic occlusion of ... (6)
Endoscopic occlusion of ... (6)
Endoscopic ultrasound of... (6)
Endoscopic biopsy of lar... (5)
Pancreas endoscopic oper... (5)
Arthroscopic meniscectom... (4)
Arthroscopic synovial bi... (4)
Arthroscopy of knee with... (4)
Laparoscopic salpingecto... (4)
Arthroscopic debridement... (3)
Arthroscopically aided a... (3)
Arthroscopically aided p... (3)
Laparoscopic excision of... (3)
Laparoscopic salpingotom... (3)
Laparoscopic ultrasound ... (3)
Laparoscopic-assisted le... (3)
Laparoscopic-assisted ri... (3)
Bilateral endoscopic des... (2)
Endoscopic biopsy of duo... (2)
Endoscopic biopsy of lar... (2)
Endoscopic biopsy of pan... (2)
Endoscopic excision of e... (2)
Endoscopic excision of s... (2)
Endoscopic operation on ... (2)
Endoscopic ultrasonograp... (2)
Endoscopic ultrasound ex... (2)
Endoscopic ultrasound ex... (2)
Esophagoscopy for biopsy... (2)
Laparoscopic partial exc... (2)
Laparoscopic total excis... (2)
Laparoscopic total excis... (2)
Laparoscopic-assisted ab... (2)
Therapeutic ureteroscopi... (2)
Thoracoscopic lobectomy ... (2)
Urethroscopic urethral d... (2)
Arthroscopic acromioplas... (1)
Arthroscopic capsular re... (1)
Arthroscopic chondroplas... (1)
Arthroscopic chondroplas... (1)
Arthroscopic division of... (1)
Arthroscopic excision of... (1)
Arthroscopic excision of... (1)
Arthroscopic excision of... (1)
Arthroscopic harvest of ... (1)
Arthroscopic lateral pat... (1)
Arthroscopic reattachmen... (1)
Arthroscopic repair of s... (1)
Arthroscopic shaving of ... (1)
Arthroscopic shoulder de... (1)
Arthroscopic synovectomy... (1)
Biopsy of nose by rhinos... (1)
Colposcopic biopsy of ce... (1)
Cystoscopic anastomosis ... (1)
Cystoscopic hydrostatic ... (1)
Diagnostic endoscopic ex... (1)
Diagnostic endoscopic ul... (1)
Diagnostic endoscopic ul... (1)
Endoscopic ablation of e... (1)
Endoscopic biopsy of fal... (1)
Endoscopic biopsy of ile... (1)
Endoscopic biopsy of jej... (1)
Endoscopic biopsy of nas... (1)
Endoscopic biopsy of sto... (1)
Endoscopic carpal tunnel... (1)
Endoscopic Congo Red Tes... (1)
Endoscopic decompression... (1)
Endoscopic decompression... (1)
Endoscopic destruction o... (1)
Endoscopic dilatation of... (1)
Endoscopic dilation of p... (1)
Endoscopic drilling of e... (1)
Endoscopic drilling of o... (1)
Endoscopic excision of m... (1)
Endoscopic excision of m... (1)
Endoscopic incision of e... (1)
Endoscopic incision of s... (1)
Endoscopic occlusion of ... (1)
Endoscopic reattachment ... (1)
Endoscopic reconstructio... (1)
Endoscopic repair of int... (1)
Endoscopic resection of ... (1)
Endoscopic reversal of f... (1)
Endoscopic shaving of ar... (1)
Endoscopic suspension of... (1)
Endoscopic trachea biops... (1)
Endoscopic transaxillary... (1)
Endoscopic transaxillary... (1)
Endoscopic ultrasonograp... (1)
Endoscopic ultrasonograp... (1)
Endoscopic ultrasound ex... (1)
Endoscopic urodynamic st... (1)
Endoscopic uteroplasty (1)
Endoscopic vocal cord me... (1)
FESS - Functional endosc... (1)
FESS - Functional endosc... (1)
Fetoscopic biopsy of fet... (1)
Fiberoptic endoscopic co... (1)
Fiberoptic endoscopic en... (1)
Fixation of retina using... (1)
Functional endoscopic si... (1)
Functional endoscopic si... (1)
Functional endoscopic si... (1)
Functional endoscopic si... (1)
Functional endoscopic si... (1)
Functional endoscopic si... (1)
Gittes endoscopic bladde... (1)
Gynecological endoscopic... (1)
Hysteroscopic destructio... (1)
Hysteroscopic excision o... (1)
Laparoscopic adrenalecto... (1)
Laparoscopic anastomosis... (1)
Laparoscopic appendectom... (1)
Laparoscopic biopsy of l... (1)
Laparoscopic biopsy of u... (1)
Laparoscopic biopsy of u... (1)
Laparoscopic cholecystec... (1)
Laparoscopic D1 lymph no... (1)
Laparoscopic D2 lymph no... (1)
Laparoscopic diathermy o... (1)
Laparoscopic excision of... (1)
Laparoscopic excision of... (1)
Laparoscopic excision of... (1)
Laparoscopic excision of... (1)
Laparoscopic excision of... (1)
Laparoscopic fixation of... (1)
Laparoscopic fixation of... (1)
Laparoscopic fundoplicat... (1)
Laparoscopic gastroenter... (1)
Laparoscopic high ligati... (1)
Laparoscopic incision of... (1)
Laparoscopic partial exc... (1)
Laparoscopic partial exc... (1)
Laparoscopic pelvic lymp... (1)
Laparoscopic pyloromyoto... (1)
Laparoscopic reconstruct... (1)
Laparoscopic reconstruct... (1)
Laparoscopic selective t... (1)
Laparoscopic sleeve gast... (1)
Laparoscopic total excis... (1)
Laparoscopic total splen... (1)
Laparoscopic transection... (1)
Laparoscopic transverse ... (1)
Laparoscopic uterosacral... (1)
Laparoscopic wedge resec... (1)
Mediastinoscopic thymect... (1)
Microtherapeutic endosco... (1)
Non-surgical otoscopic e... (1)
Partial bilateral salpin... (1)
Percutaneous pyelolysis (1)
Pereyra Raz endoscopic b... (1)
Pulmonary artery wedge p... (1)
Thoracoscopic biopsy of ... (1)
Thoracoscopic biopsy of ... (1)
Thoracoscopic pleurodesi... (1)
Thoracoscopic pneumonect... (1)
Thoracoscopic splanchnic... (1)
Thoracoscopic wedge rese... (1)
Transcatheter therapy fo... (1)
Transurethral external s... (1)
Transurethral fulguratio... (1)
Transurethral resection ... (1)
Transurethral resection ... (1)
Uchida fimbriectomy with... (1)
Upper endoscopic ultraso... (1)
Ureteroscopic pyelolysis... (1)
Video assisted thoracosc... (1)

Figure 3.6 The only third-level area of the sub-taxonomy with respect to the relationships *Method*, *Procedure site – direct*, and *Using access device*. *Source: [28]*

The partial-area *Neck excision* with 118 concepts selected from {*Method*, *Procedure site - direct*} (Level 2) added concepts from a large partial-area. These concepts were chosen for review because they have different numbers of relationships and are from different sized partial-areas.

The inferred view of SNOMED CT (January 2011 release) was used throughout the auditing process. The focus was on errors and inconsistencies involving incorrect or missing parents or children—errors that were deemed to be most troublesome in a study of SNOMED CT users' preferences [122]. Due to their definitional role in modeling a concept, such basic errors and inconsistencies may cause additional problems with relationships due to inheritance. It should be noted that missing or incorrect child errors can be equally well interpreted as missing or incorrect parent errors for the child concept. However, the errors were reported as missing or incorrect children according to the interpretation of the domain expert.

Out of the total of 550 concepts reviewed, 67 (12.2%) were found to have at least one error by the domain expert. Table 3.2 illustrates four examples of errors found. Table 3.3 provides the distribution of errors based on partial-area size. Out of the 67 errors, the domain expert found 31 concepts with at least one incorrect or redundant parent and 27 concepts missing at least one parent. Forty-four (66% = 44/67) of the problematic concepts were found to be primitives, indicating that certain knowledge about these concepts may be missing from the terminology.

Three was chosen as a threshold between small and large because it maximized the statistical significance of error rates between small and large partial-areas (15.4% vs. 8.8% erroneous, respectively), with $p<0.019$ according to the Fisher exact 2-tailed statistical test [123]. Therefore, an auditor reviewing partial-areas of size three or less is expected to uncover more inconsistencies than if they reviewed other partial-areas. Thresholds of five and seven were also found to be statistically significant with p<0.047 and p<0.031, respectively. A threshold of seven had a slightly higher ratio of errors (1.76 = 14.4/8.2) in small partial-areas (14.4%) versus large partial-areas (8.2%) compared to a threshold of three (1.75 = 15.4/8.8). In Table 3.4, results are shown with respect to small partial-areas (1–3 concepts) and large partial-areas (4–118 concepts).

Table 3.2 Four Examples of Inconsistencies Identified by the Domain Expert in her Quality Assurance Review of the Relationship Subtaxonomy

Concept	Partial-area	Problem Type	Correction
Endoscopic Congo Red Test	Endoscopic Congo Red Test (1)	Missing parent: *Congo Red Test*	Add IS-A directed to *Congo Red Test*
Ureteroscopic pyelolysis	Ureteroscopic pyelolysis (1)	Missing parent: *ureteroscopic operation*	Add IS-A directed to *ureteroscopic operation*
Endoscopic drilling of ovary	Endoscopic drilling of ovary (1)	Incorrect parent: *cauterization of ovary*	Replace with IS-A directed to *drilling of ovary*
Convulsive therapy	Convulsive therapy (11)	Missing parent: *Therapeutic procedure*	Add IS-A directed to *Therapeutic procedure*

Source: [28]

For each of the 67 erroneous concepts identified, a follow-up review using the January 2013 release of SNOMED CT was performed. Sixty-five concepts were unchanged and still erroneous and two concepts had minor changes and were still erroneous.

Table 3.3 Summary of Errors Found in the Audit of the Sample Partial-areas of the Relationship Subtaxonomy

Partial-area Size	**# of Partial-areas**	**Total # of Concepts**	**# of Erroneous Concepts**	**% Erroneous Concepts**
118	1	118	12	10
14	1	14	1	7
12	2	24	1	4
11	2	22	1	5
10	1	10	1	10
9	1	9	0	0
7	1	7	2	29
6	4	24	2	8
5	4	20	3	15
4	6	24	1	4
3	12	36	7	19
2	26	52	7	13
1	190	190	29	15
Total:	251	550	67	12

Source: [28]

Table 3.4 Summary of Erroneous Concepts Broken Down into Small and Large Partial-areas

	# of Partial-areas	**# of Concepts**	**# of Erroneous**	**% Errors**
Small Partial-areas (1–3)	228	278	43	15.4
Large Partial-areas (4–118)	23	272	24	8.8
Total:	251	550	67	12.2

Source: [28]

Of course, there are also errors in large partial-areas (8.8% in the sample). Due to the better auditing yield as measured by the ratio of the number of errors to the number of concepts in the sample, it is recommended that an auditor start with small partial-areas, where the cumulative number of concepts is relatively small. For the large partial-areas, it is recommended that an auditor exploit previously developed strategies (e.g., review concepts in the intersections of two or more partial-areas [29]) that have been shown to increase the efficiency of terminology quality assurance efforts.

In conclusion, this study confirmed that, by utilizing a relationship subtaxonomy, previously developed taxonomy-based quality assurance methodologies can successfully be applied to large SNOMED CT hierarchies. Within a relationship subtaxonomy, concepts in small partial-areas are statistically significantly more likely to contain errors than concepts in large partial-areas. The quality assurance methodology implied is to audit all small partial-areas in a relationship subtaxonomy first. There are, in total, only 734 concepts out of 17,706 (4%) in small partial-areas (of 1–3 concepts) in the relationship subtaxonomy chosen for this study. This is a limited auditing effort expected to uncover erroneous concepts at a rate of 15.1%.

3.1.2 Root-constrained Partial-area Subtaxonomy

Relationship subtaxonomies create taxonomy subsets containing structurally similar concepts. An alternate paradigm, which results in semantically similar subsets of concepts, is called the *root-constrained partial-area subtaxonomy*, or *root subtaxonomy* for short [118]. A root subtaxonomy is a subset of a partial-area taxonomy based on the *child-of* links between partial-areas. Partial-area taxonomies have traditionally been rooted at the partial-area containing the root concept of the whole hierarchy, e.g., *Specimen* or *Procedure*, and the taxonomy contained all of the descendant partial-areas of the root, i.e., the entire taxonomy. Hierarchies of SNOMED CT have unique root concepts and there is only one partial-area in each root area.

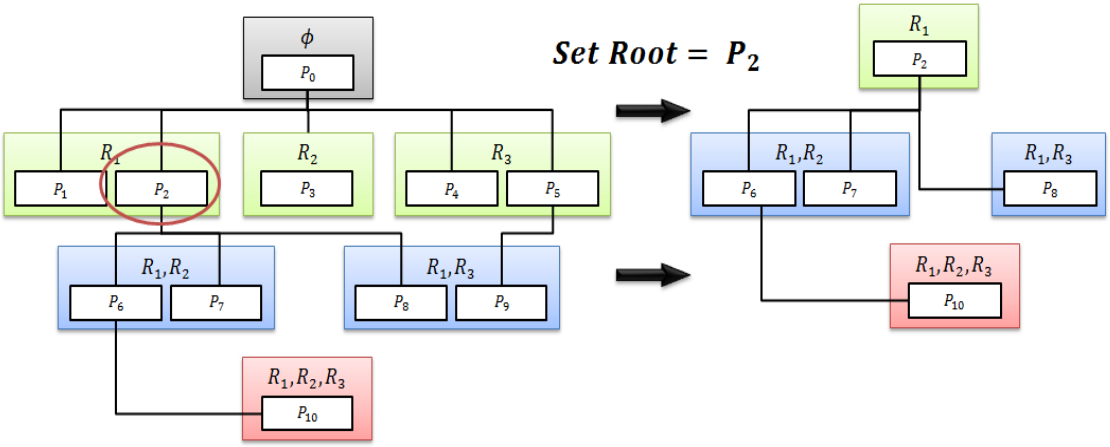


Figure 3.7 The process of deriving a root-constrained partial-area subtaxonomy. Partial-area P_2 was chosen as the root.
Source: [118]

In a root subtaxonomy one defines which partial-area is the root. The resulting root subtaxonomy consists of the selected root partial-area and all of its descendant partial-areas. The root subtaxonomy summarizes how a set of semantically-related concepts are modeled within a large SNOMED CT hierarchy.

If a partial-area is not a descendant of the chosen root (and thus, is not semantically related), it is not included in the root subtaxonomy. Figure 3.7 illustrates the general process of creating a root subtaxonomy for an example partial-area taxonomy. In this figure, partial-area P_2 is chosen as the root of the root subtaxonomy. All of P_2's descendant partial-areas are included in the root subtaxonomy.

One method of creating a root subtaxonomy is to perform a breadth-first traversal [124] of the hierarchy of inverse *child-of* links within the complete partial-area taxonomy, starting at the selected root partial-area. Partial-areas that can be reached during the traversal are considered members of the root subtaxonomy. Partial-areas with identical sets of lateral relationships are regrouped back into areas. This process is equivalent to applying the partial-area taxonomy derivation process on the chosen root partial-area's root concept and all of its descendants.

Figure 3.8 shows an example of a root subtaxonomy. Using the *Specimen* hierarchy's taxonomy (Figure 2.9), the partial-area *Lesion sample* in {*Morphology*} was selected as the root partial-area. The resulting root subtaxonomy is shown in Figure 3.8. The subtaxonomy consists of 93 concepts in 65 partial-areas, which are separated into eight areas. This root subtaxonomy summaries the major types of lesion samples in SNOMED CT.

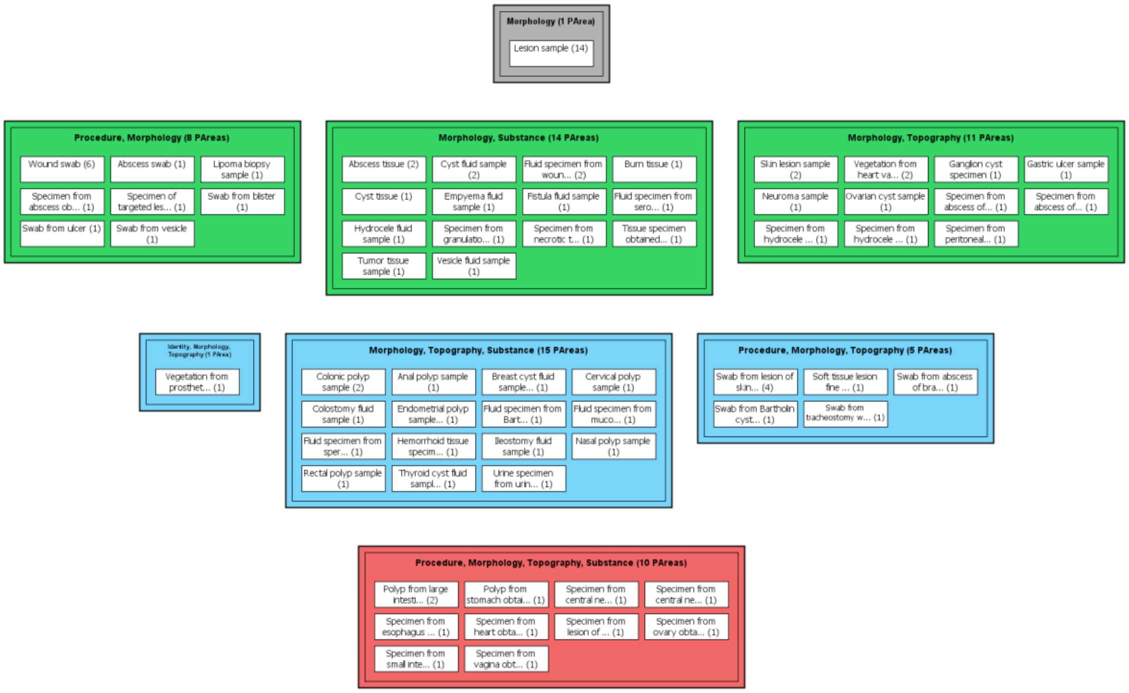

Morphology (1 PArea)
Lesion sample (14)
Procedure, Morphology (8 PAreas)
Wound swab (6)
Abscess swab (1)
Lipoma biopsy sample (1)
Specimen from abscess ob... (1)
Specimen of targeted les... (1)
Swab from blister (1)
Swab from ulcer (1)
Swab from vesicle (1)
Morphology, Substance (14 PAreas)
Abscess tissue (2)
Cyst fluid sample (2)
Fluid specimen from woun... (2)
Burn tissue (1)
Cyst tissue (1)
Empyema fluid sample (1)
Fistula fluid sample (1)
Fluid specimen from sero... (1)
Hydrocele fluid sample (1)
Specimen from granulatio... (1)
Specimen from necrotic t... (1)
Tissue specimen obtained... (1)
Tumor tissue sample (1)
Vesicle fluid sample (1)
Morphology, Topography (11 PAreas)
Skin lesion sample (2)
Vegetation from heart va... (2)
Ganglion cyst specimen (1)
Gastric ulcer sample (1)
Neuroma sample (1)
Ovarian cyst sample (1)
Specimen from abscess of... (1)
Specimen from abscess of... (1)
Specimen from hydrocele ... (1)
Specimen from hydrocele ... (1)
Specimen from peritoneal... (1)
Identity, Morphology, Topography (1 PArea)
Vegetation from prosthet... (1)
Morphology, Topography, Substance (15 PAreas)
Colonic polyp sample (2)
Anal polyp sample (1)
Breast cyst fluid sample... (1)
Cervical polyp sample (1)
Colostomy fluid sample (1)
Endometrial polyp sample... (1)
Fluid specimen from Bart... (1)
Fluid specimen from muco... (1)
Fluid specimen from sper... (1)
Hemorrhoid tissue specim... (1)
Ileostomy fluid sample (1)
Nasal polyp sample (1)
Rectal polyp sample (1)
Thyroid cyst fluid sampl... (1)
Urine specimen from urin... (1)
Procedure, Morphology, Topography (5 PAreas)
Swab from lesion of skin... (4)
Soft tissue lesion fine ... (1)
Swab from abscess of bra... (1)
Swab from Bartholin cyst... (1)
Swab from tracheostomy w... (1)
Procedure, Morphology, Topography, Substance (10 PAreas)
Polyp from large intesti... (2)
Polyp from stomach obtai... (1)
Specimen from central ne... (1)
Specimen from central ne... (1)
Specimen from esophagus ... (1)
Specimen from heart obta... (1)
Specimen from lesion of ... (1)
Specimen from ovary obta... (1)
Specimen from small inte... (1)
Specimen from vagina obt... (1)

Figure 3.8 An example of a root-constrained partial-area subtaxonomy for the *Specimen* hierarchy, with *Lesion sample* selected as the root partial-area. *Child-of* links are hidden for readability.

3.1.3 Subject Subtaxonomy

When taxonomy derivation and quality assurance methodologies were applied to large SNOMED CT hierarchies, e.g., *Procedure* and *Clinical finding* with 53,147 and 99,440 concepts, respectively, significant issues were encountered. First, the *Procedure* and

Clinical finding taxonomies contain 10,828 and 10,614 partial-areas, respectively, too many to review individually. Relationship subtaxonomies and root subtaxonomies partially addressed this issue by enabling an auditor to select a subset of partial-areas based on similar relationship structure or similar semantics, respectively.

A second, more significant, issue is thousands of concepts being categorized into partial-areas that are rooted at very general groups. For example, the partial-area *Finding by site* in the *Clinical finding* taxonomy summarizes 9,602 concepts. This represents a significant over summarization of the underlying hierarchy. The concepts in these large partial-areas are not easily accessible.

The *subject subtaxonomy* (introduced in Ochs et al. [119]) was developed to address this problem and to enable further scalability and flexibility of taxonomy-based summarization and quality assurance. A subject subtaxonomy is created by selecting a concept, for example, *Bleeding* or *Heart disease*, and all of its descendants. Since quality assurance for a whole hierarchy is not practical, anecdotally auditors usually concentrate on subjects of high interest. Subject subtaxonomies allow an auditor to focus on manageable portions of a hierarchy, covering specific subjects within a large hierarchy.

Given an arbitrary concept *c*, a subject subtaxonomy is derived using the derivation methodology described Section 2.2.1.1, but it is applied only to the concept subhierarchy rooted at *c*. The root area and unique root partial-area consist of *c* and all of its descendants with the same relationships. While the definition of the subject subtaxonomy is applicable to any SNOMED CT concept in a hierarchy with attribute relationships, it is recommended that *c* represent some desired subject area. If an editor wants to concentrate on a specific subject area, she can choose a concept that best represents the subject area.

The previously discussed root subtaxonomy can be considered a special case of the subject subtaxonomy. In the root subtaxonomy one picks a partial-area *p* to be the root of a subtaxonomy that includes *p* and all of its descendant partial-areas. The resulting subtaxonomy would be equivalent to a subject subtaxonomy created using *p*'s root concept.

However, the subject subtaxonomy approach is more flexible than the root subtaxonomy approach. For example, *Cancer* and many of its descendent concepts are hidden in large partial-areas in the complete *Clinical finding* taxonomy. Thus, they are also not accessible in relationship subtaxonomies and root subtaxonomies. However, *Cancer* can be selected as the root of a subject subtaxonomy, as done in Figure 3.10, making its subhierarchy of concepts more accessible in terms of summarization and quality assurance.

Subject subtaxonomies are not necessarily disjoint, because concepts may belong to multiple subject subtaxonomies. Additionally, subject subtaxonomy partial-areas are not always a subset of those in the complete taxonomy (see the *Cancer* subtaxonomy in Figure 3.10).

Figure 3.9 shows the subject subtaxonomy derived using the concept *Bleeding* from the January 2013 release of SNOMED CT. Compared to the complete *Clinical finding* partial-area taxonomy with 10,614 partial-areas, this subject subtaxonomy, with only 199 partial-areas, is significantly smaller. Over half (56%=522/932) of the concepts summarized by this subject subtaxonomy are in {*Associated morphology*, *Finding site*}.

The first row of larger partial-areas in this area indicates the major types of bleeding-related findings in SNOMED CT, such as *Hemorrhage of abdominal cavity structure* (186 concepts), *Gastrointestinal hemorrhage* (117), and *Genitourinary tract hemorrhage* (88), demonstrating the summary effect provided by the subject subtaxonomy.

Figure 3.10 shows the *Cancer* (*Malignant neoplasm disease*) subject subtaxonomy for the January 2014 SNOMED CT release. The majority of the *Cancer* subject subtaxonomy's concepts (3,124, 88.5%) are in {*Associated morphology*, *Finding site*} (like *Bleeding*). The *Cancer* subject subtaxonomy includes 64 partial-areas (highlighted in yellow in Figure 3.10) that are not in the complete *Clinical finding* taxonomy. These concepts are typically inside large partial-areas in the complete taxonomy, for example, all of the concepts in the {*Associated morphology, Finding site*} yellow partial-areas in Figure 3.10 are inside the large *Mass of body structure* (7,010 concepts) partial-area.

These partial-areas appear because the relationships *Associated morphology* and *Finding site* are introduced in the subtaxonomy at a lower descendant concept than in the complete taxonomy. In the complete taxonomy, all of the concepts in the yellow partial-areas are descendants of *Mass of body structure*, which is an introduction point for both *Associated morphology* and *Finding site* in the complete *Clinical finding* taxonomy. Thus, the *Cancer* subject subtaxonomy summarizes SNOMED CT Cancer disorders in a view that is more useful for both summarization and quality assurance.

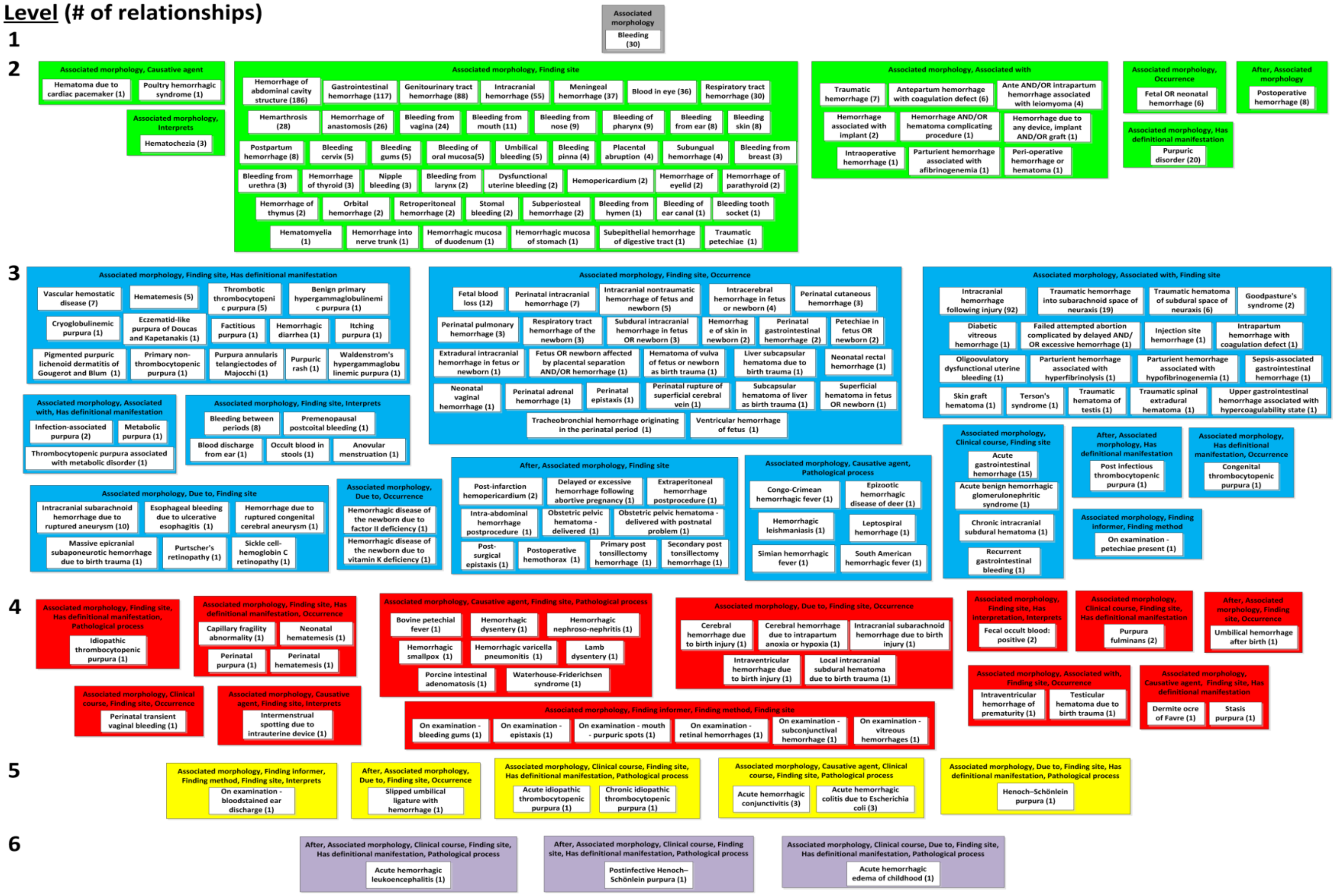

Level (# of relationships)
1
2
3
4
5
6
Associated morphology
Bleeding (30)
Associated morphology, Causative agent
Hematoma due to cardiac pacemaker (1)
Poultry hemorrhagic syndrome (1)
Associated morphology, Interprets
Hematochezia (3)
Associated morphology, Finding site
Hemorrhage of abdominal cavity structure (186)
Gastrointestinal hemorrhage (117)
Genitourinary tract hemorrhage (88)
Intracranial hemorrhage (55)
Meningeal hemorrhage (37)
Blood in eye (36)
Respiratory tract hemorrhage (30)
Hemarthrosis (28)
Hemorrhage of anastomosis (26)
Bleeding from vagina (24)
Bleeding from mouth (11)
Bleeding from nose (9)
Bleeding of pharynx (9)
Bleeding from ear (8)
Bleeding skin (8)
Postpartum hemorrhage (8)
Bleeding cervix (5)
Bleeding gums (5)
Bleeding of oral mucosa(5)
Umbilical bleeding (5)
Bleeding pinna (4)
Placental abruption (4)
Subungual hemorrhage (4)
Bleeding from breast (3)
Bleeding from urethra (3)
Hemorrhage of thyroid (3)
Nipple bleeding (3)
Bleeding from larynx (2)
Dysfunctional uterine bleeding (2)
Hemopericardium (2)
Hemorrhage of eyelid (2)
Hemorrhage of parathyroid (2)
Hemorrhage of thymus (2)
Orbital hemorrhage (2)
Retroperitoneal hemorrhage (2)
Stomal bleeding (2)
Subperiosteal hemorrhage (2)
Bleeding from hymen (1)
Bleeding of ear canal (1)
Bleeding tooth socket (1)
Hematomyelia (1)
Hemorrhage into nerve trunk (1)
Hemorrhagic mucosa of duodenum (1)
Hemorrhagic mucosa of stomach (1)
Subepithelial hemorrhage of digestive tract (1)
Traumatic petechiae (1)
Associated morphology, Associated with
Traumatic hemorrhage (7)
Antepartum hemorrhage with coagulation defect (6)
Ante AND/OR intrapartum hemorrhage associated with leiomyoma (4)
Hemorrhage associated with implant (2)
Hemorrhage AND/OR hematoma complicating procedure (1)
Hemorrhage due to any device, implant AND/OR graft (1)
Intraoperative hemorrhage (1)
Parturient hemorrhage associated with afibrinogenemia (1)
Peri-operative hemorrhage or hematoma (1)
Associated morphology, Occurrence
Fetal OR neonatal hemorrhage (6)
After, Associated morphology
Postoperative hemorrhage (8)
Associated morphology, Has definitional manifestation
Purpuric disorder (20)
Associated morphology, Finding site, Has definitional manifestation
Vascular hemostatic disease (7)
Hematemesis (5)
Thrombotic thrombocytopeni c purpura (5)
Benign primary hypergammaglobulinemi c purpura (1)
Cryoglobulinemic purpura (1)
Eczematid-like purpura of Doucas and Kapetanakis (1)
Factitious purpura (1)
Hemorrhagic diarrhea (1)
Itching purpura (1)
Pigmented purpuric lichenoid dermatitis of Gougerot and Blum (1)
Primary non-thrombocytopenic purpura (1)
Purpura annularis telangiectodes of Majocchi (1)
Purpuric rash (1)
Waldenstrom's hypergammaglobu linemic purpura (1)
Associated morphology, Finding site, Occurrence
Fetal blood loss (12)
Perinatal intracranial hemorrhage (7)
Intracranial nontraumatic hemorrhage of fetus and newborn (5)
Intracerebral hemorrhage in fetus or newborn (4)
Perinatal cutaneous hemorrhage (3)
Perinatal pulmonary hemorrhage (3)
Respiratory tract hemorrhage of the newborn (3)
Subdural intracranial hemorrhage in fetus OR newborn (3)
Hemorrhag e of skin in newborn (2)
Perinatal gastrointestinal hemorrhage (2)
Petechiae in fetus OR newborn (2)
Extradural intracranial hemorrhage in fetus or newborn (1)
Fetus OR newborn affected by placental separation AND/OR hemorrhage (1)
Hematoma of vulva of fetus or newborn as birth trauma (1)
Liver subcapsular hematoma due to birth trauma (1)
Neonatal rectal hemorrhage (1)
Neonatal vaginal hemorrhage (1)
Perinatal adrenal hemorrhage (1)
Perinatal epistaxis (1)
Perinatal rupture of superficial cerebral vein (1)
Subcapsular hematoma of liver as birth trauma (1)
Superficial hematoma in fetus OR newborn (1)
Tracheobronchial hemorrhage originating in the perinatal period (1)
Ventricular hemorrhage of fetus (1)
Associated morphology, Associated with, Finding site
Intracranial hemorrhage following injury (92)
Traumatic hemorrhage into subarachnoid space of neuraxis (19)
Traumatic hematoma of subdural space of neuraxis (6)
Goodpasture's syndrome (2)
Diabetic vitreous hemorrhage (1)
Failed attempted abortion complicated by delayed AND/ OR excessive hemorrhage (1)
Injection site hemorrhage (1)
Intrapartum hemorrhage with coagulation defect (1)
Oligoovulatory dysfunctional uterine bleeding (1)
Parturient hemorrhage associated with hyperfibrinolysis (1)
Parturient hemorrhage associated with hypofibrinogenemia (1)
Sepsis-associated gastrointestinal hemorrhage (1)
Skin graft hematoma (1)
Terson's syndrome (1)
Traumatic hematoma of testis (1)
Traumatic spinal extradural hematoma (1)
Upper gastrointestinal hemorrhage associated with hypercoagulability state (1)
Associated morphology, Associated with, Has definitional manifestation
Infection-associated purpura (2)
Metabolic purpura (1)
Thrombocytopenic purpura associated with metabolic disorder (1)
Associated morphology, Finding site, Interprets
Bleeding between periods (8)
Premenopausal postcoital bleeding (1)
Blood discharge from ear (1)
Occult blood in stools (1)
Anovular menstruation (1)
Associated morphology, Due to, Finding site
Intracranial subarachnoid hemorrhage due to ruptured aneurysm (10)
Esophageal bleeding due to ulcerative esophagitis (1)
Hemorrhage due to ruptured congenital cerebral aneurysm (1)
Massive epicranial subaponeurotic hemorrhage due to birth trauma (1)
Purtscher's retinopathy (1)
Sickle cell-hemoglobin C retinopathy (1)
Associated morphology, Due to, Occurrence
Hemorrhagic disease of the newborn due to factor II deficiency (1)
Hemorrhagic disease of the newborn due to vitamin K deficiency (1)
After, Associated morphology, Finding site
Post-infarction hemopericardium (2)
Delayed or excessive hemorrhage following abortive pregnancy (1)
Extraperitoneal hemorrhage postprocedure (1)
Intra-abdominal hemorrhage postprocedure (1)
Obstetric pelvic hematoma - delivered (1)
Obstetric pelvic hematoma - delivered with postnatal problem (1)
Post-surgical epistaxis (1)
Postoperative hemothorax (1)
Primary post tonsillectomy hemorrhage (1)
Secondary post tonsillectomy hemorrhage (1)
Associated morphology, Causative agent, Pathological process
Congo-Crimean hemorrhagic fever (1)
Epizootic hemorrhagic disease of deer (1)
Hemorrhagic leishmaniasis (1)
Leptospiral hemorrhage (1)
Simian hemorrhagic fever (1)
South American hemorrhagic fever (1)
Associated morphology, Clinical course, Finding site
Acute gastrointestinal hemorrhage (15)
Acute benign hemorrhagic glomerulonephritic syndrome (1)
Chronic intracranial subdural hematoma (1)
Recurrent gastrointestinal bleeding (1)
After, Associated morphology, Has definitional manifestation
Post infectious thrombocytopenic purpura (1)
Associated morphology, Has definitional manifestation, Occurrence
Congenital thrombocytopenic purpura (1)
Associated morphology, Finding informer, Finding method
On examination - petechiae present (1)
Associated morphology, Finding site, Has definitional manifestation, Pathological process
Idiopathic thrombocytopenic purpura (1)
Associated morphology, Finding site, Has definitional manifestation, Occurrence
Capillary fragility abnormality (1)
Neonatal hematemesis (1)
Perinatal purpura (1)
Perinatal hematemesis (1)
Associated morphology, Causative agent, Finding site, Pathological process
Bovine petechial fever (1)
Hemorrhagic dysentery (1)
Hemorrhagic nephroso-nephritis (1)
Hemorrhagic smallpox (1)
Hemorrhagic varicella pneumonitis (1)
Lamb dysentery (1)
Porcine intestinal adenomatosis (1)
Waterhouse-Friderichsen syndrome (1)
Associated morphology, Due to, Finding site, Occurrence
Cerebral hemorrhage due to birth injury (1)
Cerebral hemorrhage due to intrapartum anoxia or hypoxia (1)
Intracranial subarachnoid hemorrhage due to birth injury (1)
Intraventricular hemorrhage due to birth injury (1)
Local intracranial subdural hematoma due to birth trauma (1)
Associated morphology, Finding site, Has interpretation, Interprets
Fecal occult blood: positive (2)
Associated morphology, Clinical course, Finding site, Has definitional manifestation
Purpura fulminans (2)
After, Associated morphology, Finding site, Occurrence
Umbilical hemorrhage after birth (1)
Associated morphology, Clinical course, Finding site, Occurrence
Perinatal transient vaginal bleeding (1)
Associated morphology, Causative agent, Finding site, Interprets
Intermenstrual spotting due to intrauterine device (1)
Associated morphology, Finding informer, Finding method, Finding site
On examination - bleeding gums (1)
On examination - epistaxis (1)
On examination - mouth - purpuric spots (1)
On examination - retinal hemorrhages (1)
On examination - subconjunctival hemorrhage (1)
On examination - vitreous hemorrhages (1)
Associated morphology, Associated with, Finding site, Occurrence
Intraventricular hemorrhage of prematurity (1)
Testicular hematoma due to birth trauma (1)
Associated morphology, Causative agent, Finding site, Has definitional manifestation
Dermite ocre of Favre (1)
Stasis purpura (1)
Associated morphology, Finding informer, Finding method, Finding site, Interprets
On examination - bloodstained ear discharge (1)
After, Associated morphology, Due to, Finding site, Occurrence
Slipped umbilical ligature with hemorrhage (1)
Associated morphology, Clinical course, Finding site, Has definitional manifestation, Pathological process
Acute idiopathic thrombocytopenic purpura (1)
Chronic idiopathic thrombocytopenic purpura (1)
Associated morphology, Causative agent, Clinical course, Finding site, Pathological process
Acute hemorrhagic conjunctivitis (3)
Acute hemorrhagic colitis due to Escherichia coli (3)
Associated morphology, Due to, Finding site, Has definitional manifestation, Pathological process
Henoch–Schönlein purpura (1)
After, Associated morphology, Clinical course, Finding site, Has definitional manifestation, Pathological process
Acute hemorrhagic leukoencephalitis (1)
After, Associated morphology, Clinical course, Finding site, Has definitional manifestation, Pathological process
Postinfective Henoch–Schönlein purpura (1)
Associated morphology, Clinical course, Due to, Finding site, Has definitional manifestation, Pathological process
Acute hemorrhagic edema of childhood (1)

Figure 3.9 Top five (out of six) levels of the *Bleeding* subject subtaxonomy. A total of 932 bleeding-related concepts are summarized by 199 partial-areas in 42 areas. Over half (56%=522/932) of the concepts summarized by this subtaxonomy are in {*Associated morphology*, *Finding site*}. *Source: [119]*

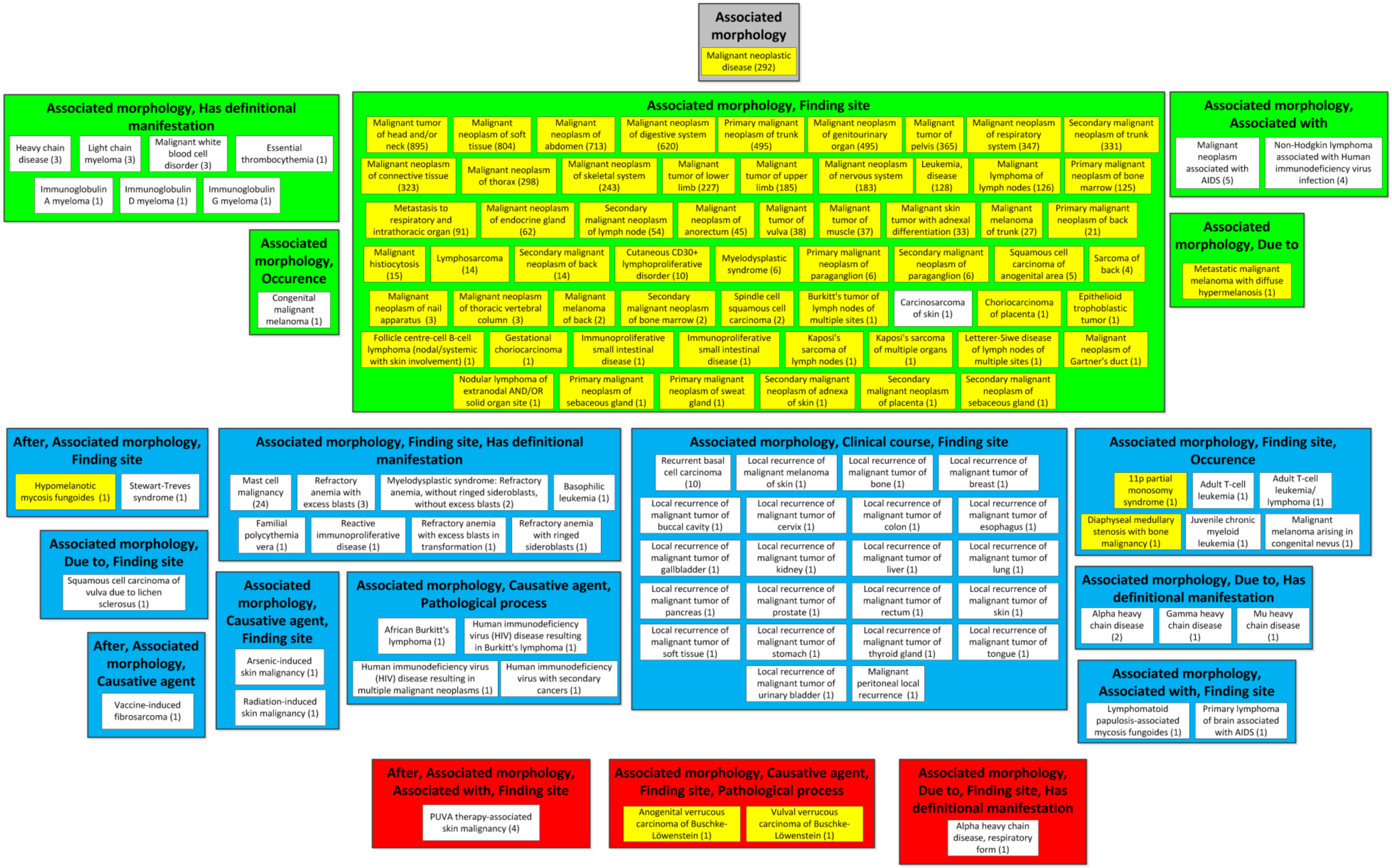

Associated morphology
Malignant neoplastic disease (292)
Associated morphology, Has definitional manifestation
Heavy chain disease (3)
Light chain myeloma (3)
Malignant white blood cell disorder (3)
Essential thrombocythemia (1)
Immunoglobulin A myeloma (1)
Immunoglobulin D myeloma (1)
Immunoglobulin G myeloma (1)
Associated morphology, Occurence
Congenital malignant melanoma (1)
Associated morphology, Finding site
Malignant tumor of head and/or neck (895)
Malignant neoplasm of soft tissue (804)
Malignant neoplasm of abdomen (713)
Malignant neoplasm of digestive system (620)
Primary malignant neoplasm of trunk (495)
Malignant neoplasm of genitourinary organ (495)
Malignant tumor of pelvis (365)
Malignant neoplasm of respiratory system (347)
Secondary malignant neoplasm of trunk (331)
Malignant neoplasm of connective tissue (323)
Malignant neoplasm of thorax (298)
Malignant neoplasm of skeletal system (243)
Malignant tumor of lower limb (227)
Malignant tumor of upper limb (185)
Malignant neoplasm of nervous system (183)
Leukemia, disease (128)
Malignant lymphoma of lymph nodes (126)
Primary malignant neoplasm of bone marrow (125)
Metastasis to respiratory and intrathoracic organ (91)
Malignant neoplasm of endocrine gland (62)
Secondary malignant neoplasm of lymph node (54)
Malignant neoplasm of anorectum (45)
Malignant tumor of vulva (38)
Malignant tumor of muscle (37)
Malignant skin tumor with adnexal differentiation (33)
Malignant melanoma of trunk (27)
Primary malignant neoplasm of back (21)
Malignant histiocytosis (15)
Lymphosarcoma (14)
Secondary malignant neoplasm of back (14)
Cutaneous CD30+ lymphoproliferative disorder (10)
Myelodysplastic syndrome (6)
Primary malignant neoplasm of paraganglion (6)
Secondary malignant neoplasm of paraganglion (6)
Squamous cell carcinoma of anogenital area (5)
Sarcoma of back (4)
Malignant neoplasm of nail apparatus (3)
Malignant neoplasm of thoracic vertebral column (3)
Malignant melanoma of back (2)
Secondary malignant neoplasm of bone marrow (2)
Spindle cell squamous cell carcinoma (2)
Burkitt's tumor of lymph nodes of multiple sites (1)
Carcinosarcoma of skin (1)
Choriocarcinoma of placenta (1)
Epithelioid trophoblastic tumor (1)
Follicle centre-cell B-cell lymphoma (nodal/systemic with skin involvement) (1)
Gestational choriocarcinoma (1)
Immunoproliferative small intestinal disease (1)
Immunoproliferative small intestinal disease (1)
Kaposi's sarcoma of lymph nodes (1)
Kaposi's sarcoma of multiple organs (1)
Letterer-Siwe disease of lymph nodes of multiple sites (1)
Malignant neoplasm of Gartner's duct (1)
Nodular lymphoma of extranodal AND/OR solid organ site (1)
Primary malignant neoplasm of sebaceous gland (1)
Primary malignant neoplasm of sweat gland (1)
Secondary malignant neoplasm of adnexa of skin (1)
Secondary malignant neoplasm of placenta (1)
Secondary malignant neoplasm of sebaceous gland (1)
Associated morphology, Associated with
Malignant neoplasm associated with AIDS (5)
Non-Hodgkin lymphoma associated with Human immunodeficiency virus infection (4)
Associated morphology, Due to
Metastatic malignant melanoma with diffuse hypermelanosis (1)
After, Associated morphology, Finding site
Hypomelanotic mycosis fungoides (1)
Stewart-Treves syndrome (1)
Associated morphology, Finding site, Has definitional manifestation
Mast cell malignancy (24)
Refractory anemia with excess blasts (3)
Myelodysplastic syndrome: Refractory anemia, without ringed sideroblasts, without excess blasts (2)
Basophilic leukemia (1)
Familial polycythemia vera (1)
Reactive immunoproliferative disease (1)
Refractory anemia with excess blasts in transformation (1)
Refractory anemia with ringed sideroblasts (1)
Associated morphology, Clinical course, Finding site
Recurrent basal cell carcinoma (10)
Local recurrence of malignant melanoma of skin (1)
Local recurrence of malignant tumor of bone (1)
Local recurrence of malignant tumor of breast (1)
Local recurrence of malignant tumor of buccal cavity (1)
Local recurrence of malignant tumor of cervix (1)
Local recurrence of malignant tumor of colon (1)
Local recurrence of malignant tumor of esophagus (1)
Local recurrence of malignant tumor of gallbladder (1)
Local recurrence of malignant tumor of kidney (1)
Local recurrence of malignant tumor of liver (1)
Local recurrence of malignant tumor of lung (1)
Local recurrence of malignant tumor of pancreas (1)
Local recurrence of malignant tumor of prostate (1)
Local recurrence of malignant tumor of rectum (1)
Local recurrence of malignant tumor of skin (1)
Local recurrence of malignant tumor of soft tissue (1)
Local recurrence of malignant tumor of stomach (1)
Local recurrence of malignant tumor of thyroid gland (1)
Local recurrence of malignant tumor of tongue (1)
Local recurrence of malignant tumor of urinary bladder (1)
Malignant peritoneal local recurrence (1)
Associated morphology, Finding site, Occurence
11p partial monosomy syndrome (1)
Adult T-cell leukemia (1)
Adult T-cell leukemia/ lymphoma (1)
Diaphyseal medullary stenosis with bone malignancy (1)
Juvenile chronic myeloid leukemia (1)
Malignant melanoma arising in congenital nevus (1)
Associated morphology, Due to, Finding site
Squamous cell carcinoma of vulva due to lichen sclerosus (1)
After, Associated morphology, Causative agent
Vaccine-induced fibrosarcoma (1)
Associated morphology, Causative agent, Finding site
Arsenic-induced skin malignancy (1)
Radiation-induced skin malignancy (1)
Associated morphology, Causative agent, Pathological process
African Burkitt's lymphoma (1)
Human immunodeficiency virus (HIV) disease resulting in Burkitt's lymphoma (1)
Human immunodeficiency virus (HIV) disease resulting in multiple malignant neoplasms (1)
Human immunodeficiency virus with secondary cancers (1)
Associated morphology, Due to, Has definitional manifestation
Alpha heavy chain disease (2)
Gamma heavy chain disease (1)
Mu heavy chain disease (1)
Associated morphology, Associated with, Finding site
Lymphomatoid papulosis-associated mycosis fungoides (1)
Primary lymphoma of brain associated with AIDS (1)
After, Associated morphology, Associated with, Finding site
PUVA therapy-associated skin malignancy (4)
Associated morphology, Causative agent, Finding site, Pathological process
Anogenital verrucous carcinoma of Buschke-Löwenstein (1)
Vulval verrucous carcinoma of Buschke-Löwenstein (1)
Associated morphology, Due to, Finding site, Has definitional manifestation
Alpha heavy chain disease, respiratory form (1)

Figure 3.10 The *Cancer* subject subtaxonomy. The Cancer subject subtaxonomy summarizes 3,531 concepts by 125 partial-areas in 19 areas. The 64 partial-areas that do not appear in the complete *Clinical finding* taxonomy are highlighted in yellow. *Source: [119]*

Table 3.5 Subject Subtaxonomy Metrics for the Ten Leading Causes of Death in the US

#	**Cause of death**	**Subject Subtaxonomy concept**	**# of Concepts**	**# of Partial-areas**	**# of Areas**	**Relative Size (Concepts / Partial-areas)**
1	Heart disease	*Heart disease*	2,402	316	61	2.4% / 3.0%
2	Cancer	*Malignant neoplastic disease*	3,531	125	19	3.6% / 1.2%
3	Chronic lower respiratory diseases	*Disorder of lower respiratory system*	1,414	354	51	1.4% / 3.4%
4	Stroke	*Cerebrovascular disease*	262	75	15	0.3% / 0.7%
5	Accidents	*Injury due to exposure to external cause*	267	65	11	0.3% / 0.6%
6	Alzheimer’s disease	*Disorder of brain*	2,300	396	67	2.3% / 3.8%
7	Diabetes	*Diabetes mellitus*	112	30	14	0.1% / 0.2%
8	Nephritis, nephrotic syndrome, and nephrosis	*Kidney disease*	909	243	47	0.9% / 2.3%
9	Influenza and Pneumonia	*Pneumonitis*	334	73	24	0.3% / 0.7%
10	Suicide	*Suicide*	16	9	2	0.4% / 29%

Source: [119]

Table 3.5 lists the metrics for subtaxonomies for the ten most common causes of death [125], along with their sizes relative to the complete partial-area taxonomy in terms of number of concepts and partial-areas. Nine are from *Clinical finding* and *Suicide* is from *Event*. From Table 3.5 one can see that, in general, subtaxonomies are significantly smaller, and thus, more manageable.

3.1.3.1 Subject Disjoint Partial-area Subtaxonomy. Disjoint partial-area taxonomies have been shown to successfully support improved content summarization [23] and quality assurance [29]. Subject subtaxonomies may contain overlapping concepts. For example, the largest area in the *Bleeding* subtaxonomy, {*Associated morphology*, *Finding site*}, has 290 overlapping concepts (55.5%).

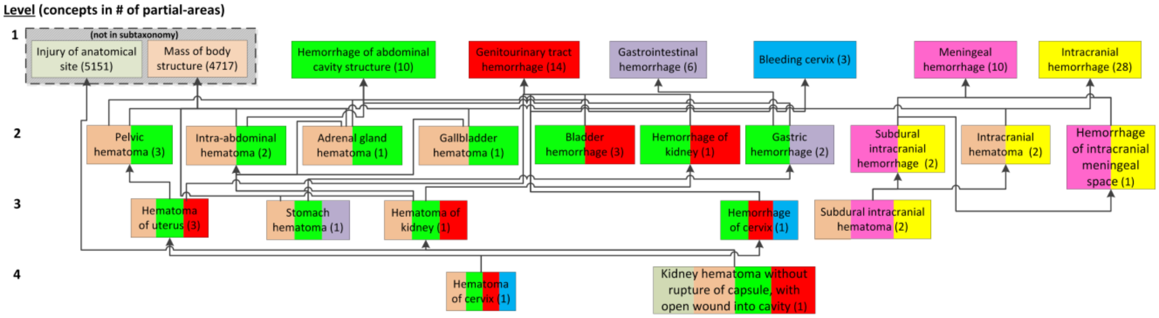

Level (concepts in # of partial-areas)
1
2
3
4
(not in subtaxonomy)
Injury of anatomical site (5151)
Mass of body structure (4717)
Hemorrhage of abdominal cavity structure (10)
Genitourinary tract hemorrhage (14)
Gastrointestinal hemorrhage (6)
Bleeding cervix (3)
Meningeal hemorrhage (10)
Intracranial hemorrhage (28)
Pelvic hematoma (3)
Intra-abdominal hematoma (2)
Adrenal gland hematoma (1)
Gallbladder hematoma (1)
Bladder hemorrhage (3)
Hemorrhage of kidney (1)
Gastric hemorrhage (2)
Subdural intracranial hemorrhage (2)
Intracranial hematoma (2)
Hemorrhage of intracranial meningeal space (1)
Hematoma of uterus (3)
Stomach hematoma (1)
Hematoma of kidney (1)
Hemorrhage of cervix (1)
Subdural intracranial hematoma (2)
Hematoma of cervix (1)
Kidney hematoma without rupture of capsule, with open wound into cavity (1)

Figure 3.11 An excerpt of 23 disjoint partial-areas from the disjoint partial-area subtaxonomy derived for the concepts in {*Associated morphology*, *Finding by site*}.
Source: [119]

For subject subtaxonomies, the disjoint partial-area taxonomy derivation methodology must be altered to account for overlapping concepts in a subject subtaxonomy that are in partial-areas that are outside of the subject subtaxonomy. For example, the concept *Intra-abdominal hematoma* has two parents in its area in the complete *Clinical finding* taxonomy: *Hemorrhage of abdominal cavity structure (*in *Bleeding's* subject subtaxonomy) and *Mass of abdominal cavity structure* (in the partial-area *Mass of body structure*, outside the *Bleeding* subject subtaxonomy). *Intra-abdominal hematoma* inherits the semantics of both partial-area roots and belongs in the disjoint taxonomy.

Disjoint partial-area subtaxonomy derivation accounts for this by (1) ensuring that all of the concepts in the disjoint partial-area taxonomy are semantically related to the subject *c* by considering only concepts that are descendants of *c*, and (2) considering overlapping concepts that overlap with partial-areas outside of the subject subtaxonomy, since such concepts are complex.

Figure 3.11 shows an excerpt of 23 disjoint partial-areas from the disjoint partial-area subtaxonomy for {*Associated morphology*, *Finding site*}. The disjoint partial-areas *Mass of body structure* and *Injury of anatomical site*, shown in a gray box, are not part of the *Bleeding* subject subtaxonomy, but many *Bleeding* concepts overlap with them in the complete *Clinical finding* taxonomy. Partial-areas outside of the subject subtaxonomy, such as *Mass of body structure*, which overlap with partial-areas in the subject subtaxonomy, for example, *Hemorrhage of abdominal cavity structure*, are not part of the subject subtaxonomy and can be hidden, but are important for quality assurance to capture the complexity of the overlapping concepts. For example, the disjoint partial-area *Pelvic hematoma* (3) would not exist if such overlap was not considered.

The disjoint partial-area taxonomy for the *Bleeding* subtaxonomy's {*Associated morphology*, *Finding site*} area contains 236 disjoint partial-areas. Most of the disjoint partial-areas are small: 176 (78.8%) are singletons (one concept). The disjoint partial-area subtaxonomy more accurately summarizes the concepts in this area than the partial-area taxonomy (Figure 3.9) at the cost of there being more summarizing groups. For example, there are 186 concepts in the partial-area *Hemorrhage of body cavity structure*, but only ten are descendants of just this root. The other 176 concepts also belong to other partial-areas. The overlapping disjoint partial-areas are made explicit in Figure 3.11.

For the *Cancer* subject subtaxonomy, the majority of the concepts in the largest area, {*Associated morphology*, *Finding site*}, are overlapping concepts. In total, 2,398 overlapping concepts (76.8%) are in this area. However, the *Cancer* subject subtaxonomy's overlapping concepts are different from those in the *Bleeding* subject subtaxonomy. Many of the partial-areas in the *Cancer* subject subtaxonomy are not found in the complete taxonomy (yellow partial-areas in Figure 3.10). The overlapping concepts in these partial-areas are not necessarily overlapping concepts in the complete *Clinical finding* taxonomy, since they are all contained in the large *Mass of body structure* partial-area. A future study, described in Section 4.1, will investigate the characteristics of these concepts, which are only overlapping concepts within a subject subtaxonomy.

3.1.3.2 In Support of Quality Assurance. Previous SNOMED CT quality assurance studies have focused on [29] *complex* concepts, e.g., overlapping concepts [23], which were shown to have more errors with high statistical significance for the small *Specimen* hierarchy due to the difficulty in modeling complex concepts. Overlapping concepts are more complex than non-overlapping concepts, since they are specializations of all the roots of the partial-areas they are contained in.

However, the number of overlapping concepts in the complete *Clinical finding* taxonomy (14,450) is overwhelming and reviewing all of them is impractical. The number of overlapping concepts in a subject subtaxonomy may be significantly smaller. For example, the *Bleeding* subject subtaxonomy has only a few hundred overlapping concepts. However, the error rates for these concepts had to be investigated. The analysis of overlapping concepts (Hypothesis H1) was repeated and three new refined hypotheses (H2-H4) were tested for a subject subtaxonomy of a large hierarchy.

Hypothesis H1: Overlapping concepts are more likely to have errors than non-overlapping concepts.

Another group of concepts, which was also shown to have more errors with high statistical significance, are *uncommonly classified* concepts, e.g., those in small partial-areas [26]. A possible reason for their uncommon classification may be a modeling error. Once the error is corrected (e.g., by adding a parent or relationship) a concept may join another common classification according to its revised modeling. However, to account for concepts that overlap between a small partial-area and a large partial-area H2 is introduced:

> *Hypothesis H2:* Concepts in small disjoint partial-areas are more likely to have errors than concepts in large disjoint partial-areas.

H1 and H2 can be compounded into H3.

> *Hypothesis H3:* Concepts in small overlapping disjoint partial-areas are more likely to have errors than concepts in large overlapping disjoint partial-areas.

H3 expresses that concepts that are both complex and uncommonly classified tend to have more errors than concepts that are just complex.

The number of partial-areas a concept belongs to is called its "*degree of overlap*."

> *Hypothesis H4: Concepts with a higher degree of overlap exhibit a higher error rate.*

Concepts that overlap between more partial-areas inherit the semantics of more roots, and thus, are more complex than concepts that overlap between fewer partial-areas.

Even the number of overlapping concepts in a subject subtaxonomy may be overwhelming when only limited resources are available to audit them. The above hypotheses can guide a quality assurance methodology by prioritizing which overlapping concepts should be reviewed first to maximize yield.

To test the hypotheses, a sample of 300 concepts was reviewed for errors by three domain experts who are trained in medicine and have extensive terminology auditing experience. The review process consisted of two phases. First, each auditor was given the complete sample as a list of concepts in alphabetical order and worked independently. The auditors were blind to the methodology and the hypotheses. Auditors were not aware of which disjoint partial-area a given concept was summarized by and the review process performed by each auditor was the same for all concepts. Each auditor then reported all errors found. As shown in [126], there are substantial differences among quality assurance reports from several auditors and a report from one auditor is not reliable. However, a consensus among several auditors' reports was shown to result in a reliable quality assurance report.

Thus, the second phase was used for consensus building. Each auditor was given a complete list of errors from all auditors. Each auditor then marked "agree" or "disagree" for each error. A concept was considered erroneous if all auditors agreed on the error. A similar consensus quality assurance protocol was used when auditing overlapping concepts in the *Specimen* hierarchy [29].

To test H1-H4, three auditors reviewed a sample of 300 concepts from the {*Associated morphology, Finding site*} area in the *Bleeding* subtaxonomy for errors: 200 randomly selected overlapping concepts (70%=200/290) and 100 randomly selected non-overlapping concepts (43%=100/232). The latter were taken from partial-areas that had overlapping concept.

The auditors reviewed the January 2013 inferred version of SNOMED CT. Together, the auditors first found 131 erroneous concepts. Next all auditors agreed that 87 (66%) of these concepts had at least one same error (Table 3.7). Among the erroneous concepts, 36 were primitives and 51 were fully defined. The auditors all agreed on 123 errors in these 87 concepts (1.41 errors per erroneous concept). Table 3.6 provides a summary of the types of errors found in this study.

Table 3.6 Errors Found in *Bleeding* Subject Subtaxonomy by Error Type

Error Type	**Number of Errors**
Missing parent	50
Incorrect parent	11
Missing relationship	4
Incorrect relationship	10
Incorrect synonym (+ missing concept)	1
Duplicate concepts	2 pairs

Source: [119]

For H1, 39% (=78/200) of overlapping concepts were determined to be erroneous, versus 9% (=9/100) of non-overlapping concepts. Thus, in the sample, overlapping concepts were 4.33 times more likely to be erroneous. For statistical analysis the double bootstrap approach was used to account for potential dependency of errors in the sample. H1 was found to be statistically significant (*p*=0.0016). For H2, several boundary points between small and large were tested (see Table 3.7, Figure 3.12). Using a boundary point of seven [22, 26], 37.3% (=85/228) of concepts in small disjoint partial-areas were erroneous versus 2.78% (=2/72) of concepts in large disjoint partial-areas. H2 was also significant (*p*=0.0394).

Table 3.7 Auditing Results for Overlapping Concepts and Non-overlapping Concepts in Small and Large Disjoint Partial-areas for Several Boundary Points Between "Small" and "Large"

Disjoint partial-area size	**Overlapping (Levels 2-8)**		**Non-overlapping (Level 1)**		**Total**	
	# Sample	**# Erroneous**	**# Sample**	**# Erroneous**	**# Sample**	**# Erroneous**
Boundary of 2 (Singletons)						
Small (= 1 concept)	146	61 (41.8%)	7	0 (0%)	153	61 (39.9%)
Large (> 1 concept)	54	17 (31.5%)	93	9 (9.68%)	147	16 (10.9%)
Boundary of 3						
Small (< 3 concepts)	168	68 (40.5%)	15	2 (13.3%)	183	70 (38.3%)
Large (>=3 concepts)	32	10 (31.3%)	85	7 (8.23%)	117	17 (14.5%)
Boundary of 5						
Small (< 5 concepts)	184	75 (40.8%)	28	4 (14.3%)	212	79 (37.3%)
Large (>= 5 concepts)	16	3 (18.8%)	72	5 (6.94%)	88	8 (9.09%)
Boundary of 7						
Small (< 7 concepts)	194	78 (40.2%)	34	7 (20.6%)	228	85 (37.3%)
Large (>= 7 concepts)	6	0 (0%)	66	2 (3.0%)	72	2 (2.78%)
Boundary of 10						
Small (< 10 concepts)	200	78 (39%)	43	7 (16.3%)	243	85 (35.0%)
Large (>= 10 concepts)	0	-	57	2 (3.51%)	57	2 (3.51%)
Total	200	78 (39%)	100	9 (9%)	300	87 (29%)

Source: [119]

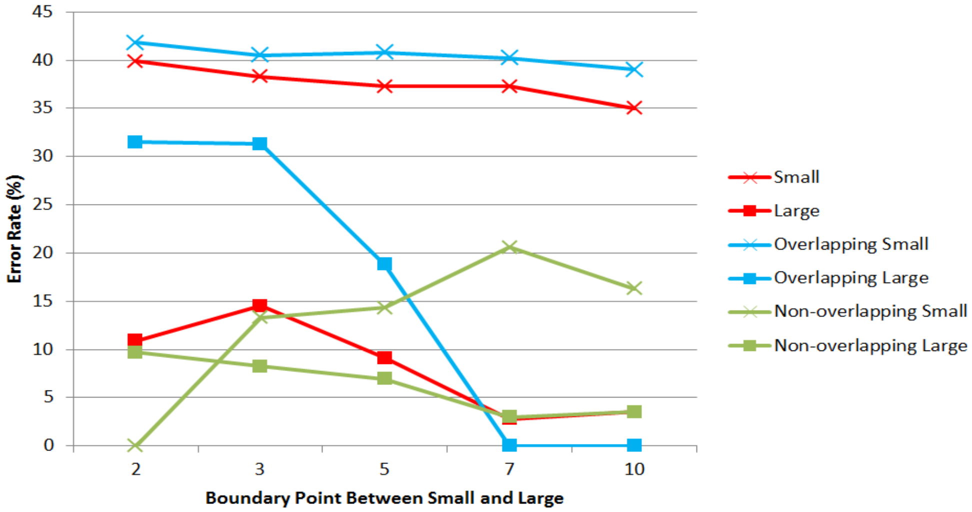


Figure 3.12 The error rates for small overlapping disjoint partial-areas for the various boundary points between small and large.
Source: [119]

In the *Bleeding* subtaxonomy, the disjoint partial-area taxonomy for {*Associated morphology*, *Finding site*} had only one large overlapping disjoint partial-area when a boundary of seven was used. The six concepts sampled from this disjoint partial-area had no errors. Small overlapping disjoint partial-areas, on the other hand, had an error rate of 40.2% (=78/194). But due to the small sample size for large overlapping disjoint partial-areas, H3 had no significance (*p*=0.2601).

Table 3.8 provides a breakdown of errors by overlap level of the disjoint partial-area taxonomy. To test H4, each level is compared to the previous level. From Level 1 to Level 7 the error rate is increasing, as expected. This hypothesis was statistically significant when comparing Level 1 to Level 2 (*p*=0.0322) and Level 2 to 3 (*p*=0.0336). Other comparisons were not significant due to the smaller sample sizes of Level 4 and above; changes in error rate were too small to detect. When Level 3 was compared to Levels 4-8 combined (error rate of 24/39=61.5%), the hypothesis was significant (*p*=0.0116). Table 3.9 shows five examples of errors and their proposed solutions.

Table 3.8 Auditing Results Broken Down by Disjoint Partial-area Taxonomy Level

Level	# Concepts in Sample	# of Erroneous Concepts	% Erroneous Concepts
1	100	9	9%
2	90	24	26.7%
3	71	29	40.8%
4	18	9	50%
5	10	7	70%
6	6	5	83.3%
7	2	2	100%
8	3	2	66.7%
Total	300	87	32.3%
Total for 4-8	39	25	64.1%

Source: [119]

Table 3.9 Five Examples of Errors Reported by the Auditors

Concept	Error	Proposed Solution
Bleeding varices of prostate	***Missing relationships*****:** *associated morphology* and *finding site*, with target concepts *varix* and *venous structure*, respectively.	Add the two new relationships in a role group.
Hemorrhage of cervix	***Incorrect parent*****:** *Hemorrhage of abdominal cavity structure*	Remove IS-A to *Hemorrhage of abdominal cavity structure* (corrected independently in Jan 2014 release)
Hematoma of pinna	***Missing child*****:** *Chronic hematoma of pinna* (which is incorrectly a synonym of the concept *Cauliflower ear*).	Add *Chronic hematoma of pinna* concept and remove the synonym from *Cauliflower ear*
Peptic ulcer with hemorrhage AND obstruction	***Incorrect relationship target*****:** *associated morphology* relationship with a target concept *Hemorrhage*	Make target concept of *associated morphology* relationship *Bleeding ulcer* to be consistent with *Esophageal bleeding*
Bleeding gastric varices	***Missing parent*****:** *Venous hemorrhage*	Add IS-A to *Venous hemorrhage*

Note: James T. Case, the head of the US Extension of SNOMED CT, confirmed all of these errors and forwarded the corrections to the IHTSDO.
Source: [119]

3.1.3.2.1 External Review of Error Report. All erroneous concepts and proposed corrections were reported to James T. Case, the head of the US Extension of SNOMED CT. He confirmed 78 (out of 87, 89.7%) of the erroneous concepts had at least one error, a high percentage considering the known variability of auditor reports [126]. In cases where an error was corrected independently of the submitted auditing report the concept was still counted erroneous, as the modeling was changed in releases after the one audited in this study (January 2013 SNOMED CT release). The nine concepts that were judged correct by James T. Case were in small overlapping disjoint partial-areas (using a boundary of seven).

Statistical analysis, using only the 78 concepts identified as erroneous by James T. Case, was repeated for H1-H4. H1 and H2 were statistically significant (p=0.0048 and p=0.0168, respectively), H3 was again not significant (p=0.2563), and H4 was still significant, except for Level 1 vs. Level 2, which was almost statistically significant (p=0.0628).

James T. Case stated that the review of the *Bleeding* subhierarchy's audit report concepts identified at least two areas where variations in modeling or constraints of the existing concept model were an obstacle to consistent and uniform modeling of concepts.

First, the auditors found 45 concepts, all of which had fully specified names (FSNs) starting with "acute" or "chronic," that were missing the ancestor *Acute disease* or *Chronic disease*, respectively. This finding is similar to the results of Rector et al. [85], who analyzed many of the *Clinical finding* hierarchy's acute and chronic concepts and discovered many such cases. These errors arose out of a lack of distinction between "acute" and "chronic" as a morphology and "acute" and "chronic" as a clinical course.

There are specific structures associated with pathological lesions that allow them to be classified as acute or chronic (e.g., *lesion fibrosis*, *infiltration with inflammatory cells*, etc.). This is sometimes orthogonal to the temporal aspect of the clinical course. A number of concepts in the audit report had “acute” or “chronic” in the FSN and an associated morphology relationship assigned, but did not have a clinical course relationship assigned, so they did not auto-classify under *Acute disease* or *Chronic disease*, as would be expected.

The second major issue was that many concepts associated with traumatic injuries did not auto-classify under *Traumatic injury* because they did not have an *associated morphology* relationship assigned that was a child of *Traumatic abnormality* (morphologic abnormality). This would best be handled by modeling the current *Traumatic injury* concept with a “*Pathological process* = *Traumatic*” relationship, and then applying that same relationship to all concepts that were caused by trauma, but the current quality assurance rules in the SNOMED CT editing environment do not allow that, even though it is in the allowed value hierarchy for the *pathological process* relationship. A question related to this was forwarded to IHTSDO for clarification.

One issue that arose from the audit was errors being uncovered (and suggested corrections being made) according to the inferred version of SNOMED CT. Often, the auditors would identify a particular concept missing a parent and would suggest adding a new IS-A relationship to correct the problem. However, when James T. Case investigated the error in the stated view of SNOMED CT, he found that the issue was not the missing parent, but instead incomplete or incorrect modeling of attribute relationships. Correctly modeling the attribute relationships would then lead to the auto-classification of the missing parent. The auditors correctly identified the concept as erroneous, but their suggested solution did not necessarily correct the problem.

Thus, one of the areas of the subtaxonomy quality assurance methodology that needs improvement is the process of auditors suggesting corrections. Since the auditors in this study reviewed the inferred version of SNOMED CT, as opposed to the stated version, the source of errors was often hidden and the proposed solution was often erroneous. Being able to see both the stated and inferred views of SNOMED CT is extremely important for correcting errors.

Training domain experts to be familiar with the SNOMED CT concept model is difficult. James T. Case confirmed that the current process of having the initial review performed by a group of auditors, and then submitting the findings via the USCRS [127] for final review by an editor familiar with the SNOMED CT concept model leads to a proper correction. However, the main value of an external quality assurance report is exposing errors, even if the suggested corrections are not accepted by a SNOMED CT editor familiar with the SNOMED CT concept model. In future studies the impact of providing auditors with both the stated and inferred versions of SNOMED CT will be investigated.

3.1.3.3 Conclusions. The scalability of taxonomy-based terminology maintenance to large SNOMED CT hierarchies was demonstrated using subject subtaxonomies. This represents a significant improvement over the previous approach of reviewing complete taxonomies, which may have thousands of partial-areas (e.g., *Clinical finding*). Such large taxonomies are hard for humans to visualize, which prevents effective taxonomy-based quality assurance, based on reviewing groups of concepts that have higher error rates (e.g., small partial-areas [22, 26, 28]). There are thousands of such concepts in a large hierarchy, e.g., 14,450 (14.3%) concepts in "small" partial-areas and 14,220 (14.5%) overlapping concepts in the *Clinical finding* hierarchy. Available quality assurance resources do not typically enable a thorough review of so many concepts.

These difficulties were addressed by combining several novel techniques. The first technique is to concentrate on a subject subtaxonomy, which is intuitive for terminology curators because it summarizes all descendants of a chosen broad concept, e.g., *Bleeding* or *Cancer*. This way, the attention of a curator is focused on a comprehensible subtaxonomy that still summarizes a sizable subject-based portion of the hierarchy. Second, refined hypotheses were formulated regarding concepts with high likelihood of errors. Third, the review of concepts is prioritized according to the ratios for the refined hypotheses.

Table 3.10 Recommended Order of Auditing in the *Bleeding* Subject Subtaxonomy

Rank	**Hypothesis**	**Group**	**Error Rate**
1	H4	Overlap levels 4-8	64.1%
2	H4	Overlap level 3	40.8%
3	H3	Small overlapping disjoint partial-areas	40.2%
4	H4	Overlap level 2	26.7%
5	H2	Small non-overlapping disjoint partial-areas	20.6%

Source: [119]

When applying taxonomy-based quality assurance methodologies to small hierarchies, two kinds of groups were discovered with higher likelihood of errors in small hierarchies: concepts in small partial-areas [26, 28] and overlapping concepts [29]. The challenges for scalability included whether this still holds true for concepts in subject subtaxonomies and prioritizing among the groups' concepts.

New hypotheses (H2-H4) were formulated and tested, while confirming the previously established hypothesis (H1), for the *Bleeding* subtaxonomy. When there are many overlapping concepts and a relatively extensive level of overlap, as for the *Bleeding* and *Cancer* subtaxonomies, resources for reviewing overlapping concepts need to be prioritized.

According to this study, the quality assurance methodology steps corresponding to the hypotheses should be applied in decreasing error percentage order (Table 3.10). Thus, an editor will achieve a higher yield for a given effort. Future studies will investigate error rates in other subject-based subtaxonomies, e.g., *Cancer* with 2,398 overlapping concepts, to verify this order.

The study confirmed most of the hypotheses and the feasibility of the subject subtaxonomy paradigm to support scalability of taxonomy-based maintenance of large SNOMED CT hierarchies. More experiments will be performed, using other subtaxonomies, where the sample sizes in the *Bleeding* subtaxonomy were not sufficient to achieve statistical significance (i.e., H3).

3.1.4 Focus Subtaxonomy

A variation of the subject subtaxonomy is the *focus subtaxonomy*. A focus subtaxonomy is a subject subtaxonomy that includes all of the ancestors of the chosen subject concept *c*. The focus subtaxonomy allows an editor to view the chosen subject concept in the context of its ancestor, summarizing how *c* obtained its structure.

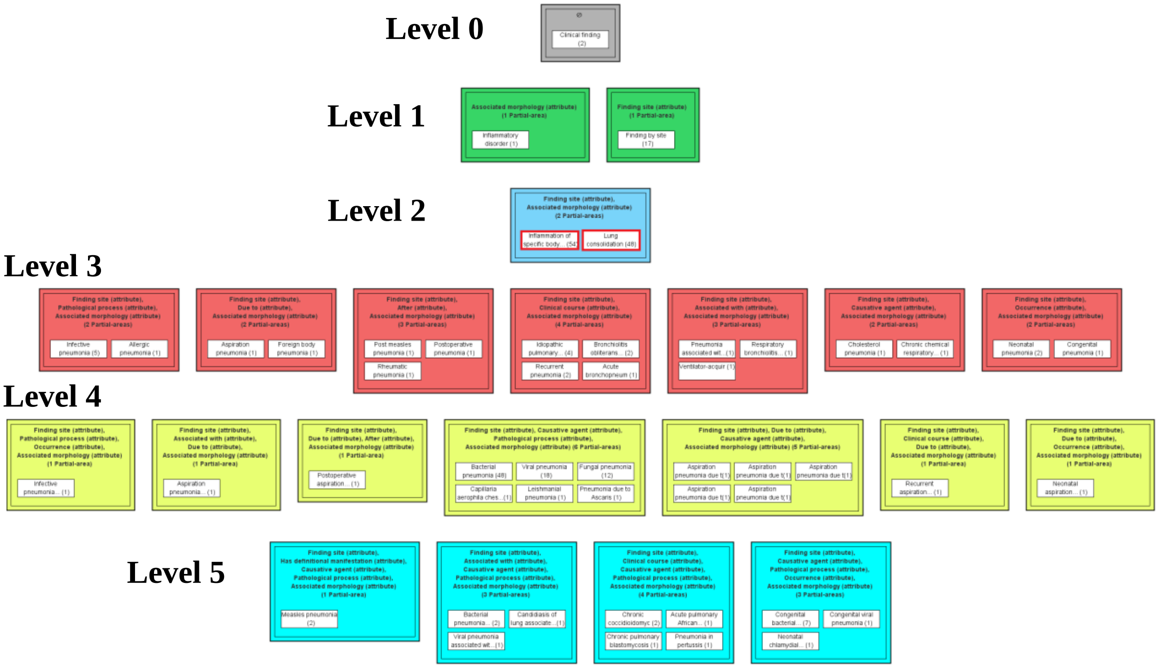

Level 0
Clinical finding (2)
Level 1
Associated morphology (attribute) (1 Partial-area)
Inflammatory disorder (1)
Finding site (attribute) (1 Partial-area)
Finding by site (17)
Level 2
Finding site (attribute), Associated morphology (attribute) (2 Partial-areas)
Inflammation of specific body... (54)
Lung consolidation (48)
Level 3
Finding site (attribute), Pathological process (attribute), Associated morphology (attribute) (2 Partial-areas)
Infective pneumonia (5)
Allergic pneumonia (1)
Finding site (attribute), Due to (attribute), Associated morphology (attribute) (2 Partial-areas)
Aspiration pneumonia (1)
Foreign body pneumonia (1)
Finding site (attribute), After (attribute), Associated morphology (attribute) (3 Partial-areas)
Post measles pneumonia (1)
Postoperative pneumonia (1)
Rheumatic pneumonia (1)
Finding site (attribute), Clinical course (attribute), Associated morphology (attribute) (4 Partial-areas)
Idiopathic pulmonary... (4)
Bronchiolitis obliterans... (2)
Recurrent pneumonia (2)
Acute bronchopneum (1)
Finding site (attribute), Associated with (attribute), Associated morphology (attribute) (3 Partial-areas)
Pneumonia associated wit... (1)
Respiratory bronchiolitis... (1)
Ventilator-acquir (1)
Finding site (attribute), Causative agent (attribute), Associated morphology (attribute) (2 Partial-areas)
Cholesterol pneumonia (1)
Chronic chemical respiratory... (1)
Finding site (attribute), Occurrence (attribute), Associated morphology (attribute) (2 Partial-areas)
Neonatal pneumonia (2)
Congenital pneumonia (1)
Level 4
Finding site (attribute), Pathological process (attribute), Occurrence (attribute), Associated morphology (attribute) (1 Partial-area)
Infective pneumonia... (1)
Finding site (attribute), Associated with (attribute), Due to (attribute), Associated morphology (attribute) (1 Partial-area)
Aspiration pneumonia... (1)
Finding site (attribute), Due to (attribute), After (attribute), Associated morphology (attribute) (1 Partial-area)
Postoperative aspiration... (1)
Finding site (attribute), Causative agent (attribute), Pathological process (attribute), Associated morphology (attribute) (6 Partial-areas)
Bacterial pneumonia (48)
Viral pneumonia (18)
Fungal pneumonia (12)
Capillaria aerophila ches... (1)
Leishmanial pneumonia (1)
Pneumonia due to Ascaris (1)
Finding site (attribute), Due to (attribute), Causative agent (attribute), Associated morphology (attribute) (5 Partial-areas)
Aspiration pneumonia due t(1)
Aspiration pneumonia due t(1)
Aspiration pneumonia due t(1)
Aspiration pneumonia due t(1)
Aspiration pneumonia due t(1)
Finding site (attribute), Clinical course (attribute), Due to (attribute), Associated morphology (attribute) (1 Partial-area)
Recurrent aspiration... (1)
Finding site (attribute), Due to (attribute), Occurrence (attribute), Associated morphology (attribute) (1 Partial-area)
Neonatal aspiration... (1)
Level 5
Finding site (attribute), Has definitional manifestation (attribute), Causative agent (attribute), Pathological process (attribute), Associated morphology (attribute) (1 Partial-area)
Measles pneumonia (2)
Finding site (attribute), Associated with (attribute), Causative agent (attribute), Pathological process (attribute), Associated morphology (attribute) (3 Partial-areas)
Bacterial pneumonia... (2)
Candidiasis of lung associate...(1)
Viral pneumonia associated wit...(1)
Finding site (attribute), Clinical course (attribute), Causative agent (attribute), Pathological process (attribute), Associated morphology (attribute) (4 Partial-areas)
Chronic coccidioidomyc (2)
Acute pulmonary African... (1)
Chronic pulmonary blastomycosis (1)
Pneumonia in pertussis (1)
Finding site (attribute), Causative agent (attribute), Pathological process (attribute), Occurrence (attribute), Associated morphology (attribute) (3 Partial-areas)
Congenital bacterial... (7)
Congenital viral pneumonia (1)
Neonatal chlamydial... (1)

Figure 3.13 The *Pneumonia* focus subtaxonomy.

The derivation of the focus subtaxonomy begins by identifying all of *c*'s ancestor and descendant concepts (along with the IS-A relationships between them). The partial-area taxonomy derivation algorithm described in Section 2.2.1.1 is applied to the subhierarchy consisting of all the ancestor and descendant concepts (and *c*). The result is a partial-area taxonomy that compactly summarizes all of *c*'s ancestors and descendants.

Figure 3.13 shows the focus subtaxonomy for the concept *Pneumonia*. The partial-areas where *Pneumonia* resides are outlined in red in the focus subtaxonomy, allowing an editor to quickly see how the subject concept is categorized. Alternatively the chosen subject concept can be displayed as a separate child node of its respective partial-areas. From the focus subtaxonomy in Figure 3.13 one can see that *Pneumonia* has 53 ancestors that are summarized by *Inflammation of specific body site*, 48 ancestors categorized under *Lung consolidation*, and several ancestors at smaller level partial-areas, e.g., 17 ancestors in *Finding by site*.

3.2 Tribal Abstraction Network

The derivation of area and partial-area taxonomies requires a hierarchy to have outgoing attribute relationships. Within SNOMED CT, twelve hierarchies have no relationships and serve only as targets for incoming relationships ("target hierarchies" for short). Thus, an alternative paradigm is required to derive abstraction networks for target hierarchies, specifically target hierarchies with concepts that have multiple parents. In SNOMED CT, 102,826 concepts (34.5%) have multiple parents and the average number of parents is 1.822.

Table 3.11 shows the number of concepts in each hierarchy having multiple parents as well as their percentage of each hierarchy. Eight of these 12 hierarchies contain more than 10 concepts with multiple parents. However, the numbers of concepts with multiple parents varies widely between different hierarchies. Almost half (45.26%) of the concepts in *Clinical finding* have multiple parents, compared to only 5.33% of the concepts in *Observable entity.*

Ochs et al. [27] introduced the *Tribal Abstraction Network* (TAN), a new type of abstraction network designed for SNOMED CT target hierarchies. The TAN is derived assuming only the existence of multiple parents in a hierarchy. The TAN can be used to summarize the content and structure of such SNOMED CT hierarchies, as well as support their quality assurance, by identifying groups of concepts with a higher likelihood of incorrect or missing IS-A relationships.

Table 3.11 A Breakdown by Hierarchy of Active SNOMED CT Concepts with Multiple Parents

Hierarchy	**# of Active Concepts**	**# w/ Multiple Parents**	**% of Hierarchy**
*Body structure**	31,117	13,339	42.9
*Clinical finding**	99,440	45,139	45.4
Environment or geographical location	1,712	28	1.6
*Event**	3,662	88	2.4
Linkage concept	1,131	0	0.0
Observable entity	8,274	439	5.3
Organism	32,776	1,195	3.6
*Pharmaceutical/biologic product**	17,146	7,727	45.1
Physical force	171	11	6.4
Physical object	4,522	383	8.5
*Procedure**	53,147	27,286	51.3
Qualifier value	8,984	750	8.4
Record artifact	223	2	0.9
*Situation with explicit context**	3,350	403	12.0
Social context	4,806	767	16.0
Special concept	802	0	0.0
*Specimen**	1,422	828	58.2
Staging and scales	1,305	1	0.08
Substance	23,822	4,445	18.7

Note: An asterisk indicates that the hierarchy has lateral relationships.
Source: [27]

3.2.1 Derivation

The TAN is derived as follows. The children of a hierarchy's root are named *patriarchs*. A *tribe* is defined as a subhierarchy consisting of a patriarch and all its descendants. The use of the words "tribe" and "patriarch" follows the family tree paradigm (e.g., parents, children, and siblings). A tribe is named after its patriarch, since all its concepts are specializations of the patriarch. Every concept in a hierarchy, except for the hierarchy root, belongs to at least one tribe. In a TAN, all concepts belonging to a common set of tribes are grouped together. A necessary but not sufficient condition for a hierarchy to have concepts in multiple tribes is that there are concepts with multiple parents.

These definitions are illustrated using an excerpt from the *Observable entity* target hierarchy, which consists of concepts “representing a question or procedure which can produce an answer or a result” [38]. In the January 2013 release of SNOMED CT, this hierarchy contained 8,274 concepts linked by 8,726 IS-A relationships.

Figure 3.14 shows a graphical representation for an excerpt of 20 concepts. Concepts are represented as nodes labeled with their respective names. Each of the three example children of *Observable entity*, i.e., *Process*, *Function*, and *Clinical history/examination observable* (shortened to *Clinical history/exam*), is a patriarch of a tribe. The tribal names are abbreviated P for *Process*, F for *Function*, and C for *Clinical history/exam* within braces below each name. Hierarchical IS-A links are represented as arrows. For example, *Digestive system function* IS-A *Function*. *Physiological action, Activity, Ingestion, Drinking, Feeding,* and *Breastfeeding (mother)* belong to the *Process* tribe since they are all descendants of *Process.*

Each concept is labeled by the set of tribes it belongs to, called its *tribal set*. To assign all concepts in a hierarchy to tribes, the hierarchy is traversed using *topological sort* [124] starting from the hierarchy’s patriarchs. Each patriarch is by definition only assigned its own tribe. In a topological sort procedure any non-patriarch concept is processed only after all of its parents have been processed.

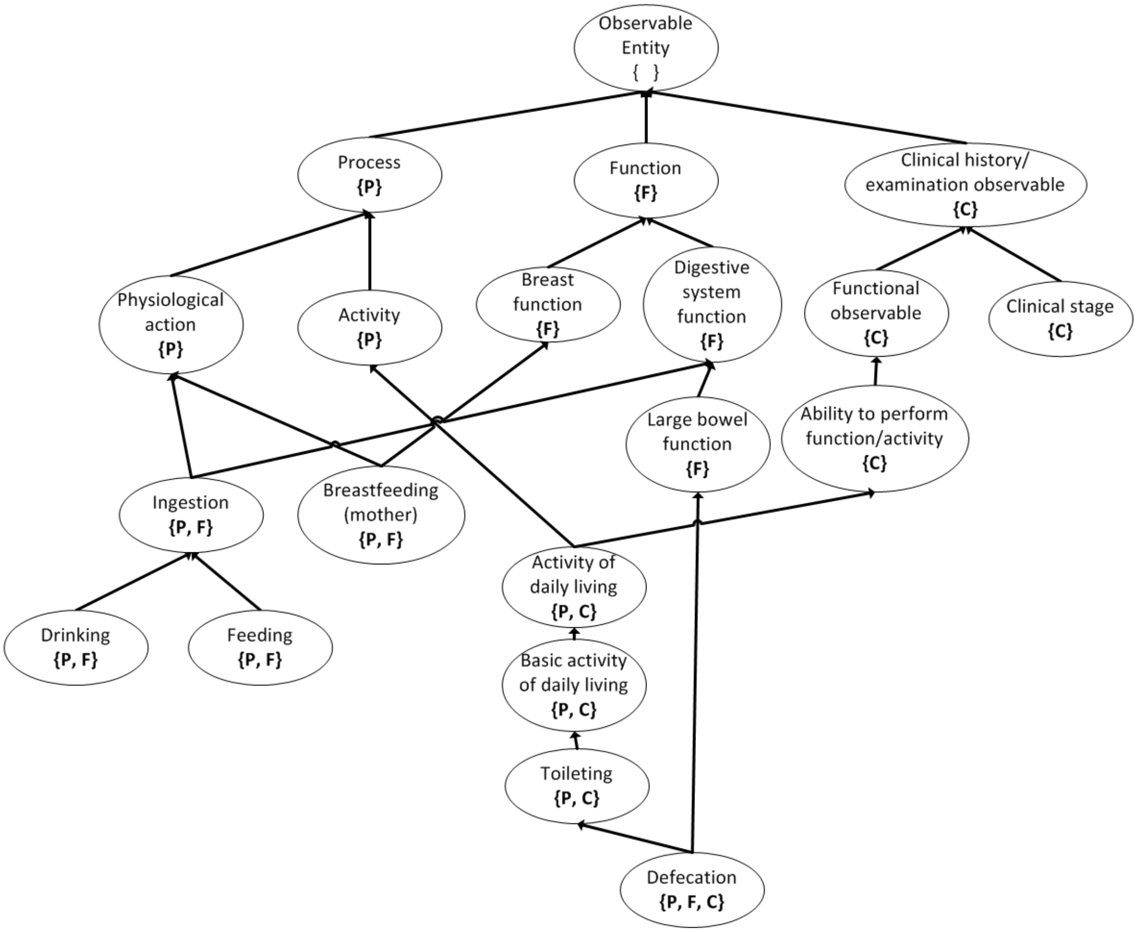

Observable Entity { }
Process {P}
Function {F}
Clinical history/ examination observable {C}
Physiological action {P}
Activity {P}
Breast function {F}
Digestive system function {F}
Functional observable {C}
Clinical stage {C}
Large bowel function {F}
Ability to perform function/activity {C}
Ingestion {P, F}
Breastfeeding (mother) {P, F}
Activity of daily living {P, C}
Drinking {P, F}
Feeding {P, F}
Basic activity of daily living {P, C}
Toileting {P, C}
Defecation {P, F, C}

Figure 3.14 An excerpt of 20 concepts from the *Observable entity* hierarchy with abbreviated tribal names in braces.
Source: [27]

If a concept *c* has one parent p_1 belonging to the tribe *A* and another parent p_2 belonging to the tribe *B*, *c* belongs to both tribes *A* and *B*, because it is a descendant of both patriarchs *A* and *B*. Once all parents of a concept *c* have been processed, *c* is assigned the union of its parents' tribal sets. This procedure is equivalent to, but generally more efficient than, performing a separate graph traversal from each patriarch, since each concept is only processed once. If a standard graph traversal, such as breadth first search [124] were performed from each patriarch, concepts would have been processed multiples times, according to the number of tribes they belong to. For example, *Defecation* would have been processed three times, instead of only once using topological sort.

Figure 3.14 shows the results of applying the tribe assignment process for an excerpt of 20 concepts. Tribal sets are shown in braces below each concept's name. Figure 3.15 groups together the concepts with identical tribal sets. Each group is represented by a dashed bubble and contains the name(s) of the tribes, separated by commas.

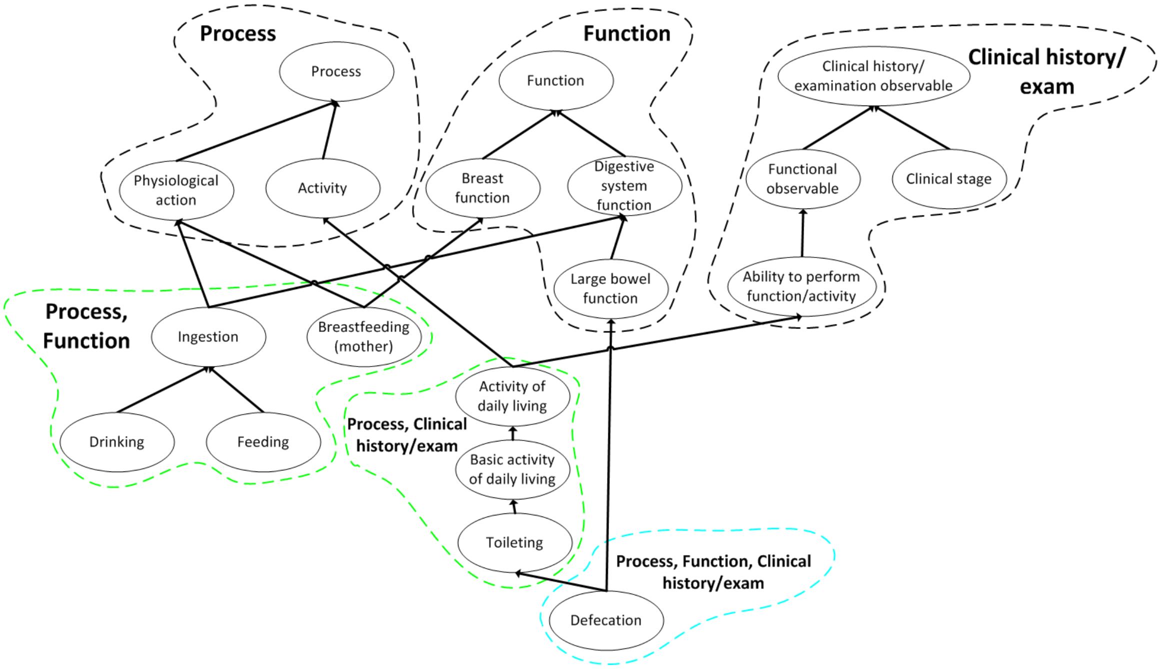


Figure 3.15 The concepts from Figure 3.14 grouped by common tribal sets.
Source: [27]

Concepts that are descendants of only one patriarch belong to only one tribe. In Figure 3.15 *Large bowel function* belongs only to the *Function* tribe. On the other hand, (Figure 3.15), *Ingestion*, *Breastfeeding (mother)*, *Activity of daily living*, and *Defecation* all belong to more than one tribe, because each has multiple parents in different tribes.

For example, *Ingestion* has two parents, *Physiological action* and *Digestive system function*, which belong to the *Process* and *Function* tribes, respectively. *Ingestion*, therefore, belongs to both the *Process* and *Function* tribes. *Defecation* belongs to all three tribes of this hierarchy.

Even though *Drinking, Feeding, Basic activity of daily living* and *Toileting* each have only one parent, they belong to multiple tribes because each has an ancestor that belongs to multiple tribes.

Generally, concepts that belong to more than one tribe are more complex than those belonging to only one tribe, since they are specializations of several patriarch concepts. A concept that belongs to multiple tribes is called a *joint* concept. Joint-ness can be used to group concepts into sets. These sets can be used to derive two kinds of Tribal Abstraction Networks: the *Band Tribal Abstraction Network* (“Band TAN”) and the more refined *Cluster Tribal Abstraction Network* (“Cluster TAN”).

3.2.1.1 Band Tribal Abstraction Network. A tribal band, or *band* for short, is a set of all concepts that are members of the exact same tribes. A band is named after the set of tribes each concept within the band belongs to. A root of a band is a concept that has no parents within the band, though it may have parents in other bands. A band may have multiple roots. Each set of concepts, surrounded by a dashed bubble (Figure 3.15), defines a band.

A band TAN consists of one node for each band. These nodes are linked by hierarchical *child-of* links derived from the underlying IS-A hierarchy of the terminology.

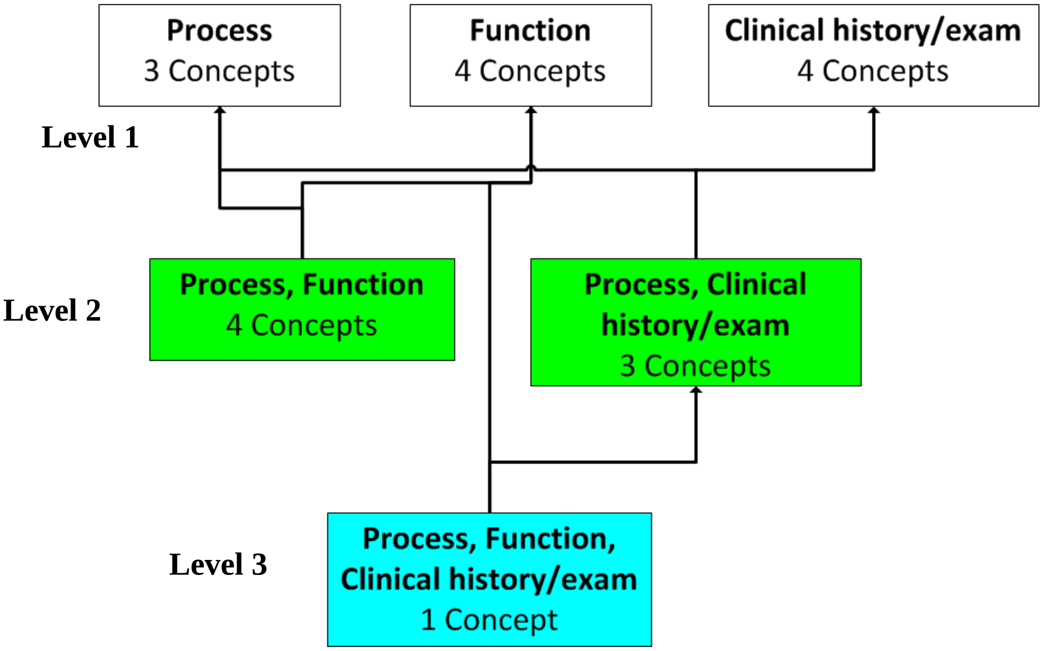


Figure 3.16 The band TAN derived from Figure 3.15. Each box represents a band. *Child-of* links are represented using arrows between bands.
Source: [27]

A band *A* is a *child-of* another band *B* if and only a root concept in *A* has an IS-A link to a concept in *B*. A band may be *child-of* multiple bands. The band TAN provides a compact, abstract view of a target hierarchy.

Figure 3.16 shows the band TAN for Figure 3.15 obtained using the tribal sets from Figure 3.14. The number of concepts is listed under each band's name. The four concepts *Ingestion*, *Feeding*, *Drinking*, and *Breastfeeding (mother)* belong to the band named {*Process*, *Function*}. *Ingestion* and *Breastfeeding (mother)* are the roots of the {*Process, Function*} band, because neither has parents in the {*Process, Function*} band. The band {*Process, Function*} is a *child-of* two bands, {*Process*} and {*Function*}, because both roots *Ingestion* and *Breastfeeding (mother)* have parents in both of these bands. By the definition of the band TAN, roots are not displayed in Figure 3.16.

The band {*Process, Function, Clinical history/exam*} is a *child-of* both bands {*Process, Clinical history/exam*} and {*Function*} because its root *Defecation* has two parents, *Toileting* in {*Process, Clinical history/exam*} and *Large bowel function* in {*Function*}.

Each band has a degree of “joint-ness” according to the number of tribes its members belong to. Bands containing concepts of only one tribe consist of the tribal patriarch and all of its descendants which are not descendants of a second patriarch.

In visualizations of band TANs (Figures 3.16 and 3.18), tribal bands are organized into levels according to their degrees of joint-ness and are color-coded. Bands of degree 1 are located at the top of the figure. Bands of degree 2, with concepts that belong to two tribes, are below.

3.2.1.2 Cluster Tribal Abstraction Network. A tribal band may have multiple roots. Each root defines a different subhierarchy of concepts within the band. A *tribal cluster*, or *cluster* for short, consists of one root of a band and all its descendants within the same band. A tribal cluster is named after its root, because all other concepts in the cluster are specializations of the root.

Clusters are used to further refine the band TAN into the *cluster TAN*. In a cluster TAN, the clusters serve as the nodes, where all the clusters of a band are drawn within that band node. Clusters, like bands, are linked by *child-of* relationships based on the underlying IS-A hierarchy. A cluster *A* is a *child-of* another cluster *B* if the root concept of *A* has an IS-A link to any concept in *B*. A cluster may be a *child-of* of multiple clusters. Clusters are not necessarily disjoint in terms of the concepts they summarize. A given concept may be summarized by more than one cluster.

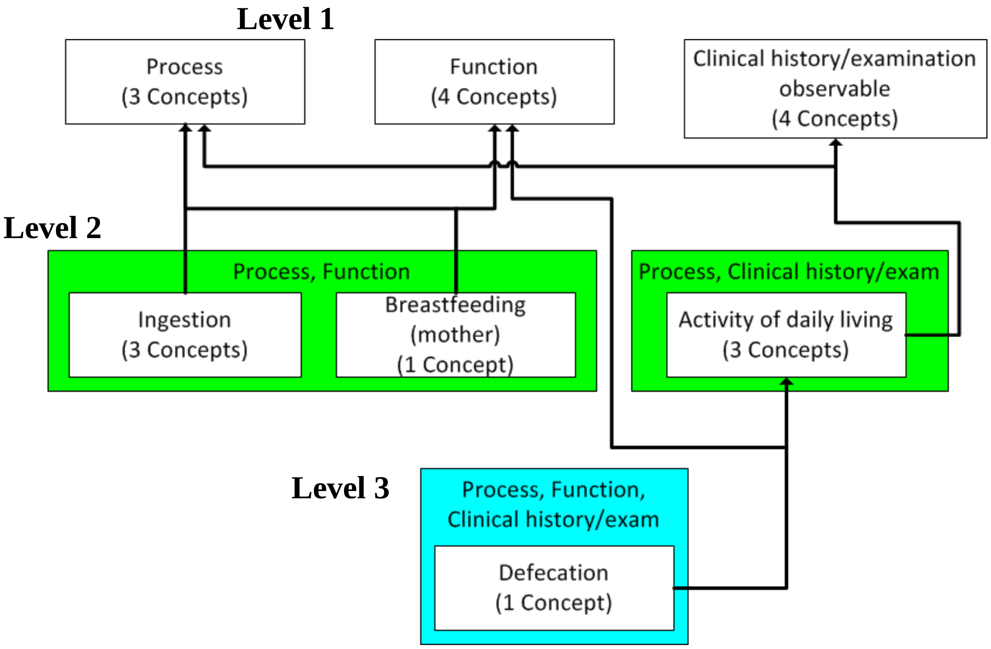


Figure 3.17 The cluster TAN derived from Figure 3.15. *Child-of* links are represented by arrows between clusters.
Source: [27]

In Figure 3.15, *Ingestion* and *Breastfeeding (mother)* are the two roots of the {*Process*, *Function*} band. In visualizations of a cluster TAN (Figures 3.16 and 3.19), clusters are represented as white boxes within a band box, labeled by their roots, with their numbers of concepts below the root names. The root *Ingestion* and its two descendants are represented as a cluster named *Ingestion* with three concepts in the {*Process*, *Function*} band (Figure 3.17). The *Ingestion* cluster is a *child-of* the *Process* and *Function* clusters because the root concept *Ingestion* has parents in these two clusters.

3.2.2 In Support of Quality Assurance

Quality assurance of large terminologies is difficult and time consuming. By focusing efforts on a subset of concepts that are likely to be more error prone, quality assurance resources can be utilized more effectively. TANs can be used to support SNOMED CT quality assurance efforts by identifying concepts more likely to have more hierarchical errors. Such errors were deemed to be the most problematic in a study of SNOMED CT's users [122]. IS-A relationships play an important definitional role for concepts in SNOMED CT. For target hierarchies the correctness of the IS-A hierarchy is important because the concepts of these hierarchies serve as targets for relationships with source concepts in other hierarchies. There are 18,839 attribute relationships with targets in *Observable entity*. Proper placement of target concepts in a hierarchy is crucial since the target of a relationship should be as specific as possible.

> *Hypothesis 1:* In a cluster TAN, concepts in large clusters are more likely to have errors than concepts in small clusters.

The rationale for Hypothesis 1 is as follows. For a concept in a target hierarchy (without relationships) to be erroneous, the errors can occur only in the hierarchy. An IS-A relationship for a concept may be either wrong or missing and the concept is misplaced in the hierarchy. There is a greater chance for such situations to occur in large clusters, because as the number of hierarchically closely related concepts increases, the chance of a concept being misplaced in the hierarchy also increases. In clusters with fewer concepts, there is less chance of a concept being misplaced in the hierarchy.

To reiterate, the goal is to minimize the number of concepts that should be the focus of a quality assurance review by selecting few concepts with a high likelihood of errors. Such a portion can be reviewed with available limited quality assurance resources and yield a large number of errors, relative to the effort spent. However, auditing all large clusters is generally not practical because of their large number of concepts. Therefore, a second hypothesis is introduced based on the level a concept belongs to. (Reminder: Level numbers grow higher when moving downward in a band diagram.)

Hypothesis 2: Among the large clusters, those concepts belonging to higher-numbered levels are more likely to be erroneous.

The rationale for this hypothesis is that concepts belonging to more tribes tend to be more complex due to their specialization of more patriarchs. The modeling of more complex concepts is more prone to errors. Assuming there is support for these two hypotheses, the following auditing methodology emerges. Start reviewing the large clusters of the highest-numbered levels. As long as quality assurance resources remain, continue to review large clusters moving up in the TAN.

To test both hypotheses, a cluster TAN was derived for the July 2011 version of the *Observable entity* hierarchy. Even though *Observable entity* has few concepts with multiple parents (Table 3.11), a cluster TAN summarizes the content and structure of this hierarchy well (Table 3.12). There are 27 children of *Observable entity* and therefore 27 tribes with 16 (59.3%) of these tribes having joint concepts while 11 tribes do not. The maximum number of tribes a concept belongs to is three, while 6,627 (80.5%) concepts of a unique tribe belong to the 27 tribal bands on the first level. The second level comprises 1,236 concepts (15%) of the hierarchy and the third level 368 (4.47%). The percentage of concepts with multiple parents is much higher in Levels 2 and 3 (14% and 20%) than in Level 1 (2.5%). Figures 3.18 and 3.19 provide visualizations of the band TAN and the cluster TAN, respectively.

The TAN summarizes a target hierarchy. The bands of Level 1 indicate the major types of concepts in a hierarchy; Level 1 of Figure 3.18 contains many *Clinical* *history/examination* and *Function* concepts. Levels 2 and 3 show how the bands of Level 1 intersect in the hierarchy, e.g., the *Clinical history/examination* band intersects with most other bands. Figure 3.19 allows identifying common concept groups of multiple tribes. For example, looking at the very large clusters, such as *Female genital feature* (152), *Cardiac feature (145), Eye observable (143)*, followed by the large clusters *Blood pressure* (86), and *Activity of daily living* (79), *Joint movement (86)*, *Feature of lower limb (84),* and *Feature of upper limb (84)*, provides a summarization of the major types of concepts in the *Observable entity* hierarchy.

Table 3.12 Metrics for the Three Levels of the *Observable entity* Hierarchy's Band and Cluster Tribal Abstraction Networks

Level	**# of Bands**	**# of Clusters**	**# (%) of Concepts w/ Multiple Parents**	**Avg # of Parents**	**# of Concepts**
1	27	27	169 (2.5%)	1.03	6,643
2	23	101	170 (14%)	1.14	1,220
3	13	52	73 (20%)	1.21	368
TOTA	63	180	412 (5.3%)	1.06	8231

Source: [27]

For a finer summary, one should view the "medium" sized clusters of 25-50 concepts, e.g., *Device of eye observable* (39), *Tumor size (35), Shoulder joint – range of movement* (28), and *Anesthetic agent concentration (26)*. Hence, by looking at the 15 clusters with at least 25 concepts, the TAN summarizes 1084 concepts (68.3%) of the major subjects in Levels 2 and 3.

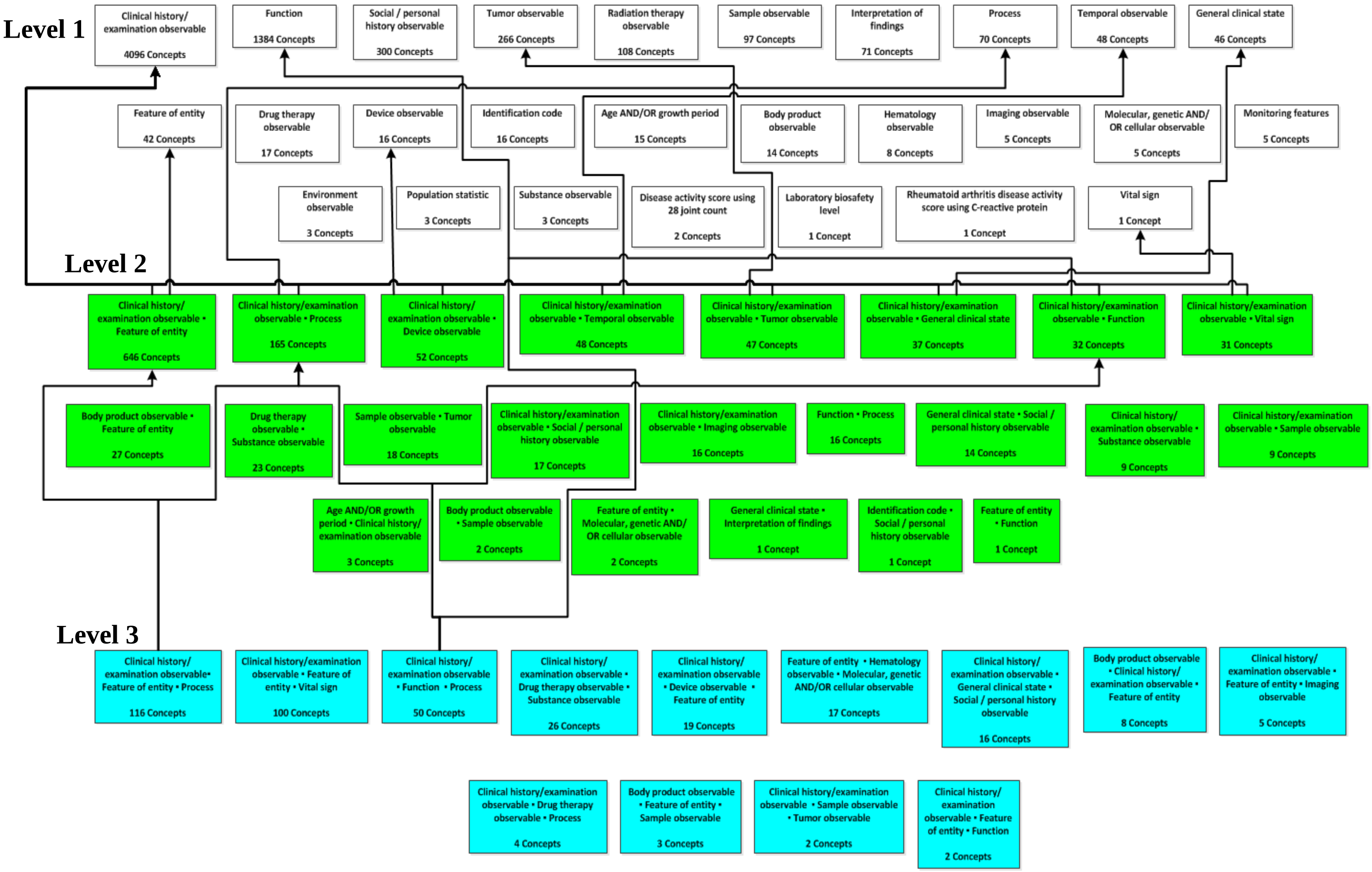

Level 1
Clinical history/ examination observable 4096 Concepts
Function 1384 Concepts
Social / personal history observable 300 Concepts
Tumor observable 266 Concepts
Radiation therapy observable 108 Concepts
Sample observable 97 Concepts
Interpretation of findings 71 Concepts
Process 70 Concepts
Temporal observable 48 Concepts
General clinical state 46 Concepts
Feature of entity 42 Concepts
Drug therapy observable 17 Concepts
Device observable 16 Concepts
Identification code 16 Concepts
Age AND/OR growth period 15 Concepts
Body product observable 14 Concepts
Hematology observable 8 Concepts
Imaging observable 5 Concepts
Molecular, genetic AND/ OR cellular observable 5 Concepts
Monitoring features 5 Concepts
Environment observable 3 Concepts
Population statistic 3 Concepts
Substance observable 3 Concepts
Disease activity score using 28 joint count 2 Concepts
Laboratory biosafety level 1 Concept
Rheumatoid arthritis disease activity score using C-reactive protein 1 Concept
Vital sign 1 Concept
Level 2
Clinical history/ examination observable • Feature of entity 646 Concepts
Clinical history/examination observable • Process 165 Concepts
Clinical history/ examination observable • Device observable 52 Concepts
Clinical history/examination observable • Temporal observable 48 Concepts
Clinical history/examination observable • Tumor observable 47 Concepts
Clinical history/examination observable • General clinical state 37 Concepts
Clinical history/examination observable • Function 32 Concepts
Clinical history/examination observable • Vital sign 31 Concepts
Body product observable • Feature of entity 27 Concepts
Drug therapy observable • Substance observable 23 Concepts
Sample observable • Tumor observable 18 Concepts
Clinical history/examination observable • Social / personal history observable 17 Concepts
Clinical history/examination observable • Imaging observable 16 Concepts
Function • Process 16 Concepts
General clinical state • Social / personal history observable 14 Concepts
Clinical history/ examination observable • Substance observable 9 Concepts
Clinical history/examination observable • Sample observable 9 Concepts
Age AND/OR growth period • Clinical history/ examination observable 3 Concepts
Body product observable • Sample observable 2 Concepts
Feature of entity • Molecular, genetic AND/ OR cellular observable 2 Concepts
General clinical state • Interpretation of findings 1 Concept
Identification code • Social / personal history observable 1 Concept
Feature of entity • Function 1 Concept
Level 3
Clinical history/ examination observable• Feature of entity • Process 116 Concepts
Clinical history/examination observable • Feature of entity • Vital sign 100 Concepts
Clinical history/ examination observable • Function • Process 50 Concepts
Clinical history/ examination observable • Drug therapy observable • Substance observable 26 Concepts
Clinical history/ examination observable • Device observable • Feature of entity 19 Concepts
Feature of entity • Hematology observable • Molecular, genetic AND/OR cellular observable 17 Concepts
Clinical history/ examination observable • General clinical state • Social / personal history observable 16 Concepts
Body product observable • Clinical history/ examination observable • Feature of entity 8 Concepts
Clinical history/ examination observable • Feature of entity • Imaging observable 5 Concepts
Clinical history/examination observable • Drug therapy observable • Process 4 Concepts
Body product observable • Feature of entity • Sample observable 3 Concepts
Clinical history/examination observable • Sample observable • Tumor observable 2 Concepts
Clinical history/ examination observable • Feature of entity • Function 2 Concepts

Figure 3.18 The band tribal abstraction network for the *Observable entity* hierarchy. Levels are organized into rows due to space limitations. Some *child-of* edges are hidden for readability. *Source: [27]*

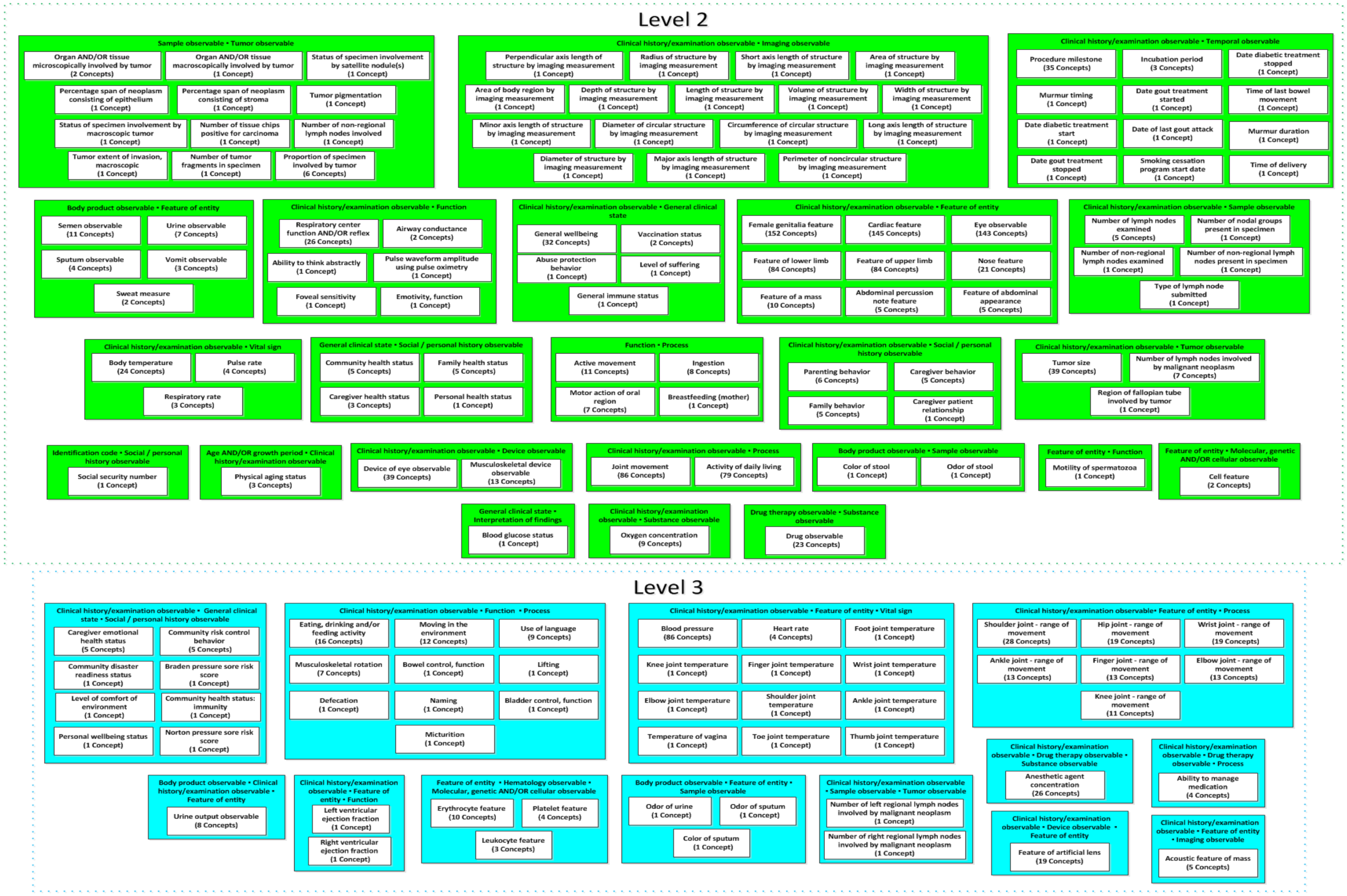

Level 2
Sample observable • Tumor observable
Organ AND/OR tissue microscopically involved by tumor (2 Concepts)
Organ AND/OR tissue macroscopically involved by tumor (1 Concept)
Status of specimen involvement by satellite nodule(s) (1 Concept)
Percentage span of neoplasm consisting of epithelium (1 Concept)
Percentage span of neoplasm consisting of stroma (1 Concept)
Tumor pigmentation (1 Concept)
Status of specimen involvement by macroscopic tumor (1 Concept)
Number of tissue chips positive for carcinoma (1 Concept)
Number of non-regional lymph nodes involved (1 Concept)
Tumor extent of invasion, macroscopic (1 Concept)
Number of tumor fragments in specimen (1 Concept)
Proportion of specimen involved by tumor (6 Concepts)
Clinical history/examination observable • Imaging observable
Perpendicular axis length of structure by imaging measurement (1 Concept)
Radius of structure by imaging measurement (1 Concept)
Short axis length of structure by imaging measurement (1 Concept)
Area of structure by imaging measurement (1 Concept)
Area of body region by imaging measurement (1 Concept)
Depth of structure by imaging measurement (1 Concept)
Length of structure by imaging measurement (1 Concept)
Volume of structure by imaging measurement (1 Concept)
Width of structure by imaging measurement (1 Concept)
Minor axis length of structure by imaging measurement (1 Concept)
Diameter of circular structure by imaging measurement (1 Concept)
Circumference of circular structure by imaging measurement (1 Concept)
Long axis length of structure by imaging measurement (1 Concept)
Diameter of structure by imaging measurement (1 Concept)
Major axis length of structure by imaging measurement (1 Concept)
Perimeter of noncircular structure by imaging measurement (1 Concept)
Clinical history/examination observable • Temporal observable
Procedure milestone (35 Concepts)
Incubation period (3 Concepts)
Date diabetic treatment stopped (1 Concept)
Murmur timing (1 Concept)
Date gout treatment started (1 Concept)
Time of last bowel movement (1 Concept)
Date diabetic treatment start (1 Concept)
Date of last gout attack (1 Concept)
Murmur duration (1 Concept)
Date gout treatment stopped (1 Concept)
Smoking cessation program start date (1 Concept)
Time of delivery (1 Concept)
Body product observable • Feature of entity
Semen observable (11 Concepts)
Urine observable (7 Concepts)
Sputum observable (4 Concepts)
Vomit observable (3 Concepts)
Sweat measure (2 Concepts)
Clinical history/examination observable • Function
Respiratory center function AND/OR reflex (26 Concepts)
Airway conductance (2 Concepts)
Ability to think abstractly (1 Concept)
Pulse waveform amplitude using pulse oximetry (1 Concept)
Foveal sensitivity (1 Concept)
Emotivity, function (1 Concept)
Clinical history/examination observable • General clinical state
General wellbeing (32 Concepts)
Vaccination status (2 Concepts)
Abuse protection behavior (1 Concept)
Level of suffering (1 Concept)
General immune status (1 Concept)
Clinical history/examination observable • Feature of entity
Female genitalia feature (152 Concepts)
Cardiac feature (145 Concepts)
Eye observable (143 Concepts)
Feature of lower limb (84 Concepts)
Feature of upper limb (84 Concepts)
Nose feature (21 Concepts)
Feature of a mass (10 Concepts)
Abdominal percussion note feature (5 Concepts)
Feature of abdominal appearance (5 Concepts)
Clinical history/examination observable • Sample observable
Number of lymph nodes examined (5 Concepts)
Number of nodal groups present in specimen (1 Concept)
Number of non-regional lymph nodes examined (1 Concept)
Number of non-regional lymph nodes present in specimen (1 Concept)
Type of lymph node submitted (1 Concept)
Clinical history/examination observable • Vital sign
Body temperature (24 Concepts)
Pulse rate (4 Concepts)
Respiratory rate (3 Concepts)
General clinical state • Social / personal history observable
Community health status (5 Concepts)
Family health status (5 Concepts)
Caregiver health status (3 Concepts)
Personal health status (1 Concept)
Function • Process
Active movement (11 Concepts)
Ingestion (8 Concepts)
Motor action of oral region (7 Concepts)
Breastfeeding (mother) (1 Concept)
Clinical history/examination observable • Social / personal history observable
Parenting behavior (6 Concepts)
Caregiver behavior (5 Concepts)
Family behavior (5 Concepts)
Caregiver patient relationship (1 Concept)
Clinical history/examination observable • Tumor observable
Tumor size (39 Concepts)
Number of lymph nodes involved by malignant neoplasm (7 Concepts)
Region of fallopian tube involved by tumor (1 Concept)
Identification code • Social / personal history observable
Social security number (1 Concept)
Age AND/OR growth period • Clinical history/examination observable
Physical aging status (3 Concepts)
Clinical history/examination observable • Device observable
Device of eye observable (39 Concepts)
Musculoskeletal device observable (13 Concepts)
Clinical history/examination observable • Process
Joint movement (86 Concepts)
Activity of daily living (79 Concepts)
Body product observable • Sample observable
Color of stool (1 Concept)
Odor of stool (1 Concept)
Feature of entity • Function
Motility of spermatozoa (1 Concept)
Feature of entity • Molecular, genetic AND/OR cellular observable
Cell feature (2 Concepts)
General clinical state • Interpretation of findings
Blood glucose status (1 Concept)
Clinical history/examination observable • Substance observable
Oxygen concentration (9 Concepts)
Drug therapy observable • Substance observable
Drug observable (23 Concepts)
Level 3
Clinical history/examination observable • General clinical state • Social / personal history observable
Caregiver emotional health status (5 Concepts)
Community risk control behavior (5 Concepts)
Community disaster readiness status (1 Concept)
Braden pressure sore risk score (1 Concept)
Level of comfort of environment (1 Concept)
Community health status: immunity (1 Concept)
Personal wellbeing status (1 Concept)
Norton pressure sore risk score (1 Concept)
Clinical history/examination observable • Function • Process
Eating, drinking and/or feeding activity (16 Concepts)
Moving in the environment (12 Concepts)
Use of language (9 Concepts)
Musculoskeletal rotation (7 Concepts)
Bowel control, function (1 Concept)
Lifting (1 Concept)
Defecation (1 Concept)
Naming (1 Concept)
Bladder control, function (1 Concept)
Micturition (1 Concept)
Clinical history/examination observable • Feature of entity • Vital sign
Blood pressure (86 Concepts)
Heart rate (4 Concepts)
Foot joint temperature (1 Concept)
Knee joint temperature (1 Concept)
Finger joint temperature (1 Concept)
Wrist joint temperature (1 Concept)
Elbow joint temperature (1 Concept)
Shoulder joint temperature (1 Concept)
Ankle joint temperature (1 Concept)
Temperature of vagina (1 Concept)
Toe joint temperature (1 Concept)
Thumb joint temperature (1 Concept)
Clinical history/examination observable• Feature of entity • Process
Shoulder joint - range of movement (28 Concepts)
Hip joint - range of movement (19 Concepts)
Wrist joint - range of movement (19 Concepts)
Ankle joint - range of movement (13 Concepts)
Finger joint - range of movement (13 Concepts)
Elbow joint - range of movement (13 Concepts)
Knee joint - range of movement (11 Concepts)
Clinical history/examination observable • Drug therapy observable • Substance observable
Anesthetic agent concentration (26 Concepts)
Clinical history/examination observable • Drug therapy observable • Process
Ability to manage medication (4 Concepts)
Body product observable • Clinical history/examination observable • Feature of entity
Urine output observable (8 Concepts)
Clinical history/examination observable • Feature of entity • Function
Left ventricular ejection fraction (1 Concept)
Right ventricular ejection fraction (1 Concept)
Feature of entity • Hematology observable • Molecular, genetic AND/OR cellular observable
Erythrocyte feature (10 Concepts)
Platelet feature (4 Concepts)
Leukocyte feature (3 Concepts)
Body product observable • Feature of entity • Sample observable
Odor of urine (1 Concept)
Odor of sputum (1 Concept)
Color of sputum (1 Concept)
Clinical history/examination observable • Sample observable • Tumor observable
Number of left regional lymph nodes involved by malignant neoplasm (1 Concept)
Number of right regional lymph nodes involved by malignant neoplasm (1 Concept)
Clinical history/examination observable • Device observable • Feature of entity
Feature of artificial lens (19 Concepts)
Clinical history/examination observable • Feature of entity • Imaging observable
Acoustic feature of mass (5 Concepts)

Figure 3.19 The cluster tribal abstraction network for *Observable entity*. *Child-of* edges are hidden for readability. Each level is organized into several rows due to space limitations. Level 1 (not shown) is the same as in Figure 3.18. *Source: [27]*

To test Hypothesis 1, a domain expert reviewed 1,160 concepts (14.1%) from *Observable entity*. The domain expert audited 410 concepts from Level 1; 477 from Level 2; and 266 from Level 3. At each level the domain expert audited all concepts from clusters of nine concepts or fewer (284 in total) and randomly selected 876 concepts from clusters containing 10 or more concepts. In total, the domain expert found 39 errors (3.36%) in the sample. Twenty-one concepts had incorrect IS-A relationships and 18 had missing IS-A relationships. These errors were submitted to the curator of the SNOMED CT US Extension at the National Library of Medicine for review and inclusion in the International Release of SNOMED CT. Only three corrections were not accepted by the US Extension's curator and all but one of the corrections was accepted by the IHTSDO.

For the 39 erroneous concepts, a total of 42 errors were found. These erroneous concepts served as targets for 42 different relationships from source hierarchies. A follow up review of these erroneous concepts using the January 2013 release of SNOMED CT was performed and all of the errors were still present.

To test Hypothesis 1, the relationship between cluster size and error rate was studied as follows. To handle correlation of concepts within clusters, the data was analyzed at the cluster level by calculating the error rate per cluster (i.e., for each cluster, the total number of erroneous concepts divided by the total number of sample concepts in the cluster). To better visualize the effect of cluster size, and because the relation between cluster size and error rate might not be linear, the clusters were stratified into six bins.

Table 3.13 shows the distribution of clusters, concepts, sample concepts, and erroneous concepts among the six bins. The mean cluster error rate column shows the average error rate of clusters in each bin.

Table 3.13 The Distribution of Concepts, Errors, and Error Rates Among the Six Bins

Bin	Cluster Size	Clusters	Concepts	Concepts/ Clusters	Samples	Erroneous Concepts (%)	Mean cluster error rate
1	> 150	5	6,198	1239.6	219	10 (4.56%)	5.1%
2	86-150	6	665	110.83	221	16 (7.24%)	4.3%
3	46-85	7	482	68.86	186	3 (1.08%)	1%
4	11-45	27	572	21.19	231	5 (2.16%)	1%
5	2-10	46	225	5	214	3 (1.40%)	1.8%
6	1	89	89	1	89	2 (2.25%)	2.3%
Tot.		180	8,231	45.98	1160	39 (3.36%)	2.0%

Source: [27]

The error rates and 95% confidence intervals versus cluster size were calculated between all bins. Bin 1 (clusters with more than 150 concepts) had an error rate statistically significantly higher than Bin 3 (46-85 concepts) and Bin 4 (clusters with 11-45 concepts), with *p*=0.019 and *p*=0.009, respectively. Furthermore, Bin 2 (86-150 concepts) had an error rate statistically significantly higher than Bin 4 (*p*=0.039). Error rates between other pairs of bins were not significantly different. However, in general, Bin 1 and 2 clusters have higher mean error rates than clusters in Bins 3-6.

To test Hypothesis 2 the mean error rates among the “large” clusters in the three levels was analyzed. Various boundaries between small and large were tested. No boundary resulted in significance due to the relatively small number of “large” clusters in the cluster TAN, e.g., there are no Bin 1 clusters and just one Bin 2 cluster at Level 3. However, it is observed that, at higher levels, larger clusters tended to have higher error rates. Using the result from Hypothesis 1, Bin 1 or 2 clusters are treated as large clusters.

Table 3.14 An Analysis of Bins 1, 2 ("Large Clusters") Broken Down by Level

	Large Clusters (Bins 1, 2)				
Level	**# of Clusters**	**# of Concepts**	**# of Sample Concepts**	**# of Erroneous Concepts (%)**	**Mean Cluster Error Rate**
1	6	6,251	183	6 (3.28%)	3.08%
2	4	526	171	9 (5.26%)	4.95%
3	1	86	86	11 (12.8%)	12.79%
Total	11	6,863	440	26 (5.9%)	4.64%

Source: [27]

Table 3.15 An Analysis of Bins 3-6 ("small clusters") Broken Down by Level

	Small Clusters (Bins 3-6)				
Level	**# of Clusters**	**# of Concepts**	**# of Sample Concepts**	**# of Erroneous Concepts (%)**	**Mean Cluster Error Rate**
1	21	392	237	7 (2.95%)	1.11%
2	97	694	303	4 (1.32%)	1.88%
3	51	282	180	2 (1.11%)	2.12%
Total	169	1,368	720	13 (1.81%)	1.86%

Source: [27]

Table 3.16 A Sample of Five Errors Taken from the Auditing Results

Concept(s)	Error	Suggested solution
Sitting systolic blood pressure and *Sitting diastolic blood pressure*	**Missing parent:** *Sitting blood pressure*	Add *IS-A* relationships from *sitting systolic blood pressure* and *sitting diastolic blood pressure* to *Sitting blood pressure*.
Ankle joint temperature	**Incorrect parent:** *Body temperature*	Replace *IS-A* to *Body temperature* by *IS-A* to *Joint temperature*
Date chemotherapy completed	**Missing parent:** *Temporal observable*	Add *IS-A* to *Temporal observable*.
Dorsalis pedis arterial pressure	**Incorrect parent:** *Blood pressure*	Replace *IS-A* to *Blood pressure* by *IS-A* to *Arterial blood pressure*
Autonomic bladder function	**Missing parent:** *Bladder function*	Add *IS-A* to *Bladder function*

Source: [27]

Tables 3.14 and 3.15 provide a breakdown of auditing results by level and by large vs. small clusters. It is observed that higher leveled large clusters have a higher error rate. For example, the single Level 3 large cluster has a mean error rate of 12.79, Level 2 large clusters have a mean error rate of 4.95%, and Level 1 large clusters have a mean error rate of 3.08%. A similar trend is observed in the small clusters (e.g., small Level 3 clusters have a slightly higher mean error rate than Level 2 clusters). For large clusters, the error rate among concepts (E) also increases with their level (i.e., 3.28%, 5.26%, and 12.8%). Table 3.16 provides five examples of errors identified.

3.2.2.1 Comparative Quality Assurance Study. The primary goal of the TAN-based quality assurance methodology is to identify groups of concepts within a SNOMED CT hierarchy without attribute relationships that are statistically more likely to be erroneous than other concepts. Auditors should focus quality assurance efforts on these concepts to increase auditing yields, as measured by the number of erroneous concepts corrected versus total number of concepts reviewed. It was observed that concepts in larger clusters (e.g., in Bins 1 and 2) were statistically significantly more likely to contain errors than concepts in smaller clusters in the *Observable entity* hierarchy. Furthermore, concepts in larger clusters at higher indexed levels were more likely to be erroneous than concepts in large clusters at lower indexed levels.

To compare the effectiveness of the TAN-based quality assurance methodology with other quality assurance methodologies, three other methods were applied to the January 2013 release *of the Observable entity* hierarchy. Note that no quality assurance techniques that use attribute relationships can be applied to a hierarchy without attribute relationships. For each methodology 200 sample concepts and a 100 control concepts were audited by a domain expert. In total, 832 unique concepts were reviewed for errors (there was some overlap among the samples due to random sampling). When a randomly sampled concept was previously reviewed for the TAN quality assurance study the result of the previous audit was used. The auditing was conducted by the same auditor from the TAN study. Each of these three techniques, one lexical, one fully-specified-name-based, and one hierarchy-based, supposedly identify concepts with a higher expected error ratio than other concepts. The last two identify complex concepts, as does the TAN.

3.2.2.1.1 Lexical Containment. The first approach involved testing *lexical containment* between pairs of concepts. This approach is used to identify missing IS-A relationships in the following way. Given two concepts c_1 and c_2, if c_1's fully specified name lexically contains c_2's fully specified name (the words consists of a subsequence and stop words are removed), then c_1 may be a descendant of c_2. For example *Intestinal absorption, function* is lexically contained within *Intestinal protein absorption, function* so the first concept should be a parent (or an ancestor) of the second concept. A total of 2,942 such pairs of concepts were identified in the *Observable entity* hierarchy. Only concepts with four or more significant words, including semantic tag, were considered.

A total of 5,884 concepts (3,058 unique concepts) were considered in the 2,942 pairs. Of the 2,942 pairs, 1,811 pairs were found to have an ancestor-successor relationship, e.g., for *Intestinal protein absorption, function*. The remaining 1,131 pairs did not have an ancestor-successor relationship. A random sample of 200 such pairs was audited to check if the base concept of each pair should be an ancestor of the second concept in the pair. A concept may be lexically contained in more than one concept. In the sample there were 188 unique base concepts. A random sample of 100 concepts that are *not* in a lexical containment pair was audited as a control sample.

Table 3.17 Lexical Containment Audit Results

	# of Erroneous Concepts (%)	**# of Sample Concepts**
Control	3 (3%)	100
Lexical containment pairs	40 (20%)	200

Source: [27]

Table 3.17 summarizes the results of the lexical containment review. Only three erroneous concepts were found in the control sample. The control sample included all concepts that are not a base concept in a lexical pair, thus all of the errors found involved missing IS-A relationships where the erroneous concept was not lexically contained in the parent (e.g., *Behavior to maintain weight* should have the parent *Weight control behavior*). Among the 200 lexical containment pairs there was a missing or incorrect ancestor-successor relationship for 40 of the pairs (each with a unique base concept). Table **3.18** provides three examples of missing ancestor-successor pairs uncovered during the review.

Table 3.18 Three Examples of Errors Found in the Lexical Containment Pairs

Concept	**Error Type**	**Proposed Solution**
Adolescent/adult sensory profile score	Incorrect parent: *Functional observable*	Replace with IS-A to more specific concept *Sensory profile score*
Temporomandibular joint stability	Missing parent	Add IS-A to *Joint stability*
Cerebrospinal fluid pressure observable	Missing parent	Add IS-A to *Fluid pressure*

Source: [27]

3.2.2.1.2 Number of Parents. The second method is based on the hypothesis that concepts with a greater number of parents are more likely to be erroneous than concepts with relatively few parents [128], which was confirmed for the Problem List of SNOMED CT by Agrawal et al. [128]. A concept with a relatively large number of parents implies complexity. Complex concepts should have a higher probability of being erroneous than concepts which are less complex.

The *Observable entity* hierarchy has 7,835 concepts with 1 parent, 426 concepts with 2 parents and 13 concepts with 3 parents. All 13 concepts with three parents, a random sample of 187 concepts with two parents, and a random sample of 100 concepts with only one parent (the control) were reviewed. Table 3.19 summarizes the results. Within the sample, concepts with more parents were found to have *fewer* errors than concepts with one parent, thus, this approach does not appear practical for the *Observable entity* hierarchy.

Table 3.19 Auditing Results for Number of Parents Study

	# of Erroneous Concepts (%)	**# of Sample Concepts**
1 Parent	4 (4%)	100
> 1 Parents	1 (0.5%)	200

Source: [27]

3.2.2.1.3 Number of Words in Fully Specified Name. The third and final method is based on the hypothesis that concepts with relatively long fully specified names are more complex, and thus more likely to have errors, than concepts with shorter fully specified names, which was also confirmed for the Problem List of SNOMED CT by Agrawal et al. [128]. The Observable entity hierarchy has concepts with fully specified names of net word length three to 16 (including semantic tag). A randomly selected sample of 100 concepts was selected from concepts with a short net word length of three to five.

Another sample of 200 concepts was randomly selected from concepts with a long word length, 9 to 16 words. Table 3.20 summarizes the results. While concepts with longer word lengths are slightly more likely to be erroneous (7.5%) than concepts with shorter word lengths (6%) there was no statistically significant difference.

Table 3.20 Auditing Results for Fully Specified Name Length Study

	# of Erroneous Concepts (%)	**# of Sample Concepts**
3-5 Words	6 (6%)	100
9-16 Words	15 (7.5%)	200

Source: [27]

3.2.2.1.4 Comparative Study Discussion. Lexical containment was found to successfully identify groups of concepts that are more likely to have errors than a control sample with a higher error ratio than the TAN-based QA technique. Larger numbers of parents and longer fully specified names were found to *not* indicate a higher likelihood of error in *Observable entity*. The comparative study showed that both QA techniques can successfully be applied to the *Observable entity* hierarchy. By combining the TAN and lexical containment techniques to identify different sets of concepts, each promising a higher percentage of errors, a relatively large number of errors can be found with a relatively small QA effort.

Both the TAN and lexical containment techniques have advantages and disadvantages. The TAN, for example, provides context for the concepts being reviewed. Finding one error may lead to uncovering similar errors for other concepts in same cluster. With the lexical containment approach, an auditor is only presented with pairs of concepts. More complex (or similar) modeling errors may not be uncovered. However, the TAN approach of reviewing complex concepts had a lower error rate (7.8%) than the lexical containment approach (20%). By reviewing potentially missing ancestor-successor relationships, an auditor's work load is significantly reduced when compared to manually reviewing all of the relationships of a given concept.

However, one major weakness of lexical containment is that it can only be applied when there are lexical containment concept pairs without ancestor-successor relationships; it cannot be used to uncover missing IS-A relationships when the parent is not lexically contained in the child, e.g., *Behavior to maintain weight* and *Weight control behavior*. A total of 21 errors uncovered using the TAN review would not have been found using lexical containment.

In regards to implementation cost, both the TAN and lexical containment techniques require an initial effort to implement the algorithms that generate clusters and lexical containment pairs, respectively. However, the biggest cost for both methods is the manual auditing effort required; in the lexical containment approach the auditor only reviews the given pair of concepts, while in the TAN-based methodology the auditor reviews the entire neighborhood of a concept, potentially finding more kinds of errors.

Table 3.21 Precision, Sensitivity, and Specificity for the TAN and Lexical Containment Methodologies

	Tribal Abstraction Network	**Lexical Containment**
Error Ratio	3.71 (=0.078/0.021)	6.67 (=0.2/0.03)
Precision	0.078 (=20/(20 + 237))	0.2 (=40/(40 + 160))
Sensitivity	0.51 (=20/(20 + 19))	0.93 (=40/(40 + 3))
Specificity	0.79 (=884/(884 + 237))	0.38 (=97/(97 + 160))

Source: [27]

In Table 3.21 the precision, sensitivity, and specificity of the TAN and lexical containment methodologies are compared Since number of parents and number of words were found to be not helpful for QA of the *Observable entity* hierarchy, their measures are not reported. For the TAN methodology a positive result is a complex concept belonging to Bin 1 or Bin 2 in Level 2 or Level 3 of the *Observable entity* TAN (257 total), as described in Discussion. A negative result is any other concept (903 total). A true positive is an error in the positive group (20 total) and a false positive is a positive concept that has no error (237 total). A true negative is a negative concept with no errors (884 total) and a false negative is a negative concept with an error (19 total).

For lexical containment a true result is a lexical pair without an ancestor-successor relationship (200 concepts total). A negative result is a concept not in a lexical pair (100 total). A true positive is a lexical pair that should have an ancestor-successor relationship (40 total) and a false positive is a lexical pair where no ancestor-successor relationship should exist (160 total). A true negative is a non-erroneous concept not in a lexical containment pair (97 total) and a false negative is an erroneous concept that is not in a lexical containment pair (3 total).

3.2.3 Limitations

The TAN represents a new paradigm for summarizing SNOMED CT hierarchies. There are several open research questions regarding their use for summarization and quality assurance. One issue is the relatively low number of errors uncovered in the quality assurance review of the *Observable entity* hierarchy. There are several possible reasons for this finding. One possibility is the *Observable entity* hierarchy is more correctly modeled when compared to *Specimen* and *Procedure*, which had much higher error rates [26, 28, 29]. Alternatively, low error rates may be a common phenomenon in all target hierarchies. Another important issue is the emergence of disproportionately large clusters ("super-large clusters," for short) which summarize thousands, or tens of thousands, of concepts. These clusters represent an over summarization of a set of concepts. Future studies, discussed in Section 4.2, will be conducted to investigate these limitations.

3.2.4 Additional TAN Applications

The TAN can be used to address several open issues with partial-area taxonomy-based quality assurance methodologies. Large SNOMED CT hierarchies with attribute relationships, such as *Procedure, Clinical finding,* and *Body structure*, have very large root partial-areas (the single partial-area in each hierarchy's ∅ area). This root partial-area contains only concepts that have no relationships. The root partial-areas of *Procedure* and *Clinical finding* contain over 2,500 concepts and over 8,000 concepts, respectively. In the *Body structure* partial-area taxonomy, the root partial-area contains nearly 28,000 concepts, 90% of the hierarchy. Super-large root partial-areas represent an over summarization of the concepts in these hierarchies.

Partial-areas are singly rooted and root partial-areas in partial-area taxonomies derived for entire hierarchies contain the portion of the hierarchy which has no attribute relationships. Super-large root partial-areas may include concepts that have multiple parent concepts that are also in the same super-large partial-area. Therefore, if such concepts exist, it may be possible to derive a TAN for the concepts in a super-large root partial-area. The children of the root partial-area's root concept can be used as patriarchs. The TAN derivation methodology can then be applied using the set of concepts in the root partial-area. Thus, a root partial-area's TAN will summarize the hierarchy of concepts in a super-large partial-area. This approach enables TAN-based quality assurance of such concepts.

It is actually possible to derive a TAN for any partial-area in a partial-area taxonomy, not just super-large root partial-areas. What is common to all concepts in a partial-area is that they all share the same root concept and the same set of attribute relationships. Hence, for non-root partial-areas, it is not possible to obtain further division when using a taxonomy for a complete hierarchy. However, by ignoring the lateral relationships of the concepts in such a super-large partial-area, it is possible to derive a TAN for a non-root partial-area, enabling summarization of its concepts.

This will be particularly useful for super-large partial-areas in the *Procedure* hierarchy, e.g., *Procedure by method* (3684), *Imaging by body site* (1673), and *Measurement of substance* (3980). A study will be conducted to investigate the use of TANs to complement existing partial-area taxonomy-based quality assurance of the *Procedure* or *Clinical finding* hierarchy. TANs will be created for several non-root partial-areas and a quality assurance review of their concepts will be performed.

For hierarchies with lateral relationships, such as *Specimen*, *Procedure,* and *Clinical finding*, it is possible to derive either a partial-area taxonomy or a TAN. When compared to a partial-area taxonomy for the same hierarchy, a TAN provide an alternate, hierarchy-focused summarization of the same content. It is possible to compare the TAN summary against a partial-area taxonomy summary for such a hierarchy. A study will be performed to investigate how a TAN summary can support quality assurance for source hierarchies, complementing the existing quality assurance techniques which utilize taxonomies.

3.2.5 Discussion

The TAN addresses the need for summarization methodologies for the eight target hierarchies of SNOMED CT with multiple parents. The number of concepts with multiple parents in a hierarchy is not as important for deriving a TAN as the locations where such concepts appear. Only 412 (5.33%) of the concepts in *Observable entity* have multiple parents, a relatively small number compared to several other hierarchies (Table 3.11), but a TAN is successfully derived, since 153 such concepts are located “at the crossroads” of tribe combinations.

The TAN summary of a target hierarchy can be used to support quality assurance. The overall desired effect of using a TAN is to limit the resources for and increase the yield of QA. It was found that concepts in the *Observable entity* hierarchy are more likely to be erroneous if they belong to larger clusters (e.g., Bins 1, 2) in the TAN rather than to smaller clusters (Bins 3-6). Furthermore, the percentage of errors was highest in larger clusters at Level 3 and slightly higher in larger clusters in Level 2 than Level 1.

Following the previously described TAN quality assurance methodology, the 86 and 526 concepts in large clusters of Levels 3 and 2, respectively, should be reviewed. The 86 concepts in the larger Level 3 cluster were reviewed and 11 erroneous concepts were found. The number of erroneous concepts expected in reviewing the 526 concepts in larger Level 2 clusters is 28 (=0.0526×526) (based on E in Table 3.14). Hence, a total of 39 (=11+28) errors are expected from reviewing 612 (=86+526) concepts in the large clusters of Levels 2 and 3, according to the methodology. Coincidentally, 39 erroneous concepts were also found when reviewing the sample of 1,160 concepts. Hence, the methodology would likely yield the same number of erroneous concepts while saving the review of 548 (=1,160 [=total reviewed] – 612 [=86+526]) extra concepts.

One issue which will be investigated is the existence of concepts which overlap between multiple clusters. While no such concepts currently exist in the *Observable entity* hierarchy, there are over 18,000 concepts that overlap between multiple clusters spread throughout SNOMED CT's other hierarchies. For partial-area taxonomies, concepts that overlap between multiple partial-areas have been found to be more likely to contain errors [29]. It is hypothesized that a similar hypothesis will be true for concepts that overlap between clusters.

3.3 Abstraction Networks for OWL Ontologies

Abstraction network derivation methodologies have been created for various ontologies. However, idiosyncrasies in the underlying knowledge models limit the overall applicability of the methodologies outside of a few specific ontologies. Many abstraction networks, such as the area taxonomies and partial-area taxonomies for SNOMED CT [24] and NCIt [22], are only applicable to their associated ontology. There is a need to formulate a unified abstraction methodology that can be applied to entire families of structurally similar ontologies.

One of the major goals of this research is to develop an abstraction-network-based summarization and quality assurance framework for the ontologies in the NCBO BioPortal [50]. BioPortal does not currently provide functionality to support the summarization and quality assurance of ontologies. A preliminary step in this research was the investigation of a family-based quality assurance framework. In He et al. [55] the structure of 186 BioPortal ontologies was analyzed to determine which *structural features* were available for abstraction network derivation.

A *structural feature* is defined as a type of knowledge element or structural configuration used within an ontology. Structural features are used to define an ontology’s classes. In SNOMED CT, lateral attribute relationships are a type of structural feature used to define many concepts. Object properties and data properties can be considered structural features of OWL ontologies. Different hierarchical relationship configurations also constitute a type of structural feature. For example, the existence of classes with multiple superclasses (multiple parents) can also be considered a structural feature of an ontology. Multiple parents can be used to derive different types of abstraction networks (e.g., the TAN in Section 3.2 or the disjoint partial-area taxonomy in [23]).

Many biomedical ontologies are developed in OWL and Open Biological and Biomedical Ontologies (OBO) [52] formats. OWL and OBO provide standard frameworks for creating ontologies. Note that OBO ontologies can be converted to OWL, so the methodologies described in this section are also applicable to OBO ontologies [129]. Of the over 300 ontologies in BioPortal, the large majority are released in either OWL or OBO format.

One example of an ontology developed in OWL, and available in BioPortal, is the Ontology of Clinical Research (OCRe), which provides classes and relationships to characterize the different types of human studies in a uniform way [51]. OCRe was developed using Protégé [56] and focuses on annotating human studies according to their design and analysis. OCRe is organized as a set of modular components with the core modules being *Study protocol*, *Study design*, and *Statistics*. OCRe includes significant information on human investigations, with its Revision 258 consisting of 342 unique classes and 192 different kinds of relationships [130].

Another OWL ontology available in BioPortal is the Sleep Domain Ontology (SDO) [45], an ontology focused on the domain of sleep medicine. The SDO consists of 1390 classes and is available on BioPortal in OWL format. The SDO was developed as part of the PhysioMIMI project to support the merging of physiological and clinical data. The SDO was built by merging knowledge from several ontologies, such as the Ontology for General Medical Science (OGMS) [47] and Foundational Model of Anatomy (FMA) [131], with sleep-domain knowledge being added by its curator.

OCRe and SDO will be used to illustrate the process deriving two types of abstraction networks which are designed for OWL ontologies: the *domain-defined partial-area taxonomy* and the *restriction-defined partial-area taxonomy*.

3.3.1 Domain-defined Derivation Methodology

An important structural feature used in the development of OWL ontologies is the *object property*, which defines a directed binary relationship between two classes, allowing for their respective instances to be related. Using an example from OCRe, consider the definition of the object property *hasMember*: The example below, shown in OWL XML format, states that *hasMember* has the domain (class) *Organization* and the range (class) *Person*. This indicates that, within the OCRe ontology, any instance of an *Organization* can have a member that is an instance of *Person*. Object properties can have more than one class in their domain or range.

```
<owl:ObjectProperty rdf:ID="hasMember">
     <rdfs:domain rdf:resource="Organization"/>
     <rdfs:range rdf:resource="Person"/>
</owl:ObjectProperty>
```

The derivation of a domain-defined partial-area taxonomy for OWL-based ontologies required the reformulation of the area and partial-area taxonomic elements. For OWL ontologies, these notions are based on object properties and their explicitly defined domains (such as *Organization* in the *hasMember* example above). This represents a shift from a reliance on instantiated relationship occurrences (e.g., in SNOMED CT) to potential relationship occurrences.

Let *O* be a non-empty set of object properties. The area with respect to *O* is defined as the set of all classes that are explicitly defined (or are inferred) as being in the exact domains of *O*'s object properties. The object properties collectively are used to name the area. For example, the OCRe class *Entity* is explicitly asserted as the domain for the object properties *has part* and *part of*; therefore, it belongs to the area named {*has part, part of*}. All of the descendants of *Entity* are also implicitly within the domain of *has part* and *part of*. However, many descendants "introduce" new object properties of their own in the sense of being the asserted domain of the properties and therefore, will have different (larger) sets of object properties and reside in different areas. This inheritance and introduction of object properties within OWL ontologies (such as OCRe) is the basis for defining an area taxonomy. Areas are linked by *child-of* relationships that abstract the underlying subclass hierarchy.

A root within an area is defined as a class such that the set of object properties having the class as their domains differs from all such sets of its superclasses. An area may have more than one root. A partial-area, which is based on a root within an area, is defined as a subhierarchy of classes that share a common set of object properties and a common ancestor class, namely, the root, which introduced the partial-area's new object properties (while the rest of the object properties were inherited from ancestors of the root). Let *R* be a root of an area *A*. The set of classes consisting of *R* and all its descendants in *A* is called a partial-area and is named after the root. Partial-areas are linked by *child-of* relationships derived from the underlying subclass hierarchy in the ontology.

To illustrate the process of deriving a domain-defined partial-area taxonomy, OCRe's *Entity* hierarchy, which is the largest in the ontology and features a rich set of object properties, was utilized in Ochs et al. [25]. As of Version 244 of OCRe, there were 120 distinct classes and 75 unique types of object properties whose explicitly defined domains are subclasses of *Entity*. This hierarchy also contains the important *Study* class, which is considered the primary element of OCRe.

A preliminary step in creating a domain-defined partial-area taxonomy is to run a reasoner on the ontology to obtain the inferred view. Pellet [68], provided within Protégé [56], was applied to the stated view of OCRe. As the first step, object property introduction was analyzed within the hierarchy. Figure 3.20 shows an indented subhierarchy of 21 classes from *Entity*, along with classes that are explicitly defined as the domains for the given object properties. The object properties introduced at a given class are shown in brackets next to the class name. Background color alternates between white and light blue to help identify on which level of the subhierarchy a given class resides.

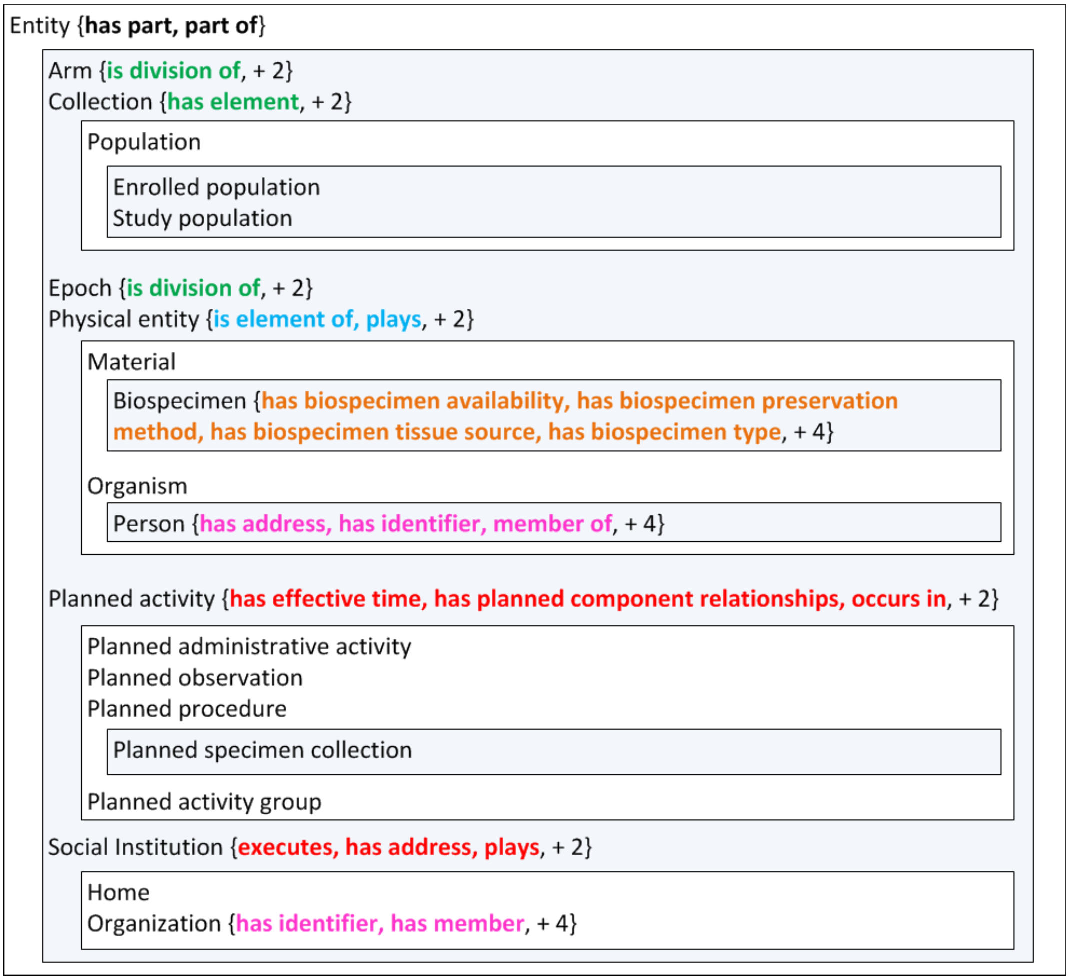

Entity {has part, part of}
Arm {is division of, + 2}
Collection {has element, + 2}
Population
Enrolled population
Study population
Epoch {is division of, + 2}
Physical entity {is element of, plays, + 2}
Material
Biospecimen {has biospecimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type, + 4}
Organism
Person {has address, has identifier, member of, + 4}
Planned activity {has effective time, has planned component relationships, occurs in, + 2}
Planned administrative activity
Planned observation
Planned procedure
Planned specimen collection
Planned activity group
Social Institution {executes, has address, plays, + 2}
Home
Organization {has identifier, has member, + 4}

Figure 3.20 A portion of OCRe's *Entity* hierarchy in an indented format with object properties introduced. The number after "+" indicates the number of inherited properties.
Source: [25]

As an example, the class Physical entity is defined as being within the domain of two object properties, is element of and plays. In addition, Physical entity has the two object properties has part and part of that are inherited from the Entity class.

Altogether, *Physical entity* is in the domain of four object properties. Object properties are color-coded according to the total number of properties for the class at which they are introduced. For example, all classes with a green-colored object property have three in total. This color coding will be utilized in the following figures.

Once the inferred hierarchy of an ontology is established, and all its object properties are identified, the class hierarchy of the ontology is traversed using a topological traversal algorithm [124] and the classes that are in the domains of the exact same set of object properties are grouped together.

Applying this second step to OCRe, starting at the class Entity, established the area taxonomy for Entity. Figure 3.21 shows the grouping of classes for the sample of OCRe's hierarchy shown in Figure 3.20. Areas are represented as colored boxes. Different colors indicate different numbers of object properties. Sets of object properties are shown at the top of each colored box. Each such list of object properties is the respective area's name. Classes with that set of properties are shown in the box, with descendants of each root shown indented. For example, the class Collection in Figure 3.21 has the object property set {has part, part of, has element} and a child class Population along with two grandchildren, Enrolled population and Study population. Edges are used to represent *child-of* links between areas. *Child-of* links indicate the chain of inheritance involved with a particular set of object properties. The edge from the area containing Organization indicates that this area inherited five properties from the area containing Social institution.

The third and final step in the domain-defined derivation methodology is to identify the roots of each area and group the descendants of these roots into partial-areas. In Figure 3.21, *Planned activity* is the ancestor of five classes that all share the same set of object properties: *has part*, *part of*, *has effective time*, *has planned component relationship*, and *occurs in*. This group of classes is joined into a single partial-area rooted at *Planned activity*.

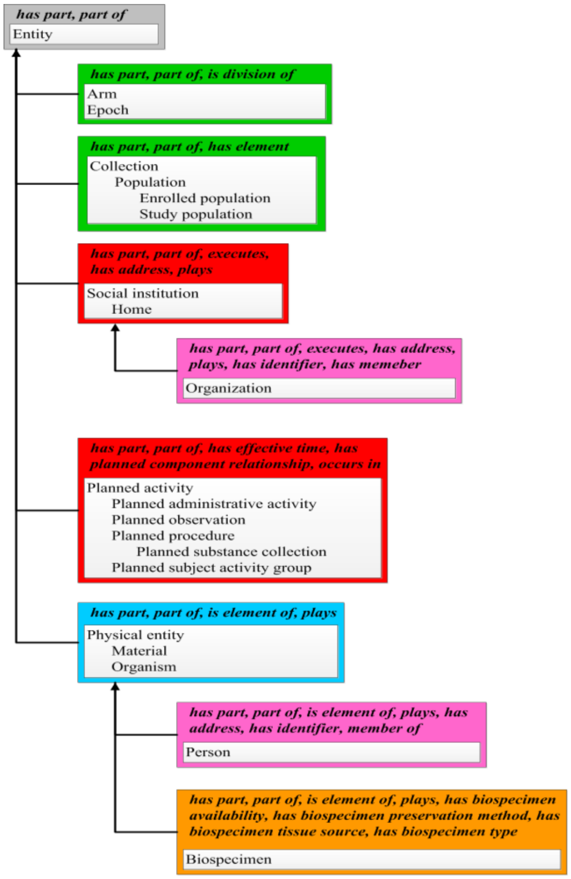

has part, part of
Entity
has part, part of, is division of
Arm
Epoch
has part, part of, has element
Collection
Population
Enrolled population
Study population
has part, part of, executes, has address, plays
Social institution
Home
has part, part of, executes, has address, plays, has identifier, has memeber
Organization
has part, part of, has effective time, has planned component relationship, occurs in
Planned activity
Planned administrative activity
Planned observation
Planned procedure
Planned substance collection
Planned subject activity group
has part, part of, is element of, plays
Physical entity
Material
Organism
has part, part of, is element of, plays, has address, has identifier, member of
Person
has part, part of, is element of, plays, has biospecimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type
Biospecimen

Figure 3.21 The grouping of classes from Figure 3.20 into areas based on each class' set of object properties. Edges are *child-of* links between areas.
Source: [25]

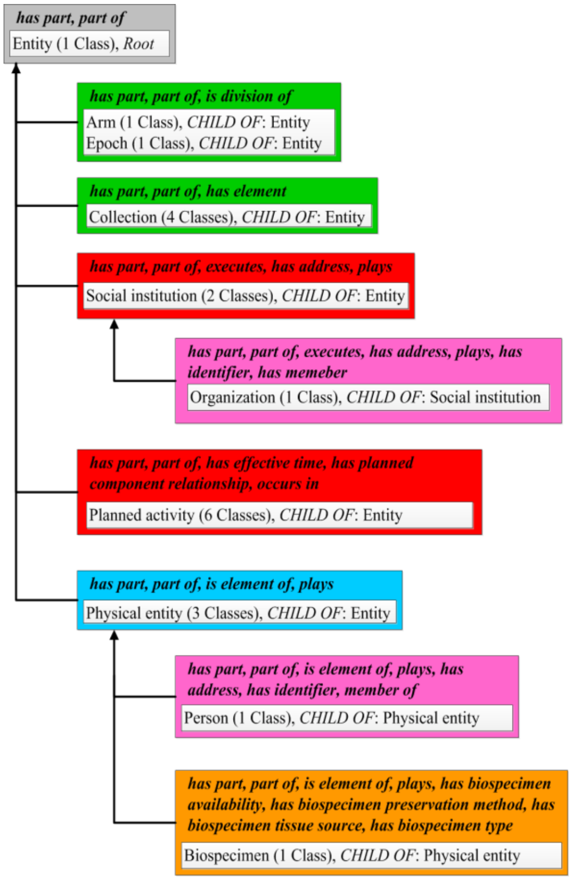

has part, part of
Entity (1 Class), Root
has part, part of, is division of
Arm (1 Class), CHILD OF: Entity
Epoch (1 Class), CHILD OF: Entity
has part, part of, has element
Collection (4 Classes), CHILD OF: Entity
has part, part of, executes, has address, plays
Social institution (2 Classes), CHILD OF: Entity
has part, part of, executes, has address, plays, has identifier, has memeber
Organization (1 Class), CHILD OF: Social institution
has part, part of, has effective time, has planned component relationship, occurs in
Planned activity (6 Classes), CHILD OF: Entity
has part, part of, is element of, plays
Physical entity (3 Classes), CHILD OF: Entity
has part, part of, is element of, plays, has address, has identifier, member of
Person (1 Class), CHILD OF: Physical entity
has part, part of, is element of, plays, has biospecimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type
Biospecimen (1 Class), CHILD OF: Physical entity

Figure 3.22 Partial-areas derived for each area in Figure 3.21. The targets of *child-of* links within partial-areas are indicated after "*CHILD OF*."
Source: [25]

Figure 3.22 shows the partial-area taxonomy for the subset of classes from Figure 3.21. In Figure 3.22, the root of each partial-area is explicitly identified. The numbers in parentheses indicate the total numbers of classes summarized by the respective partial-areas. The targets of the *child-of*'s between partial-areas are listed in the boxes after "*CHILD OF*." Figure 3.23 shows the final diagram representation of the domain-defined partial-area taxonomy created for OCRe's *Entity* subhierarchy given in Figure 3.20. It is a representation that is more compact than the original hierarchy.

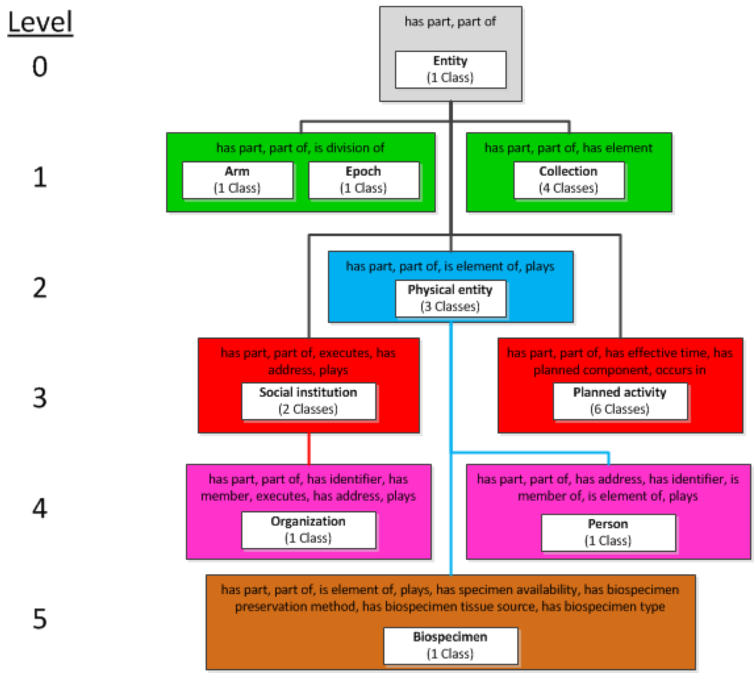

Level
0
1
2
3
4
5
has part, part of
Entity
(1 Class)
has part, part of, is division of
Arm
(1 Class)
Epoch
(1 Class)
has part, part of, has element
Collection
(4 Classes)
has part, part of, is element of, plays
Physical entity
(3 Classes)
has part, part of, executes, has address, plays
Social institution
(2 Classes)
has part, part of, has effective time, has planned component, occurs in
Planned activity
(6 Classes)
has part, part of, has identifier, has member, executes, has address, plays
Organization
(1 Class)
has part, part of, has address, has identifier, is member of, is element of, plays
Person
(1 Class)
has part, part of, is element of, plays, has specimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type
Biospecimen
(1 Class)

Figure 3.23 Partial-area taxonomy for the subset of classes from OCRe's *Entity* hierarchy in Figure 3.20.
Source: [25]

Figure 3.23 condenses the 21 classes of Figure 3.20 into a structure of ten partial-areas residing in nine areas. In the figure, the colored boxes represent areas. The white boxes in an area are the partial-areas. The number of classes for each partial-area is shown. Areas are organized into levels based on the number of object properties in each area. Areas containing classes with the fewest object properties are at the top.

Within the partial-area taxonomy, edges are used to represent the *child-of*'s links between partial-areas. As a graphical simplification, edges are drawn between areas only when all of the partial-areas of a given area are children of the same parent area. As the level structure is now explicit, arrow heads may be omitted. For additional clarity, edges may be color coded based on which area the parent partial-area resides in. For example, *Biospecimen* is *child-of Physical entity*. Since *Physical entity* is at Level 2 (blue), a blue line is used to connect the two areas that contain these two partial-areas.

The advantage of the domain-defined partial-area taxonomy derivation methodology is that it is applicable to any ontology expressed in OWL which has object properties with explicitly defined domains. This is in contrast to the previously developed methodologies which were only applicable to one ontology at a time (e.g., SNOMED CT). For example, in [55] a domain-defined partial-area taxonomy was created for the Cancer Chemoprevention Ontology (CanCo) [64]. Additionally, Ochs and Perl [132] illustrate several other examples of domain-defined partial-area taxonomies created for ontologies such as the Sleep Domain Ontology (SDO) [45] and Top-Meneles [133]. Based on the analysis in [55], the domain-defined partial-area taxonomies derivation methodology is applicable to over 80 different BioPortal ontologies.

3.3.2 Restriction-defined Derivation Methodology

Many OWL ontologies, such as the Sleep Domain Ontology (SDO), do not rigorously define domains and ranges for every object property. In such cases, the domain-defined partial-area taxonomy derivation methodology will not produce a useful summarization of the ontology. Thus, alternate structural features must be considered to derive partial-area taxonomies for these ontologies. OWL allows object properties to be used in restrictions on classes. The major difference from explicitly specifying domains and ranges is that a restriction is *local*, i.e., the restriction only applies within the context of the class with the restriction (and, implicitly, its descendants). For example, consider the following class definition from the SDO, shown in Manchester OWL syntax:

```
Class: BilateralUpperLimbMovementDuringSleep
   SubClassOf:
      UpperLimbMovementDuringSleep
      includes some
            RightUpperLimbMovementDuringSleep
      includes some
            LeftUpperLimbMovementDuringSleep
```

This states that the class *Bilateral Upper Limb Movement During Sleep* is a subclass of two restrictions that use the object property *includes*. One restriction is that *Bilateral Upper Limb Movement During Sleep* includes *Right Upper Limb Movement During Sleep*; the second is that it includes *Left Upper Limb Movement During Sleep*. Both restrictions use the constraint *some*, which requires that at least one instance of the object property used with *Bilateral Upper Limb Movement During Sleep* conform to the restriction. An alternative would be *all*, which means when the object property *includes* is used with *Bilateral Upper Limb Movement During Sleep,* all its instances must conform to the restriction. Using object properties in restrictions allows for more flexibility in ontology design. *Includes* is a high-level property used in 82 different restrictions in SDO.

Taxonomies can be derived using the defined restrictions when there are a sufficient number of them, yielding what is called a *restriction-defined partial-area taxonomy*. The derivation of the restriction-defined partial-area taxonomy was described in Ochs et al. [120]. The SDO has 44 types of object properties used in restrictions on classes, which means that there are enough of them for creating a partial-area taxonomy for the ontology.

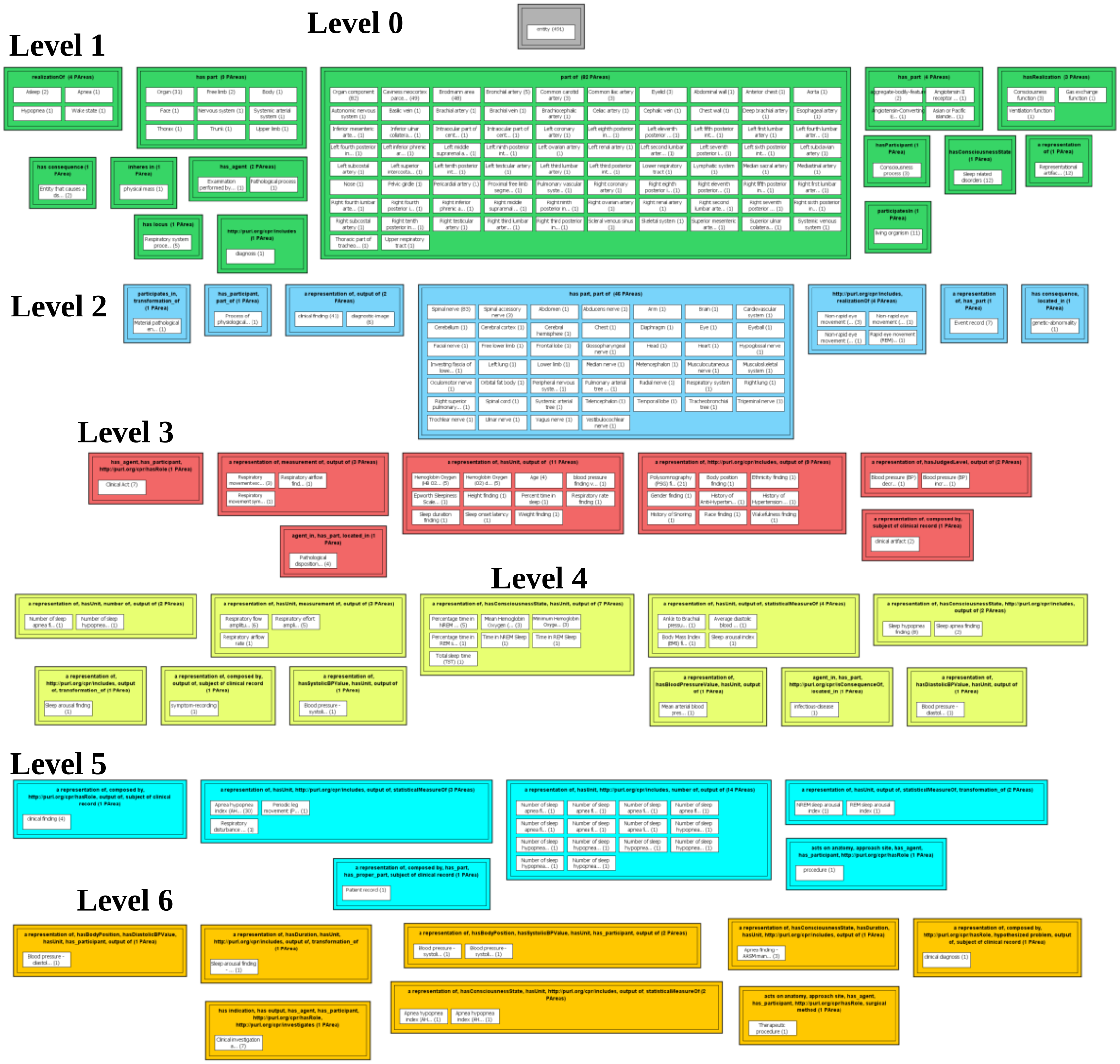


Figure 3.24 The restriction-defined partial-area taxonomy for the Sleep Domain Ontology's *Entity* hierarchy. Levels have been organized into rows and child-of edges are hidden for readability.

In a restriction-defined taxonomy, an area is defined to be the set of classes that are explicitly defined or inferred to be bound by restrictions that use the object properties in a given set *O*. A restriction can be either *allValuesFrom* or *someValuesFrom*; the methodology does not distinguish between the two. *Child-of* links are derived as with the domain-defined partial-area taxonomy. The class that has the restriction is treated as belonging to the domain of the object property.

Additionally, any descendants of the class with the restriction are considered to be implicitly in the object property's domain. The definition of the partial-areas remains unchanged from the domain-defined derivation methodology.

Figure 3.24 shows the first seven levels of the SOD's restriction-defined partial-area taxonomy. Levels 7 and 8 are not shown. The complete restriction-defined taxonomy contains 262 partial-areas in 61 areas. The full restriction-defined taxonomy can be viewed in [132].

Like the domain-defined taxonomy, the restriction-defined taxonomy is applicable to many OWL ontologies, particularly those which mostly use object properties in restrictions on classes. He et al. [55] found that 150 out of 186 BioPortal ontologies fit this criteria. Thus, the restriction-defined partial-area taxonomy can be used to summarize many of these ontologies.

3.3.3 Granularity of OWL Abstraction Networks

The *abstraction ratio* of an abstraction network is defined as the average number of ontology classes mapped to each abstraction network node (i.e. #classes / #nodes). This ratio indicates the *granularity* of an abstraction network. If there are few nodes in the abstraction network (e.g., many classes are mapped to few nodes), then the abstraction network has *coarse* granularity. Even though the abstraction network summarizes the ontology, that summary may contain too little information to be considered useful. Conversely, an abstraction network's granularity may be considered *too fine* if it has too many nodes, meaning the summarization benefits are effectively lost.

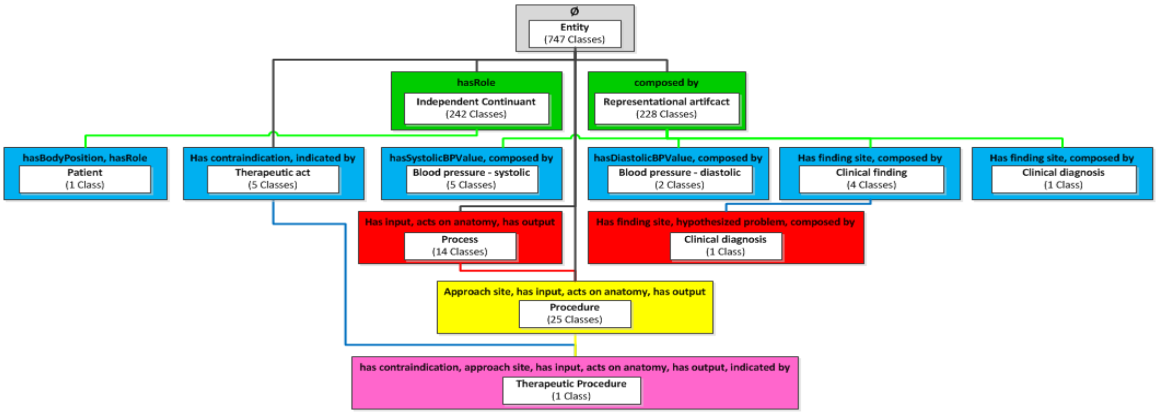

Ø
Entity
(747 Classes)
hasRole
Independent Continuant
(242 Classes)
composed by
Representational artifcact
(228 Classes)
hasBodyPosition, hasRole
Patient
(1 Class)
Has contraindication, indicated by
Therapeutic act
(5 Classes)
hasSystolicBPValue, composed by
Blood pressure - systolic
(5 Classes)
hasDiastolicBPValue, composed by
Blood pressure - diastolic
(2 Classes)
Has finding site, composed by
Clinical finding
(4 Classes)
Has finding site, composed by
Clinical diagnosis
(1 Class)
Has input, acts on anatomy, has output
Process
(14 Classes)
Has finding site, hypothesized problem, composed by
Clinical diagnosis
(1 Class)
Approach site, has input, acts on anatomy, has output
Procedure
(25 Classes)
has contraindication, approach site, has input, acts on anatomy, has output, indicated by
Therapeutic Procedure
(1 Class)

Figure 3.25 Domain-defined partial-area taxonomy for the SDO's *Entity* hierarchy.
Source: [120]

Granularity may be affected by an abstraction network's derivation methodology. Several different types of abstraction networks can potentially be derived for the same ontology. What differs among the abstraction networks is the algorithm used to define the nodes (i.e., what structural features are utilized to create the abstraction network).

The SDO has object properties with explicitly defined domains and object properties used in restrictions. Thus, both the domain-defined derivation methodology and the restriction-defined derivation methodology are applicable to the SDO. Granularity differences are expected for different types of abstraction networks. Finding the "best" abstraction network for an ontology is based on the structure of the ontology and/or the intended use of the abstraction network.

Figure 3.25 shows a domain-defined partial-area taxonomy for the SDO's *Entity* hierarchy. When compared to the SDO's restriction-defined taxonomy in Figure 3.24, there is significantly less information conveyed in the domain-defined taxonomy. The SDO's domain-defined partial-area taxonomy is considered too coarse in granularity for activities such as quality assurance. The domain-defined taxonomy contains only 13 partial-areas separated into an equal number of areas. The abstraction ratio is 98.08

(=1,275/13) classes per partial-area. This is in contrast to the 262 partial-areas in 61 areas for the restriction-defined taxonomy (abstraction ratio of 4.867). Three partial-areas, *Entity*, *Representational artifact*, and *Independent continuant*, together constitute nearly the entire hierarchy (1,217 classes). The ten other partial-areas together contain only 58 classes, 25 of which are in the partial-area *Procedure*. Hence, the granularity of the top part of the taxonomy is too coarse for quality assurance, since this portion of the taxonomy over-summarizes the content.

Domain-defined taxonomies will only provide sufficient granularity when enough object properties have explicitly defined domains and the set of classes that are in one or more object property's domains is large enough. Within the SDO's *Entity* hierarchy only 16 of the 50 object properties have explicitly defined domains. The remaining object properties are used in restrictions or have no domain information. When no domain information is given, the domain is implicitly the root of the ontology, and is not used in the derivation of a taxonomy.

When no single derivation methodology provides an abstraction network of sufficient granularity, combinations of derivation methodologies can be used to derive new kinds of abstraction networks. For example, if neither the domain-defined partial-area taxonomy or restriction-defined partial-area taxonomy derivation methodologies work well on their own, then they can be combined to create a (domain or restriction)-defined derivation methodology, which uses object properties with either explicitly defined domains or object properties used in class restrictions to create a partial-area taxonomy.

\

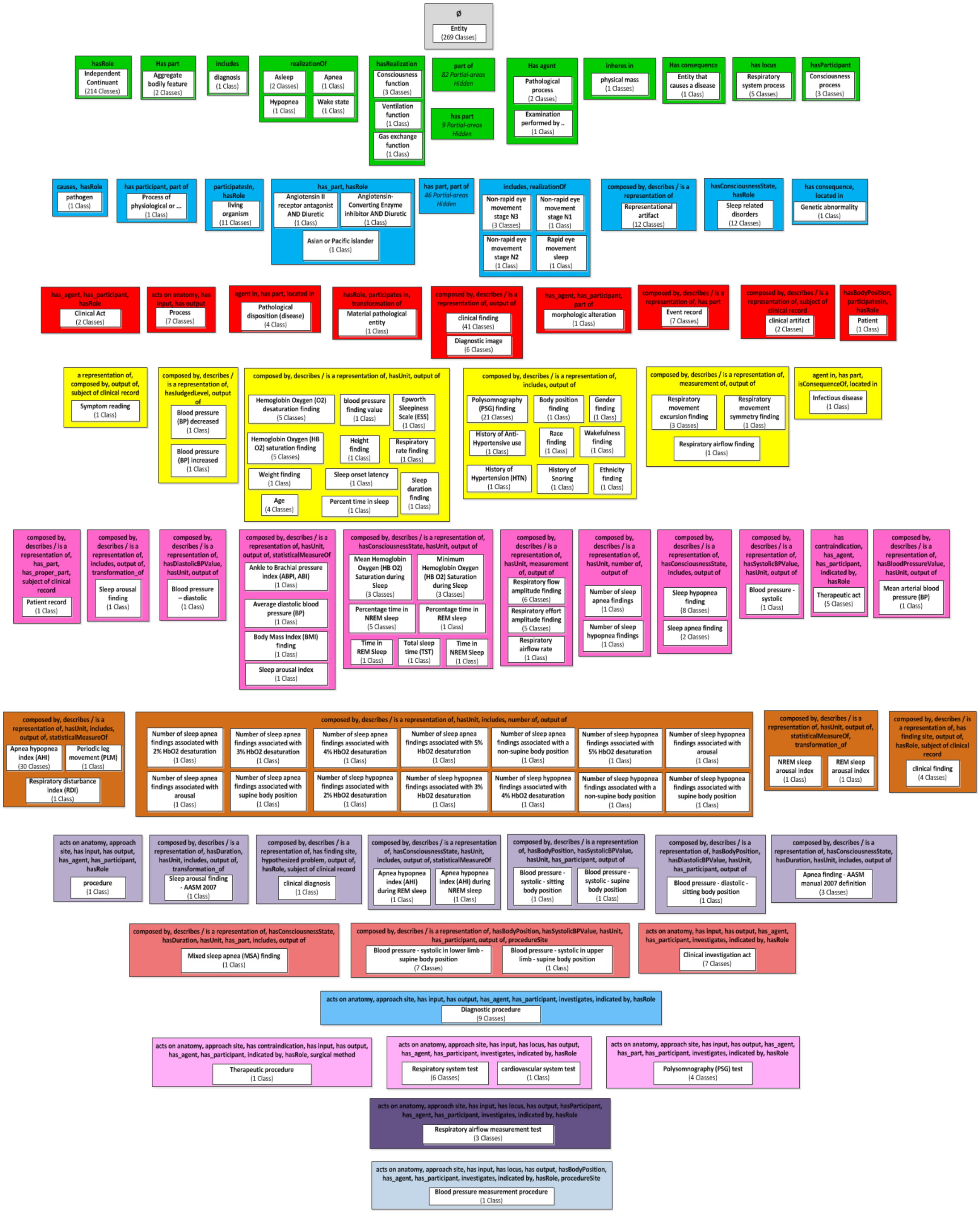

ø
Entity (269 Classes)
hasRole
Independent Continuant (214 Classes)
Has part
Aggregate bodily feature (2 Classes)
includes
diagnosis (1 Class)
realizationOf
Asleep (2 Classes)
Apnea (1 Class)
Hypopnea (1 Class)
Wake state (1 Class)
hasRealization
Consciousness function (3 Classes)
Ventilation function (1 Class)
Gas exchange function (1 Class)
part of
82 Partial-areas Hidden
has part
9 Partial-areas Hidden
Has agent
Pathological process (2 Classes)
Examination performed by .. (1 Class)
inheres in
physical mass (1 Classes)
Has consequence
Entity that causes a disease (1 Class)
has locus
Respiratory system process (5 Classes)
hasParticipant
Consciousness process (3 Classes)
causes, hasRole
pathogen (1 Class)
has participant, part of
Process of physiological or ... (1 Class)
participatesIn, hasRole
living organism (11 Classes)
has_part, hasRole
Angiotensin II receptor antagonist AND Diuretic (1 Class)
Angiotensin-Converting Enzyme inhibitor AND Diuretic (1 Class)
Asian or Pacific islander (1 Class)
has part, part of
46 Partial-areas Hidden
includes, realizationOf
Non-rapid eye movement stage N3 (3 Classes)
Non-rapid eye movement stage N1 (1 Class)
Non-rapid eye movement stage N2 (1 Class)
Rapid eye movement sleep (1 Class)
composed by, describes / is a representation of
Representational artifact (12 Classes)
hasConsciousnessState, hasRole
Sleep related disorders (12 Classes)
has consequence, located in
Genetic abnormality (1 Class)
has_agent, has_participant, hasRole
Clinical Act (2 Classes)
acts on anatomy, has input, has output
Process (7 Classes)
agent in, has part, located in
Pathological disposition (disease) (4 Class)
hasRole, participates in, transformation of
Material pathological entity (1 Class)
composed by, describes / is a representation of, output of
clinical finding (41 Classes)
Diagnostic image (6 Classes)
has_agent, has_participant, part of
morphologic alteration (1 Class)
composed by, describes / is a representation of, has part
Event record (7 Classes)
composed by, describes / is a representation of, subject of clinical record
clinical artifact (2 Classes)
hasBodyPosition, participatesIn, hasRole
Patient (1 Class)
a representation of, composed by, output of, subject of clinical record
Symptom reading (1 Class)
composed by, describes / is a representation of, hasJudgedLevel, output of
Blood pressure (BP) decreased (1 Class)
Blood pressure (BP) increased (1 Class)
composed by, describes / is a representation of, hasUnit, output of
Hemoglobin Oxygen (O2) desaturation finding (5 Classes)
blood pressure finding value (1 Class)
Epworth Sleepiness Scale (ESS) (1 Class)
Hemoglobin Oxygen (HB O2) saturation finding (5 Classes)
Height finding (1 Class)
Respiratory rate finding (1 Class)
Weight finding (1 Class)
Sleep onset latency (1 Class)
Sleep duration finding (1 Class)
Age (4 Classes)
Percent time in sleep (1 Class)
composed by, describes / is a representation of, includes, output of
Polysomnography (PSG) finding (21 Classes)
Body position finding (1 Class)
Gender finding (1 Class)
History of Anti-Hypertensive use (1 Class)
Race finding (1 Class)
Wakefulness finding (1 Class)
History of Hypertension (HTN) (1 Class)
History of Snoring (1 Class)
Ethnicity finding (1 Class)
composed by, describes / is a representation of, measurement of, output of
Respiratory movement excursion finding (3 Classes)
Respiratory movement symmetry finding (1 Class)
Respiratory airflow finding (1 Class)
agent in, has part, isConsequenceOf, located in
Infectious disease (1 Class)
composed by, describes / is a representation of, has_part, has_proper_part, subject of clinical record
Patient record (1 Class)
composed by, describes / is a representation of, includes, output of, transformation_of
Sleep arousal finding (1 Class)
composed by, describes / is a representation of, hasDiastolicBPValue, hasUnit, output of
Blood pressure – diastolic (1 Class)
composed by, describes / is a representation of, hasUnit, output of, statisticalMeasureOf
Ankle to Brachial pressure index (ABPI, ABI) (1 Class)
Average diastolic blood pressure (BP) (1 Class)
Body Mass Index (BMI) finding (1 Class)
Sleep arousal index (1 Class)
composed by, describes / is a representation of, hasConsciousnessState, hasUnit, output of
Mean Hemoglobin Oxygen (HB O2) Saturation during Sleep (3 Classes)
Minimum Hemoglobin Oxygen (HB O2) Saturation during Sleep (3 Classes)
Percentage time in NREM sleep (5 Classes)
Percentage time in REM sleep (1 Class)
Time in REM Sleep (1 Class)
Total sleep time (TST) (1 Class)
Time in NREM Sleep (1 Class)
composed by, describes / is a representation of, hasUnit, measurement of, output of
Respiratory flow amplitude finding (6 Classes)
Respiratory effort amplitude finding (5 Classes)
Respiratory airflow rate (1 Class)
composed by, describes / is a representation of, hasUnit, number of, output of
Number of sleep apnea findings (1 Class)
Number of sleep hypopnea findings (1 Class)
composed by, describes / is a representation of, hasConsciousnessState, includes, output of
Sleep hypopnea finding (8 Classes)
Sleep apnea finding (2 Classes)
composed by, describes / is a representation of, hasSystolicBPValue, hasUnit, output of
Blood pressure - systolic (1 Class)
has contraindication, has_agent, has_participant, indicated by, hasRole
Therapeutic act (5 Classes)
composed by, describes / is a representation of, hasBloodPressureValue, hasUnit, output of
Mean arterial blood pressure (BP) (1 Class)
composed by, describes / is a representation of, hasUnit, includes, output of, statisticalMeasureOf
Apnea hypopnea index (AHI) (30 Classes)
Periodic leg movement (PLM) (1 Class)
Respiratory disturbance index (RDI) (1 Class)
composed by, describes / is a representation of, hasUnit, includes, number of, output of
Number of sleep apnea findings associated with 2% HbO2 desaturation (1 Class)
Number of sleep apnea findings associated with 3% HbO2 desaturation (1 Class)
Number of sleep apnea findings associated with 4% HbO2 desaturation (1 Class)
Number of sleep apnea findings associated with 5% HbO2 desaturation (1 Class)
Number of sleep apnea findings associated with a non-supine body position (1 Class)
Number of sleep hypopnea findings associated with 5% HbO2 desaturation (1 Class)
Number of sleep hypopnea findings associated with arousal (1 Class)
Number of sleep apnea findings associated with arousal (1 Class)
Number of sleep apnea findings associated with supine body position (1 Class)
Number of sleep hypopnea findings associated with 2% HbO2 desaturation (1 Class)
Number of sleep hypopnea findings associated with 3% HbO2 desaturation (1 Class)
Number of sleep hypopnea findings associated with 4% HbO2 desaturation (1 Class)
Number of sleep hypopnea findings associated with a non-supine body position (1 Class)
Number of sleep hypopnea findings associated with supine body position (1 Class)
composed by, describes / is a representation of, hasUnit, output of, statisticalMeasureOf, transformation_of
NREM sleep arousal index (1 Class)
REM sleep arousal index (1 Class)
composed by, describes / is a representation of, has finding site, output of, hasRole, subject of clinical record
clinical finding (4 Classes)
acts on anatomy, approach site, has input, has output, has_agent, has_participant, hasRole
procedure (1 Class)
composed by, describes / is a representation of, hasDuration, hasUnit, includes, output of, transformation_of
Sleep arousal finding - AASM 2007 (1 Class)
composed by, describes / is a representation of, has finding site, hypothesized problem, output of, hasRole, subject of clinical record
clinical diagnosis (1 Class)
composed by, describes / is a representation of, hasConsciousnessState, hasUnit, includes, output of, statisticalMeasureOf
Apnea hypopnea index (AHI) during REM sleep (1 Class)
Apnea hypopnea index (AHI) during NREM sleep (1 Class)
composed by, describes / is a representation of, hasBodyPosition, hasSystolicBPValue, hasUnit, has_participant, output of
Blood pressure - systolic - sitting body position (1 Class)
Blood pressure - systolic - supine body position (1 Class)
composed by, describes / is a representation of, hasBodyPosition, hasDiastolicBPValue, hasUnit, has_participant, output of
Blood pressure - diastolic - sitting body position (1 Class)
composed by, describes / is a representation of, hasConsciousnessState, hasDuration, hasUnit, includes, output of
Apnea finding - AASM manual 2007 definition (3 Classes)
composed by, describes / is a representation of, hasConsciousnessState, hasDuration, hasUnit, has_part, includes, output of
Mixed sleep apnea (MSA) finding (1 Class)
composed by, describes / is a representation of, hasBodyPosition, hasSystolicBPValue, hasUnit, has_participant, output of, procedureSite
Blood pressure - systolic in lower limb - supine body position (7 Classes)
Blood pressure - systolic in upper limb - supine body position (1 Class)
acts on anatomy, has input, has output, has_agent, has_participant, investigates, indicated by, hasRole
Clinical investigation act (7 Classes)
acts on anatomy, approach site, has input, has output, has_agent, has_participant, investigates, indicated by, hasRole
Diagnostic procedure (9 Classes)
acts on anatomy, approach site, has contraindication, has input, has output, has_agent, has_participant, indicated by, hasRole, surgical method
Therapeutic procedure (1 Class)
acts on anatomy, approach site, has input, has locus, has output, has_agent, has_participant, investigates, indicated by, hasRole
Respiratory system test (6 Classes)
cardiovascular system test (1 Class)
acts on anatomy, approach site, has input, has output, has_agent, has_part, has_participant, investigates, indicated by, hasRole
Polysomnography (PSG) test (4 Classes)
acts on anatomy, approach site, has input, has locus, has output, hasParticipant, has_agent, has_participant, investigates, indicated by, hasRole
Respiratory airflow measurement test (3 Classes)
acts on anatomy, approach site, has input, has locus, has output, hasBodyPosition, has_agent, has_participant, investigates, indicated by, hasRole, procedureSite
Blood pressure measurement procedure (1 Class)

Figure 3.26 The (domain or restriction)-defined partial-area taxonomy for the Sleep Domain Ontology's *Entity* hierarchy. The {*part of*} and {*has part*} areas are not shown.
Source: [120]

Figure 3.26 illustrates the (domain or restriction)-defined partial-area taxonomy for SDO's *Entity* hierarchy. This taxonomy has 267 partial-areas in 67 areas (abstraction ratio of 4.77). While the (domain or restriction)-defined partial-area taxonomy has approximately the same abstraction ratio as the restriction-defined taxonomy, it provides a more complete summary of the SDO's structure and is still not overwhelming.

3.3.4 In Support of Quality Assurance

One way to perform quality assurance using a partial-area taxonomy is to review the taxonomy to see whether it conforms to the original conception that the designer of the ontology had. For example, do the classes in the various partial-areas indeed have the correct sets of object properties? Such a review can be done by an individual who is familiar with the content and structure of the ontology. Another way of utilizing the taxonomy is by identifying any components that display an anomaly vis-à-vis the rest of the ontology. For example, a partial-area that is much larger than all the other partial-areas might be considered an anomaly in an ontology the size of OCRe. Another example is a partial-area in which a very large number of object properties are introduced.

A further anomaly may relate to exceptions in the number of *child-of* relationships emanating from a partial-area. If, for example, most partial-areas have just one *child-of* and only a few have multiple *child-of*'s, the latter constitute an exception to the norm and are recommended for review. It is not necessarily the case that each such anomaly manifests an error, but the anomalous classes are recommended for in-depth review by a curator of the ontology. Some anomalies are the results of modeling errors that can be discovered during an in-depth review.

3.3.4.1 OCRe Quality Assurance Review. The complete domain-defined partial-area taxonomy for the *Entity* hierarchy, shown in Figure 3.27, was created for Version 244 of OCRe. This version of *Entity* consists of 120 classes and 75 unique types of object properties. Levels are numbered, with the root area {*has part*, *part of*} at Level 0. Lower levels have larger level numbers and also larger numbers of object properties; however, these numbers are not necessarily equal.

The partial-area *Physical entity* is in the area {*has part, part of, is element of, plays*} at Level 2. *Physical entity* has three classes, the root and its two children, *Material* and *Organism* (see Figure 3.21).

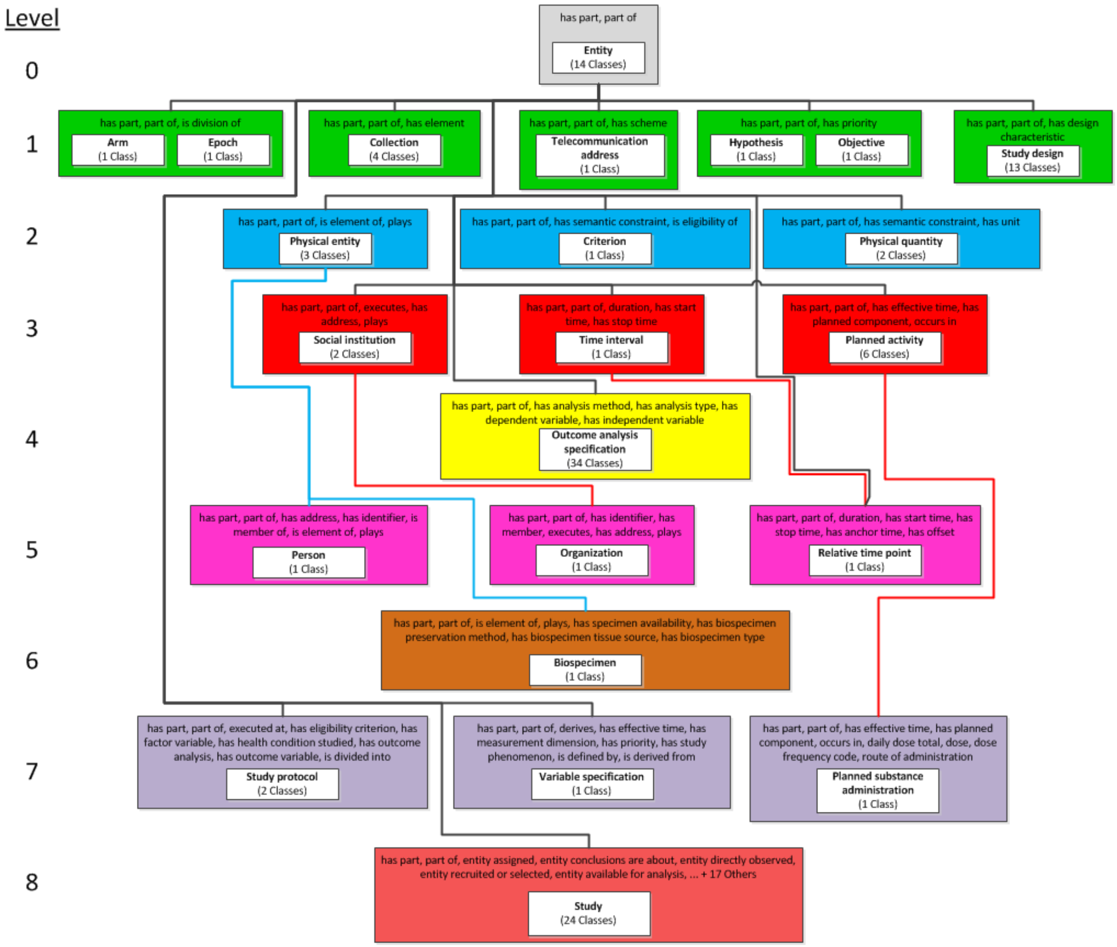

Level
0
1
2
3
4
5
6
7
8
has part, part of
Entity
(14 Classes)
has part, part of, is division of
Arm
(1 Class)
Epoch
(1 Class)
has part, part of, has element
Collection
(4 Classes)
has part, part of, has scheme
Telecommunication address
(1 Class)
has part, part of, has priority
Hypothesis
(1 Class)
Objective
(1 Class)
has part, part of, has design characteristic
Study design
(13 Classes)
has part, part of, is element of, plays
Physical entity
(3 Classes)
has part, part of, has semantic constraint, is eligibility of
Criterion
(1 Class)
has part, part of, has semantic constraint, has unit
Physical quantity
(2 Classes)
has part, part of, executes, has address, plays
Social institution
(2 Classes)
has part, part of, duration, has start time, has stop time
Time interval
(1 Class)
has part, part of, has effective time, has planned component, occurs in
Planned activity
(6 Classes)
has part, part of, has analysis method, has analysis type, has dependent variable, has independent variable
Outcome analysis specification
(34 Classes)
has part, part of, has address, has identifier, is member of, is element of, plays
Person
(1 Class)
has part, part of, has identifier, has member, executes, has address, plays
Organization
(1 Class)
has part, part of, duration, has start time, has stop time, has anchor time, has offset
Relative time point
(1 Class)
has part, part of, is element of, plays, has specimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type
Biospecimen
(1 Class)
has part, part of, executed at, has eligibility criterion, has factor variable, has health condition studied, has outcome analysis, has outcome variable, is divided into
Study protocol
(2 Classes)
has part, part of, derives, has effective time, has measurement dimension, has priority, has study phenomenon, is defined by, is derived from
Variable specification
(1 Class)
has part, part of, has effective time, has planned component, occurs in, daily dose total, dose, dose frequency code, route of administration
Planned substance administration
(1 Class)
has part, part of, entity assigned, entity conclusions are about, entity directly observed, entity recruited or selected, entity available for analysis, ... + 17 Others
Study
(24 Classes)

Figure 3.27 Complete partial-area taxonomy for OCRe's *Entity* hierarchy prior to the auditing efforts.
Source: [25]

There are two partial-areas that are *child-of* the partial-area *Physical entity*: *Person* which is a subclass of *Organism*, and *Biospecimen* which is a subclass of *Material entity* in the ontology. The corresponding *child-of* relationships are shown as (blue) lines in Figure 3.27. In total, the taxonomy has 21 areas organized into nine levels. There are 23 partial-areas in total, because two areas at Level 1 contain two partial-areas.

Twelve of the partial-areas consist of just one class. To observe the main focus of the content of the *Entity* hierarchy, one should concentrate on the larger partial-areas: *Entity* (14 classes), *Study design* (13 classes), *Outcome analysis specification* (34 classes), *Planned activity* (6 classes), and *Study* (24 classes). By reviewing the 23 partial-areas and concentrating on the large ones, one can get an orientation into the structure and content of a hierarchy. The two largest partial-areas, *Outcome analysis specification* and *Study*, could be considered anomalous. At first, the OCRe curators were surprised to see so many classes included within the former.

Upon closer inspection, it was seen that all 33 descendants of *Outcome analysis specification* describe statistical methods that clearly did not belong under this class. Furthermore, they did not even belong in the *Entity* hierarchy. The reasoner inferred the subsumption relationship because of the erroneous domain specifications of the object properties *has dependent variable* and *has independent variable*. After the error was fixed by OCRe's curator, the 33 classes no longer show up as inferred descendants of *Outcome analysis specification*.

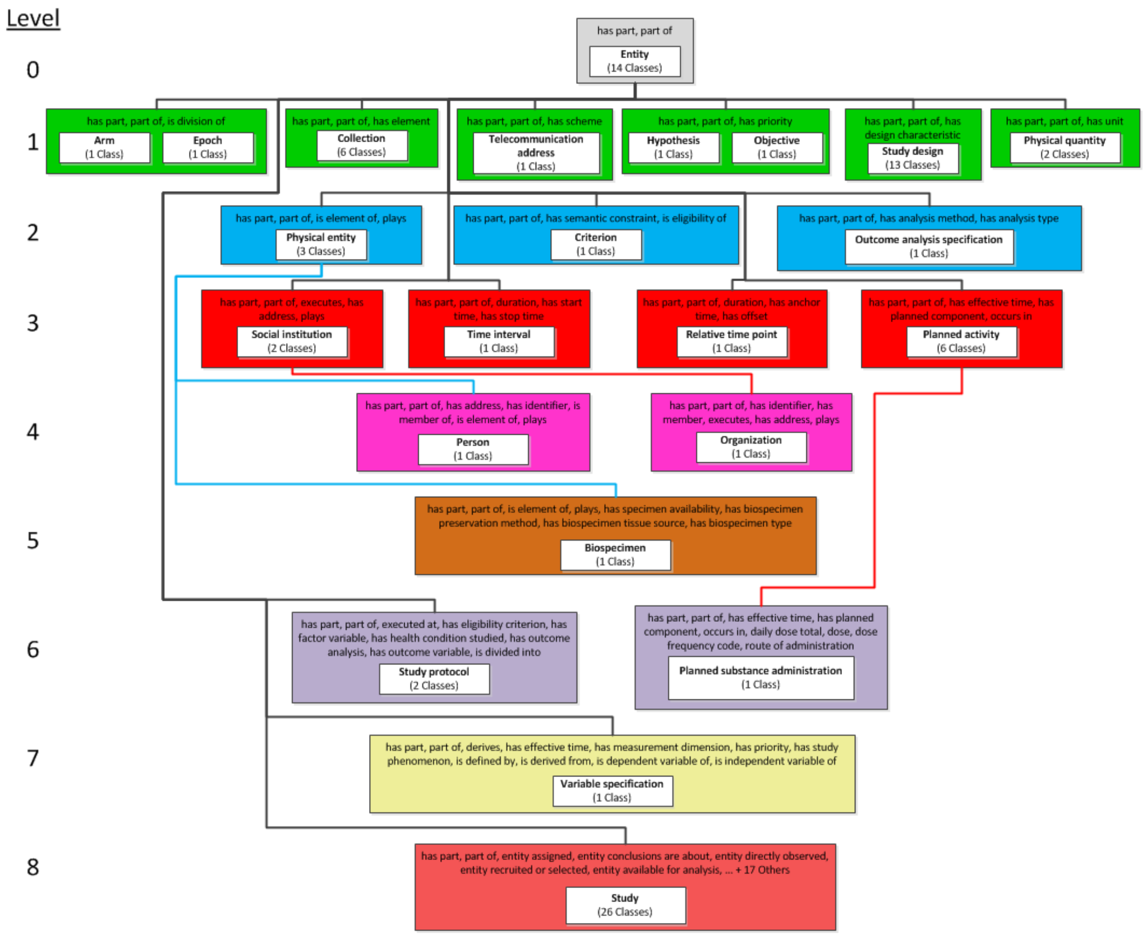

Level
0
1
2
3
4
5
6
7
8
has part, part of
Entity
(14 Classes)
has part, part of, is division of
Arm
(1 Class)
Epoch
(1 Class)
has part, part of, has element
Collection
(6 Classes)
has part, part of, has scheme
Telecommunication address
(1 Class)
has part, part of, has priority
Hypothesis
(1 Class)
Objective
(1 Class)
has part, part of, has design characteristic
Study design
(13 Classes)
has part, part of, has unit
Physical quantity
(2 Classes)
has part, part of, is element of, plays
Physical entity
(3 Classes)
has part, part of, has semantic constraint, is eligibility of
Criterion
(1 Class)
has part, part of, has analysis method, has analysis type
Outcome analysis specification
(1 Class)
has part, part of, executes, has address, plays
Social institution
(2 Classes)
has part, part of, duration, has start time, has stop time
Time interval
(1 Class)
has part, part of, duration, has anchor time, has offset
Relative time point
(1 Class)
has part, part of, has effective time, has planned component, occurs in
Planned activity
(6 Classes)
has part, part of, has address, has identifier, is member of, is element of, plays
Person
(1 Class)
has part, part of, has identifier, has member, executes, has address, plays
Organization
(1 Class)
has part, part of, is element of, plays, has specimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type
Biospecimen
(1 Class)
has part, part of, executed at, has eligibility criterion, has factor variable, has health condition studied, has outcome analysis, has outcome variable, is divided into
Study protocol
(2 Classes)
has part, part of, has effective time, has planned component, occurs in, daily dose total, dose, dose frequency code, route of administration
Planned substance administration
(1 Class)
has part, part of, derives, has effective time, has measurement dimension, has priority, has study phenomenon, is defined by, is derived from, is dependent variable of, is independent variable of
Variable specification
(1 Class)
has part, part of, entity assigned, entity conclusions are about, entity directly observed, entity recruited or selected, entity available for analysis, ... + 17 Others
Study
(26 Classes)

Figure 3.28 Partial-area taxonomy for OCRe’s *Entity* hierarchy, revised after audit.
Source: [25]

Figure 3.28 shows the taxonomy of the revised *Entity* hierarchy, which has only 88 classes. It was made available on BioPortal as Version 258 of OCRe. In the revised taxonomy, the partial-area *Outcome analysis specification* contains just one class on Level 2 (blue) with only four object properties. *Outcome analysis specification* was removed as the domain of two object properties, *has dependent variable* and *has independent variable*, which were formulated as existential restrictions on *Outcome analysis specification. Variable specification* was added as the domain of these properties’ inverses.

Another anomaly was encountered at the partial-area *Relative time point* (one class) at Level 5 of Figure 3.27. It is the only partial-area that has two *child-of*'s emanating from it, one to the partial-area *Entity*, where it is a subclass of the *Time point* class, and the other to the partial-area *Time interval*, where it is a subclass of the root class *Time interval*.

According to the definition provided within the ontology, *Relative time point* is not a *Time interval* at all, but a *Time point* in reference to some other given time point. Furthermore, the subclass to *Time interval* was not in the asserted view of OCRe, but was instead inferred by the Pellet reasoner. The error was due to an error in the specified domain of the *duration* object property, leading the reasoner to infer an unintended subsumption relationship. During the review of OCRe, the domain of the property was changed and as a result the second subclass relationship to *Time interval* is no longer inferred.

In Figure 3.28, *Relative time point* appears on Level 3 (red), with only five object properties. This is due to the fact that the two object properties *has start time* and *has stop time,* originally inherited from *Time interval*, disappeared since *Relative time point* is no longer a subclass of *Time interval*.

Another change in the ontology resulted from observations made upon review of the various partial-areas and their sets of object properties. The partial-area *Physical quantity* (two classes) on Level 2 (blue) had an irrelevant object property *has semantic constraint,* which was removed. This partial-area is now on Level 1 (green) of the taxonomy (Figure 3.28).

Comparing Figure 3.27 to Figure 3.28, it can be seen that changes occurred on all levels except Level 0. However, as it happens, Level 5 (comprising one class *Biospecimen*) in Figure 3.28 is identical to Level 6 in Figure 3.27. On every other level, there are significant changes between the two figures.

The remodeling of OCRe following the auditing that was facilitated by the taxonomy of Figure 3.27 has not yet been completed. In addition to the changes reflected in Figure 3.28 and presented in the previous section, there is some additional remodeling work underway for the partial-area *Study* (bottom of Figure 3.27 on Level 8). This partial-area has 24 object properties, a large increase versus the nine object properties at Level 7 in the taxonomy. It is surprising that besides *part of* and *has part* all 22 other object properties have the *Study* class as their domain. One could envision some of the object properties having children or grandchildren of *Study* as their domains instead.

For example, the object properties *has recruitment status* or *has biospecimen collected* may not be relevant for all 24 classes of this partial-area and should be introduced at appropriate descendants. This modification would partition the large partial-area *Study* into several smaller ones, likely improving the presentation of OCRe to users. The editorial team of OCRe is currently using this feedback to re-examine the modeling of these classes, which are critical to the purpose of OCRe. Some initial changes are reflected in Figure 3.28, where *Study* has grown from 24 to 26 classes, compared to Figure 3.27, reflecting a finer distinction between classes.

3.3.4.2 Sleep Domain Ontology Quality Assurance Review. Ochs et al. [120] describe a preliminary quality assurance review of the SDO using the (domain or restriction)-defined taxonomy shown in Figure 3.26. This preliminary review identified several errors that were later corrected by SDO's curator, Sivaram Arabandi..

The first issue, a dissimilar partial area grouping, was noticed at level 2 in the area {*has part, hasRole*}. This area has three partial-areas, one of which (*Asian or Pacific Islander*) does not match the other two partial-areas about Angiotensin. They fall under very different hierarchies – first one is a subclass of population, and the other two are classes under the medication hierarchy. Upon investigation, the class *Asian or Pacific Islander* was introduced for cases where records did not distinguish between the two races and the actual race is not known. The semantics of such a situation fits the OR logical operator, as the term describes, and does not fit a part relation [134]. An individual with this race is not part Asian and part Pacific Islander, but is one of the two. The knowledge is just not available. Thus the *has part* object property was removed from this class and it is now in the *Independent continuant* partial-area in the {*hasRole*} area, like all of its sibling races.

This modeling error was discovered only due to the dissimilar grouping in the area {*hasRole*, *has part*}. This area consists of object property *hasRole* with an explicit domain and object property *has part*, which is used in a restriction. Hence, this area did not appear in the domain-defined partial-area taxonomy of Figure 3.25. Neither did it appear in the restriction-defined partial-area taxonomy of Figure 3.24. The only taxonomy where this dissimilarity appeared was in the (domain or restriction)-defined partial-area taxonomy of Figure 3.26. This example demonstrates why granularity has to be considered when utilizing abstraction networks for quality assurance. The classes in partial-area for living organism were found to have duplicate properties –

"*participatesIn*" (from BioTop) and "*participates in"* (from RO). On examination, neither of the two relations have a description associated with them. However, based on the usage of the relations, it appears that the two are equivalent. Neither property has domain or range specified, but the RO version has a subproperty and an inverse property associated with it. The BioTop version of the relation is used only once (in the definition of living organism). Therefore, SDO was refactored to replace this relation with the one from RO.

3.3.4.3 Gene Ontology Quality Assurance Review. The Gene Ontology (GO) [42] is an important ontology utilized extensively to support annotation of genomics findings [135]. GO comprises over 40,000 terms (i.e., classes/concepts) and nearly 95,000 synonyms. GO terms are connected by about 64,000 hierarchical *is_a* links that collectively form a directed acyclic graph. GO's terms are further defined using almost 15,000 relationships, including many *part_of* and *has_part* relationships. Additionally, GO is extensively cross-mapped to other ontologies and external references. It is separated into three subsets: Biological process (BP), which describes biological events; Cellular component, which describes different cell parts; and Molecular function, which describes activities at the molecular level.

Due to the size and complexity of GO, modeling problems and inconsistencies are nearly unavoidable. Thus, it is imperative to develop quality assurance techniques for GO's content. GO is the largest member of the Open Biomedical Ontologies (OBO) Foundry [52], a collection of biomedical ontologies that adhere to a common design philosophy and implementation. Abstraction networks, e.g., partial-area taxonomies, can be used to support the quality assurance of GO's content.

A preliminary review of GO’s Biological process (BP) hierarchy was performed, using a partial-area taxonomy, in Ochs et al. [136]. Various kinds of anomalies and their impact on GO were analyzed. When a kind of anomaly is repeated multiple times, and it is proven to indicate modeling problems with a high degree of likelihood, then it *de facto* forms part of a quality assurance regimen. For example, one of the anomalies frequently found within GO’s partial-area taxonomy is the *overlapping term* (i.e., overlapping concept [23]). Such terms were found to be statistically more likely to be in error in SNOMED CT [29]. By focusing on these terms, it is expected that more errors will be identified and corrected, as compared to reviewing random terms from GO’s general population.

The partial-area taxonomy for the BP hierarchy, consisting of 25,635 terms (February 2014 release), comprises 1,653 partial-areas with 27 areas. That works out to an abstraction ratio of approximately 15:1 (terms to partial-areas). The largest area is {*part of*} on Level 1, with 1,005 partial-areas summarizing 10,934 terms (42.7% of the hierarchy). Figure 3.29 shows a significant portion of partial-area taxonomy. Due to space limitations, Level 5 is not shown, {*part of*} on Level 1 has been truncated to its 72 largest partial-areas (capturing 65.5% of its terms), and several small areas have been omitted. The complete GO partial-area taxonomy can be viewed at [132]. Long partial-area names are abbreviated using ellipses, e.g., *regulation of blood pressure* is written as “*regulation of blood*…”. When referring to such a long-named partial-area, its complete name will be written. Within each area, partial-areas are sorted into rows by their size in left-to-right order, internally sub-sorted alphabetically according to their names.

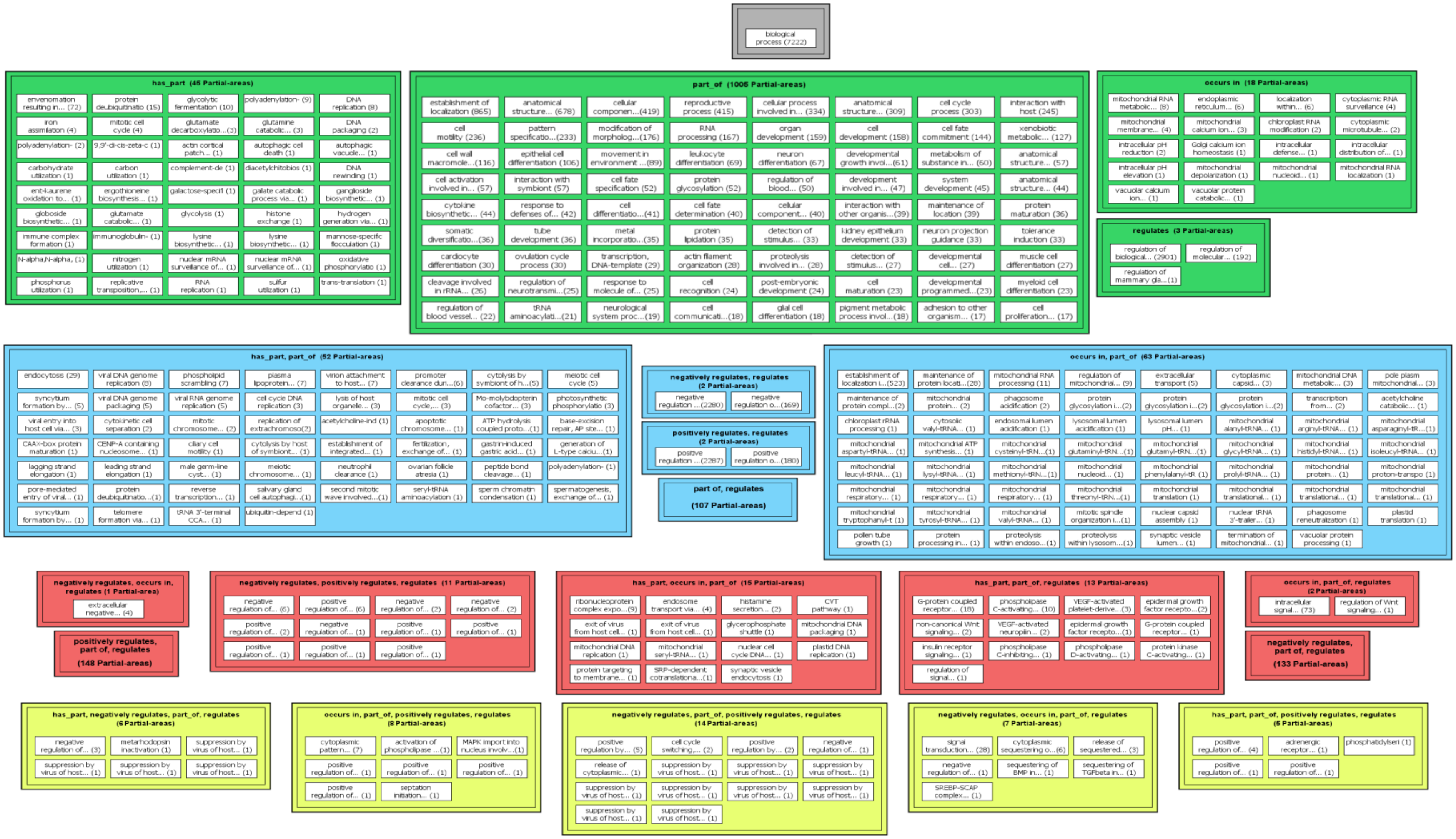

biological process (7222)
has_part (45 Partial-areas)
envenomation resulting in... (72)
protein deubiquitinatio (15)
glycolytic fermentation (10)
polyadenylation- (9)
DNA replication (8)
iron assimilation (4)
mitotic cell cycle (4)
glutamate decarboxylatio...(3)
glutamine catabolic... (3)
DNA packaging (2)
polyadenylation- (2)
9,9'-di-cis-zeta-c (1)
actin cortical patch... (1)
autophagic cell death (1)
autophagic vacuole... (1)
carbohydrate utilization (1)
carbon utilization (1)
complement-de (1)
diacetylchitobios (1)
DNA rewinding (1)
ent-kaurene oxidation to... (1)
ergothioneine biosynthesis... (1)
galactose-specifi (1)
gallate catabolic process via... (1)
ganglioside biosynthetic... (1)
globoside biosynthetic... (1)
glutamate catabolic... (1)
glycolysis (1)
histone exchange (1)
hydrogen generation via... (1)
immune complex formation (1)
immunoglobulin- (1)
lysine biosynthetic... (1)
lysine biosynthetic... (1)
mannose-specific flocculation (1)
N-alpha,N-alpha, (1)
nitrogen utilization (1)
nuclear mRNA surveillance of... (1)
nuclear mRNA surveillance of... (1)
oxidative phosphorylatio (1)
phosphorus utilization (1)
replicative transposition,... (1)
RNA replication (1)
sulfur utilization (1)
trans-translation (1)
part_of (1005 Partial-areas)
establishment of localization (865)
anatomical structure... (678)
cellular componen...(419)
reproductive process (415)
cellular process involved in... (334)
anatomical structure... (309)
cell cycle process (303)
interaction with host (245)
cell motility (236)
pattern specificatio...(233)
modification of morpholog...(176)
RNA processing (167)
organ development (159)
cell development (158)
cell fate commitment (144)
xenobiotic metabolic... (127)
cell wall macromole...(116)
epithelial cell differentiation (106)
movement in environment ...(89)
leukocyte differentiation (69)
neuron differentiation (67)
developmental growth invol...(61)
metabolism of substance in... (60)
anatomical structure... (57)
cell activation involved in... (57)
interaction with symbiont (57)
cell fate specification (52)
protein glycosylation (52)
regulation of blood... (50)
development involved in... (47)
system development (45)
anatomical structure... (44)
cytokine biosynthetic... (44)
response to defenses of... (42)
cell differentiatio...(41)
cell fate determination (40)
cellular component... (40)
interaction with other organis...(39)
maintenance of location (39)
protein maturation (36)
somatic diversificatio...(36)
tube development (36)
metal incorporatio...(35)
protein lipidation (35)
detection of stimulus... (33)
kidney epithelium development (33)
neuron projection guidance (33)
tolerance induction (33)
cardiocyte differentiation (30)
ovulation cycle process (30)
transcription, DNA-template (29)
actin filament organization (28)
proteolysis involved in... (28)
detection of stimulus... (27)
developmental cell... (27)
muscle cell differentiation (27)
cleavage involved in rRNA... (26)
regulation of neurotransmi...(25)
response to molecule of... (25)
cell recognition (24)
post-embryonic development (24)
cell maturation (23)
developmental programmed...(23)
myeloid cell differentiation (23)
regulation of blood vessel... (22)
tRNA aminoacylati...(21)
neurological system proc...(19)
cell communicati...(18)
glial cell differentiation (18)
pigment metabolic process invol...(18)
adhesion to other organism... (17)
cell proliferation... (17)
occurs in (18 Partial-areas)
mitochondrial RNA metabolic... (8)
endoplasmic reticulum... (6)
localization within... (6)
cytoplasmic RNA surveillance (4)
mitochondrial membrane... (4)
mitochondrial calcium ion... (3)
chloroplast RNA modification (2)
cytoplasmic microtubule... (2)
intracellular pH reduction (2)
Golgi calcium ion homeostasis (1)
intracellular defense... (1)
intracellular distribution of... (1)
intracellular pH elevation (1)
mitochondrial depolarization (1)
mitochondrial nucleoid... (1)
mitochondrial RNA localization (1)
vacuolar calcium ion... (1)
vacuolar protein catabolic... (1)
regulates (3 Partial-areas)
regulation of biological... (2901)
regulation of molecular... (192)
regulation of mammary gla...(1)
has_part, part_of (52 Partial-areas)
endocytosis (29)
viral DNA genome replication (8)
phospholipid scrambling (7)
plasma lipoprotein... (7)
virion attachment to host... (7)
promoter clearance duri...(6)
cytolysis by symbiont of h...(5)
meiotic cell cycle (5)
syncytium formation by... (5)
viral DNA genome packaging (5)
viral RNA genome replication (5)
cell cycle DNA replication (3)
lysis of host organelle... (3)
mitotic cell cycle,... (3)
Mo-molybdopterin cofactor... (3)
photosynthetic phosphorylatio (3)
viral entry into host cell via... (3)
cytokinetic cell separation (2)
mitotic chromosome... (2)
replication of extrachromoso(2)
acetylcholine-ind (1)
apoptotic chromosome... (1)
ATP hydrolysis coupled proto...(1)
base-excision repair, AP site... (1)
CAAX-box protein maturation (1)
CENP-A containing nucleosome... (1)
ciliary cell motility (1)
cytolysis by host of symbiont... (1)
establishment of integrated... (1)
fertilization, exchange of... (1)
gastrin-induced gastric acid... (1)
generation of L-type calciu...(1)
lagging strand elongation (1)
leading strand elongation (1)
male germ-line cyst... (1)
meiotic chromosome... (1)
neutrophil clearance (1)
ovarian follicle atresia (1)
peptide bond cleavage... (1)
polyadenylation- (1)
pore-mediated entry of viral... (1)
protein deubiquitinatio...(1)
reverse transcription... (1)
salivary gland cell autophagi...(1)
second mitotic wave involved...(1)
seryl-tRNA aminoacylation (1)
sperm chromatin condensation (1)
spermatogenesis, exchange of... (1)
syncytium formation by... (1)
telomere formation via... (1)
tRNA 3'-terminal CCA... (1)
ubiquitin-depend (1)
negatively regulates, regulates (2 Partial-areas)
negative regulation ...(2280)
negative regulation o...(169)
positively regulates, regulates (2 Partial-areas)
positive regulation ...(2287)
positive regulation o...(180)
part of, regulates (107 Partial-areas)
occurs in, part_of (63 Partial-areas)
establishment of localization i...(523)
maintenance of protein locati...(28)
mitochondrial RNA processing (11)
regulation of mitochondrial... (9)
extracellular transport (5)
cytoplasmic capsid... (3)
mitochondrial DNA metabolic... (3)
pole plasm mitochondrial... (3)
maintenance of protein compl...(2)
mitochondrial protein... (2)
phagosome acidification (2)
protein glycosylation i...(2)
protein glycosylation i...(2)
protein glycosylation i...(2)
transcription from... (2)
acetylcholine catabolic... (1)
chloroplast rRNA processing (1)
cytosolic valyl-tRNA... (1)
endosomal lumen acidification (1)
lysosomal lumen acidification (1)
lysosomal lumen pH... (1)
mitochondrial alanyl-tRNA... (1)
mitochondrial arginyl-tRNA... (1)
mitochondrial asparaginyl-tR...(1)
mitochondrial aspartyl-tRNA... (1)
mitochondrial ATP synthesis... (1)
mitochondrial cysteinyl-tRN...(1)
mitochondrial glutaminyl-tRN...(1)
mitochondrial glutamyl-tRN...(1)
mitochondrial glycyl-tRNA... (1)
mitochondrial histidyl-tRNA... (1)
mitochondrial isoleucyl-tRNA... (1)
mitochondrial leucyl-tRNA... (1)
mitochondrial lysyl-tRNA... (1)
mitochondrial methionyl-tRN...(1)
mitochondrial nucleoid... (1)
mitochondrial phenylalanyl-tR (1)
mitochondrial prolyl-tRNA... (1)
mitochondrial protein... (1)
mitochondrial proton-transpo (1)
mitochondrial respiratory... (1)
mitochondrial respiratory... (1)
mitochondrial respiratory... (1)
mitochondrial threonyl-tRN...(1)
mitochondrial translation (1)
mitochondrial translational... (1)
mitochondrial translational... (1)
mitochondrial translational... (1)
mitochondrial tryptophanyl-t (1)
mitochondrial tyrosyl-tRNA... (1)
mitochondrial valyl-tRNA... (1)
mitotic spindle organization i...(1)
nuclear capsid assembly (1)
nuclear tRNA 3'-trailer... (1)
phagosome reneutralization (1)
plastid translation (1)
pollen tube growth (1)
protein processing in... (1)
proteolysis within endoso...(1)
proteolysis within lysosom...(1)
synaptic vesicle lumen... (1)
termination of mitochondrial... (1)
vacuolar protein processing (1)
negatively regulates, occurs in, regulates (1 Partial-area)
extracellular negative... (4)
positively regulates, part of, regulates (148 Partial-areas)
negatively regulates, positively regulates, regulates (11 Partial-areas)
negative regulation of... (6)
positive regulation of... (6)
negative regulation of... (2)
negative regulation of... (2)
positive regulation of... (2)
negative regulation of... (1)
positive regulation of... (1)
positive regulation of... (1)
positive regulation of... (1)
positive regulation of... (1)
positive regulation of... (1)
has_part, occurs in, part_of (15 Partial-areas)
ribonucleoprotein complex expo...(9)
endosome transport via... (4)
histamine secretion... (2)
CVT pathway (1)
exit of virus from host cell... (1)
exit of virus from host cell... (1)
glycerophosphate shuttle (1)
mitochondrial DNA packaging (1)
mitochondrial DNA replication (1)
mitochondrial seryl-tRNA... (1)
nuclear cell cycle DNA... (1)
plastid DNA replication (1)
protein targeting to membrane... (1)
SRP-dependent cotranslationa...(1)
synaptic vesicle endocytosis (1)
has_part, part_of, regulates (13 Partial-areas)
G-protein coupled receptor... (18)
phospholipase C-activating... (10)
VEGF-activated platelet-derive...(3)
epidermal growth factor recepto...(2)
non-canonical Wnt signaling... (2)
VEGF-activated neuropilin... (2)
epidermal growth factor recepto...(1)
G-protein coupled receptor... (1)
insulin receptor signaling... (1)
phospholipase C-inhibiting... (1)
phospholipase D-activating... (1)
protein kinase C-activating... (1)
regulation of signal... (1)
occurs in, part_of, regulates (2 Partial-areas)
intracellular signal... (73)
regulation of Wnt signaling... (1)
negatively regulates, part of, regulates (133 Partial-areas)
has_part, negatively regulates, part_of, regulates (6 Partial-areas)
negative regulation of... (3)
metarhodopsin inactivation (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
occurs in, part_of, positively regulates, regulates (8 Partial-areas)
cytoplasmic pattern... (7)
activation of phospholipase ...(1)
MAPK import into nucleus involv...(1)
positive regulation of... (1)
positive regulation of... (1)
positive regulation of... (1)
positive regulation of... (1)
septation initiation... (1)
negatively regulates, part_of, positively regulates, regulates (14 Partial-areas)
positive regulation by... (5)
cell cycle switching,... (2)
positive regulation by... (2)
negative regulation of... (1)
release of cytoplasmic... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
suppression by virus of host... (1)
negatively regulates, occurs in, part_of, regulates (7 Partial-areas)
signal transduction... (28)
cytoplasmic sequestering o...(6)
release of sequestered... (3)
negative regulation of... (1)
sequestering of BMP in... (1)
sequestering of TGFbeta in... (1)
SREBP-SCAP complex... (1)
has_part, part_of, positively regulates, regulates (6 Partial-areas)
positive regulation of... (4)
adrenergic receptor... (1)
phosphatidylseri (1)
positive regulation of... (1)
positive regulation of... (1)

Figure 3.29 Excerpt of GO's *Biological process* partial-area taxonomy. Due to space limitations, certain areas are hidden and only a subset of partial-areas is shown for {*part of*}. Child-of's between partial-areas are hidden to enable readability. The number of terms in each partial-area is shown in parenthesis. *Source: [136]*

As can be seen, the BP partial-area taxonomy summarizes the content and complex structure of GO. For example, looking at larger partial-areas allows one to identify large groups of structurally and semantically similar terms.

The first row of the Level-1 area {*part_of*}, with partial-areas of size 240 (terms) and up, identifies the area's major types of terms and their frequencies. For example, looking at the first row in {*part_of*}, one sees establishment of localization (865), anatomical structure morphogenesis (678), and reproductive process (415). These are major types of terms in GO. For a more refined view, one can look at the partial-areas that are at lower rows. Table 3.22 summarizes the structure of the BP partial-area taxonomy across its six levels. For example, one can see that the majority of partial-areas and terms are on Level 1, indicating that most terms have only one relationship. Within the taxonomy, 1,474 overlapping terms were identified. The majority of overlapping terms (966, 65.5%) are in {*part of*}.

Table 3.23 summarizes {*part of*}'s overlapping terms according to their *degrees of overlap* (i.e., the number of partial-areas each term is summarized by).

Table 3.22 Structure of GO's Biological process Taxonomy by Levels

Level	# of Terms (%)	# of Areas	# of Partial-areas
0	7,222 (28%)	1	1
1	10,934 (43%)	4	1,071
2	6,420 (25%)	7	223
3	964 (4%)	7	314
4	93 (0.3%)	6	42
5	2 (0.007%)	2	2
Total:	25,635	27	1,653

Source: [136]

Table 3.23 Overlapping Terms in {*part of*} by Overlap Degree

Degree of Overlap	**# of Overlapping Terms**
2	855
3	94
4	13
5	4
Total:	966

Source: [136]

The taxonomy-based quality assurance regimen for GO is based on two heuristics that have been shown to be successful for the quality assurance of other ontologies (e.g., [22, 24, 25, 29, 119, 120]).

1. **Taxonomy anomaly:** when the taxonomy's summary of the ontology exhibits some kind of anomaly (e.g., an unexpected or irregular structural configuration that stands out), there is a higher likelihood of finding errors in that anomalous portion of the ontology.
2. **Term anomaly:** anomalous terms in the form of overlapping terms in a partial-area taxonomy have been shown to be statistically more likely to be erroneous than other terms.

The review of GO's partial-area taxonomy for anomalies can be conducted at three levels: the area level and the partial-area level (Item (1) above), and the term level (Item (2)). For the first two, an editor of GO can review the different taxonomic elements and determine if an area or partial-area stands out or summarizes terms with uncommon modeling.

When reviewing the areas of a partial-area taxonomy, the only specific data available are the areas' names (i.e., their relationship sets) and the numbers of terms summarized by the respective areas. Even so, areas can reveal potential errors in the relationship structure via anomalistic configurations.

One kind of an anomaly found in GO is a questionable combination of relationships revealed by the names of the various areas. For example, on Level 2 of the GO partial-area taxonomy (Figure 3.29), there are areas {*negatively regulates, regulates*} and {*positively regulates, regulates*}. *Positively regulates* and *negatively regulates* are both refined (child) relationships of the relationship *regulates*. Such combinations of refined relationships and more general parent relationships appearing together for the same term raise the issue of redundancy. According to ontological principles, the definition of two relationships for a given term where one is more refined than the other is only allowed to happen when the target of the refined relationship is a more specific term than the target of the more general relationship. For example, if a term has both *negatively regulates* and *regulates* relationships, then they should *not* have the same target. *Negatively regulates* should have a more refined target term. The existence of such relationship combinations raises questions about whether GO's modeling is following this ontological principle.

At the partial-area level, one can review an individual partial-area to determine if it has the correct set of relationships, it is grouped into the proper area, and its collection of summarized, member terms makes sense. One example of such an anomaly would be encountering an area with one very large partial-area and one very small partial-area. The contrasting sizes raise questions about the correctness of the terms' modeling in the smaller partial-area.

Consider the area {*regulates*} in the BP partial-area taxonomy (a portion of which is shown in Figure 3.29), with three partial-areas *regulation of biological process* (2901), *regulation of molecular function* (192), and *regulation of mammary gland cord* *elongation by mammary fat precursor cell-epithelial cell signaling* (1). This area has about 3,000 terms that are partitioned into two major groups, both of which are rooted at a general term. The existence of the singleton raises the question about why its term is special in relation to the other terms. That term has the same relationship structure, but it is not part of the other two partial-areas.

Overlapping terms can also be considered anomalous. The majority of GO's terms are summarized by only one partial-area, i.e., most terms are a specialization of only one root. Overlapping terms elaborate the semantics of multiple roots and are, thus, more complex and more difficult to model than non-overlapping terms. In SNOMED, overlapping concepts were found to be statistically more likely to harbor errors as compared to non-overlapping concepts [29].

This phenomenon was hypothesized to also occur in GO's content. If this was the case, a GO curator could focus on overlapping terms and expect to discover more errors then if they reviewed non-overlapping terms. To assess this phenomenon in GO, a preliminary study was performed to compare error rates among overlapping and non-overlapping terms. Jane Lomax, the coordinator of the GO Editorial Office, reviewed a sample of 40 overlapping terms and 20 non-overlapping terms to serve as a control. Additionally, terms with a higher degree of overlap were investigated to determine if they have higher error rates than terms with a degree of overlap of two, as seen previously for SNOMED [119].

Consider the anomaly of terms having the relationships *regulates* and *positively regulates* or *negatively regulates*, which manifests itself in the areas {*positively regulates*, *regulates*} on Level 2 and {*positively regulates*, *negatively regulates*, *regulates*} on

Level 3. In most cases, a term inherits *regulates* from a term in *regulation of biological process*, and also introduces, say, *negatively regulates*.

When a sample of the terms in these areas was analyzed, significant redundancy in the targets of the relationships was found. Many terms had either a *positively regulates* or *negatively regulates* with the same target as their *regulates* relationship. For example, *negative regulation of bone resorption* has *negatively regulates* and *regulates* to *bone resorption*. Similarly, the root term *positive regulation of molecular function* in {*positively regulates, regulates*} has both relationships targeting *molecular function*. Furthermore, in the area {*positively regulates*, *negatively regulates*, *regulates*}, the partial-area *positive regulation of molecular function in other organism* (6) has all three relationships to *molecular function*. In all these cases, *regulates* is redundant and should be removed.

This redundancy was in fact confirmed as an issue in the GO development pipeline. These redundant relationships will be automatically removed when the GO pipeline is enabled to delete no-longer inferable relationships. However, the areas of the taxonomy highlighted the potential existence of such redundancy, and many examples of redundant relationships were found by reviewing the taxonomy for anomalies.

Once GO enables the automatic removal of the redundant relationships, GO's taxonomy will change significantly. All terms that lose their *regulates* relationship and keep only *negatively regulates* or *positively regulates* will move one level up. Two new areas will then exist on Level 1: {*positively regulates*} and {*negatively regulates*}. Several other areas will likely come into existence at other levels, e.g., many terms may belong in {*part of, positively regulates*} or {*part of*, *negatively regulates*}. Of course, many terms legitimately have both *regulates* and *positively regulates*, and, thus, the areas {*positively regulates*, *regulates*}, etc., will likely still exist. Ultimately, GO's taxonomy will have a finer granularity.

Consider the anomaly regarding a singleton partial-area *regulation of mammary gland cord elongation by mammary fat precursor cell-epithelial cell signaling*, which is grouped with two larger partial-areas in the area {*regulates*}. Upon review of this term, it appeared to be missing a parent that should be in the partial-area *regulation of biological process*. The addition of this parent term to GO would imply the elimination of this singleton partial-area, leaving a large area with two large partial-areas—without anomalies. This error was confirmed, and the term should indeed have a regulation-related term as a parent. Currently, only *regulation of developmental growth* could serve as a parent term. To provide a complete fix, it would be necessary to add new intermediate terms, e.g., *regulation of mammary gland cord-elongation*. The GO editorial team will fully correct this error in due course.

To investigate the error rates of overlapping terms in GO the following samples of terms from {*part of*} were provided for domain-expert review by Jane Lomax: a random sample of 20 non-overlapping terms; a random sample of 20 terms with the minimum degree of overlap (two); and the group of all 17 terms with the highest overlap degrees (four and five) plus three randomly selected terms with a degree of overlap of three. The last group consists of the most complex terms in {*part of*}, as they elaborate the semantics of many roots.

Table 3.24 Quality Assurance Review Results According to Degree of Overlap

Degree of Overlap	# of Samples	# of Errors (%)
(none)	20	5 (25%)
2	20	7 (35%)
3–5	20	13 (65%)
Overlapping Total:	40	20 (50%)

Source: [136]

The samples were presented in alphabetical order according to term names. The degree of overlap was not given. Table 3.24 summarizes the findings. The percentages of modeling problems found were 25%, 35%, and 65% for the three groups, respectively. In total, 50% of overlapping terms had at least one problem, compared to 25% of the non-overlapping terms. Table 3.25 provides three examples of errors discovered among the overlapping terms. Several types of errors were found, including incorrect logical modeling, missing or incorrect parents, and missing relationships.

Table 3.25 Three Examples of Overlapping Term Errors

Term Name	Error
ascospore formation	Redundant parent**:** *cell development*
DNA replication termination involved in meiotic DNA replication	Incorrect logical definition
metabolism by symbiont of host xylan	Missing parent: *cell wall*

Source: [136]

The development of effective quality assurance methodologies for GO will enable improvements in its content. In the preliminary quality assurance review of GO’s BP hierarchy, relatively few terms were reviewed but a significant number of errors and inconsistencies were identified with the use of the BP partial-area taxonomy. These results are encouraging and help illustrate the feasibility of a comprehensive quality assurance review of GO based on the taxonomy. In future studies, different anomalies, such as “small” partial-areas (consisting of about 1–3 terms), already shown to be successful for the quality assurance of SNOMED CT and NCIt [22, 24], will be assessed for their usefulness in quality assurance of GO.

A limitation of focusing on the overlapping terms can be seen in their low numbers relative to the number of non-overlapping terms in BP: 1474 / 25635 = 5.7%. This limits the approach to a small portion of the hierarchy. However, reviewing the 1,474 overlapping terms for errors will require a relatively small effort that will likely result in uncovering more errors than reviewing a similar sized randomly selected sample of non-overlapping terms.

In the preliminary quality assurance study of GO all errors were counted, regardless of their severity. Some types of errors, e.g., incorrect logical modeling, will typically have a greater effect on the ontology as compared to less severe errors such as a redundant parent. The error rate for non-overlapping terms (25%) was higher than expected, and greater than what was found in the context of SNOMED [119]. The larger sample of non-overlapping terms that will be reviewed in future studies will provide further insight into their expected error rate. In regards to the error rates for overlapping sample terms, the terms with degrees of overlap of four and five (17 total terms) are the most complex in {*part of*} (and the entire BP), and it was expected to find relatively many issues with them, as seen in SNOMED [119].

The preliminary study focused on the core content of GO. However, GO on its own does not have many relationships and, thus, it has many large areas and partial-areas (e.g., {*part_of*}). One way of refining the granularity of GO’s taxonomies (see [120]) is to include equivalence axioms that reference other ontologies [137], such as ChEBI [138]. When the relationships to ChEBI, or to other ontologies, are considered in GO's taxonomy, there are many more areas and partial-areas (see [132]). Having a more refined summary of GO will likely enable the identification of more internal problems, in addition to errors in those external relationships.

A large portion of the BP taxonomy is the {*part_of*} area. GO contains many *part_of* relationships that play an integral role in the ontology. As with *is_a*'s, *part_of*'s are hierarchical. It would be useful to have a summary of GO's so-called "partonomy." In a future study the feasibility of deriving a taxonomy that summarizes GO's partonomy will be investigated.

3.4 Diff Abstraction Networks

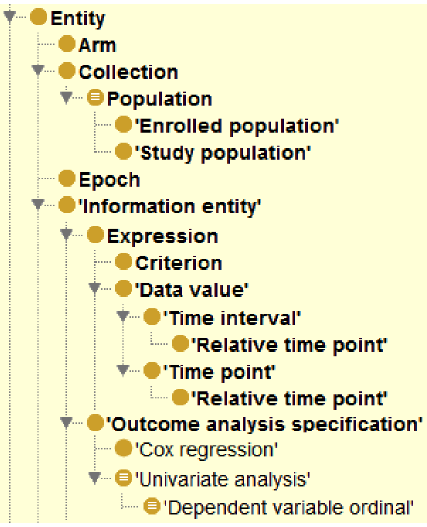
The structure of a biomedical ontology continually evolves as its content goes through cycles of editing, e.g., adding new domain-specific knowledge or importing additional knowledge from other ontologies. Classes, relationships, etc., are added, deleted, and updated. Each of these modifications affects the knowledge represented in the ontology.

A typical ontology will go through several stages of evolution. The early stage involves the initial design of the ontology, which may include importing one or more upper level ontologies, e.g., the Basic Formal Ontology (BFO) [46]. The later stages involve its maintenance, including periodic updates, which incorporate newly available knowledge into the ontology. During its evolution, an ontology may go through stages of quality assurance, where errors and inconsistencies are identified and corrected. During each of the various stages, the ontology goes through numerous release cycles, where changes are made from one release to the next. The problem is that while such changes are intended to extend the ontology’s knowledge or to correct previously discovered problems, they may have unintended, and potentially erroneous, consequences. In particular, a quality assurance phase *may introduce new errors*, while old errors are fixed. Such errors are typically not detected, due to the perception that the change is fulfilling its desired purpose. Sometimes, undesired changes may have broad effects, yet they still might go undetected because the curator “cannot see the forest for the trees.”

Not all editing operations affect an ontology in the same way. While adding a new leaf class will have no global impact, changing the domain of an object property may affect the definition of hundreds of classes. Similarly, modifying superclass axioms may lead to unintended object property inheritance. Having a global view of all of the changes that result from a series of editing operations is important for ontology maintenance and quality assurance. Ontology editing tools, such as Protégé [56], typically show an ontology as an indented hierarchy of classes. A curator can see only a few classes, or one class with its properties, at a time. It is difficult for a curator to identify the overall impact of an editing phase. To find all of the changes, a curator would have to check every potentially affected class, which is impractical for large ontologies.

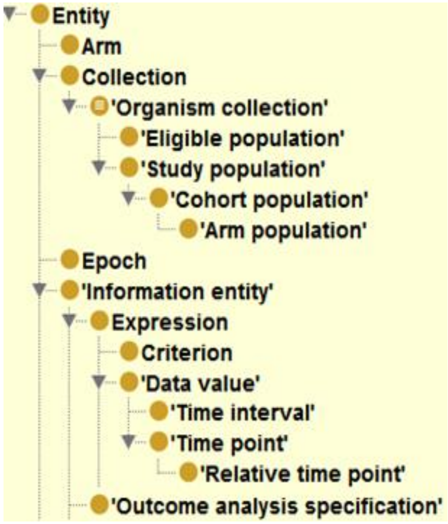
Figure 3.30 illustrates an indented hierarchy for an excerpt of 18 classes from the *Entity* hierarchy of the Ontology of Clinical Research (OCRe), Release 244 [51]. Figure 3.31 shows the same excerpt, from a later release. Clearly, a series of editing operations were applied between these two releases. While the hierarchical changes are easy to identify in this small example, it is not possible to see other changes, e.g., changes in object property inheritance. To identify unwanted changes, a curator would have to directly compare each version’s class definitions, which is a time-consuming process. If there are dozens or hundreds of classes in the ontology then this manual comparison process is not practical.

Whenever working with different versions of a document, whether it's a diagram, plain text, or an ontology, it is important to be able to identify changes between them. UNIX-based operating systems have the "diff" tool for this purpose [112]. For ontologies, the problem of identifying individual changes between two ontology versions has been extensively studied. PromptDiff [114], OWLDiff [91], and ContentCVS [115], among others, identify individual ontology changes in support of collaborative development and version control [113].



- Entity
  - Arm
  - Collection
    - Population
      - 'Enrolled population'
      - 'Study population'
  - Epoch
  - 'Information entity'
    - Expression
      - Criterion
      - 'Data value'
        - 'Time interval'
          - 'Relative time point'
        - 'Time point'
          - 'Relative time point'
    - 'Outcome analysis specification'
      - 'Cox regression'
      - 'Univariate analysis'
        - 'Dependent variable ordinal'

Figure 3.30 A subhierarchy of 18 classes taken from OCRe Version 244, as shown in Protégé.
Source: [121]



- Entity
  - Arm
  - Collection
    - 'Organism collection'
      - 'Eligible population'
      - 'Study population'
        - 'Cohort population'
          - 'Arm population'
  - Epoch
  - 'Information entity'
    - Expression
      - Criterion
      - 'Data value'
        - 'Time interval'
        - 'Time point'
          - 'Relative time point'
    - 'Outcome analysis specification'

Figure 3.31 The subhierarchy from Figure 3.30 after several editing operations have been applied to the classes. This excerpt is from OCRe Version 258.
Source: [121]

However, these tools show individual differences as a list or in an indented hierarchy. If there are hundreds of changes (both explicit and implicit) between two ontology versions, then the amount of difference information becomes overwhelming and unintended changes will remain undiscovered.

By summarizing, in a compact way, the changes that occur between any two releases, either consecutive or not, of an ontology it may be possible to detect unintended consequences of changes, due to the compact representation of the summary diff, and take steps to correct erroneous or undesired side effects of those changes.

To address this problem, it was necessary to create a new innovative structural diff technique called a Diff Abstraction Network ("Diff AbN") [121], for summarizing and visualizing differences between two versions of an ontology. A Diff AbN summarizes the difference in structure and content between two ontology releases. Unlike traditional ontology diff methods, which typically identify axiom changes for individual classes and properties, a Diff AbN shows the overall impact on the whole ontology, summarizing many explicit and implicit structural changes in a compact visualization. Thus, using a compact Diff AbN, an ontology curator can identify the global changes that result from her editing operations. By identifying unintended consequences of changes during the ontology development process, fewer errors will be introduced into the released ontology.

3.4.1 Derivation

Given two releases of an ontology, O_{from} and O_{to}, a *Diff Abstraction Network* ("Diff AbN") summarizes and visualizes, in a compact way, the global structural changes that occurred when moving from O_{from} to O_{to} due to editing operations. A Diff AbN supports the reflection of which structural changes occurred, and which classes in the ontology were affected by each change, by summarizing the changes that affect groups of structurally similar classes.

The derivation of two Diff AbNs, the Diff Area Taxonomy (DAT) and the Diff Partial-area Taxonomy (DPAT), will now be described in detail. A diff area taxonomy summarizes and visualizes the structural changes between O_{from} and O_{to}. A diff partial-area taxonomy refines the diff area taxonomy by summarizing and visualizing both structural and semantic changes to the subhierarchies of classes in each area. Object properties are an important structural feature used in the definition of many ontologies' classes [55], thus, it is important to identify the changes that occurred to the sets of object properties used to define the ontology's classes.

Various types of editing operations can alter the structure of an ontology, and thus, alter the area taxonomy and partial-area taxonomy derived from it. Any editing operation that affects object property introduction or inheritance for a set of classes will affect the taxonomies derived for the ontology. Some examples (labeled E1-E4) include: (E1) Adding or removing a class from an object property's domain; (E2) Adding or removing an object property from the ontology; (E3) Adding or removing a class from an ontology; (E4) Adding or removing a superclass axiom from a class. Multiple editing operations may be applied to a given class.

Previously ([25], Section 3.3.4.1) a quality assurance review of OCRe's *Entity* hierarchy was performed using a partial-area taxonomy. The quality assurance review identified errors in OCRe's modeling. To fix the identified errors, OCRe's curators made significant changes and a new version of OCRe was released. To illustrate the derivation of the diff taxonomies, an excerpt of classes from the version of OCRe that was reviewed for errors (Version 244, Figures 3.30 and 3.32) and the corresponding excerpt for the version released after all of the uncovered errors were corrected (Version 258, Figures 3.11 and 3.33).

Figure 3.32 illustrates the class hierarchy of Figure 3.30 and Figure 3.33 illustrates the corresponding class hierarchy of Figure 3.31, obtained from Figure 3.32 after several editing operations. Four classes have been removed from the hierarchy: *Population*, *Cox regression*, *Univariate analysis*, and *Dependent variable ordinal*. Three classes have been added: *Organism collection*, *Cohort population*, and *Arm population*. *Outcome analysis specification* was removed from the domain of two object properties and *Relative time point* is no longer a subclass of *Time interval*, thus it is no longer in the domain of *has start time* and *has stop time*. Note that these object property changes are not reflected in Figure 3.31.

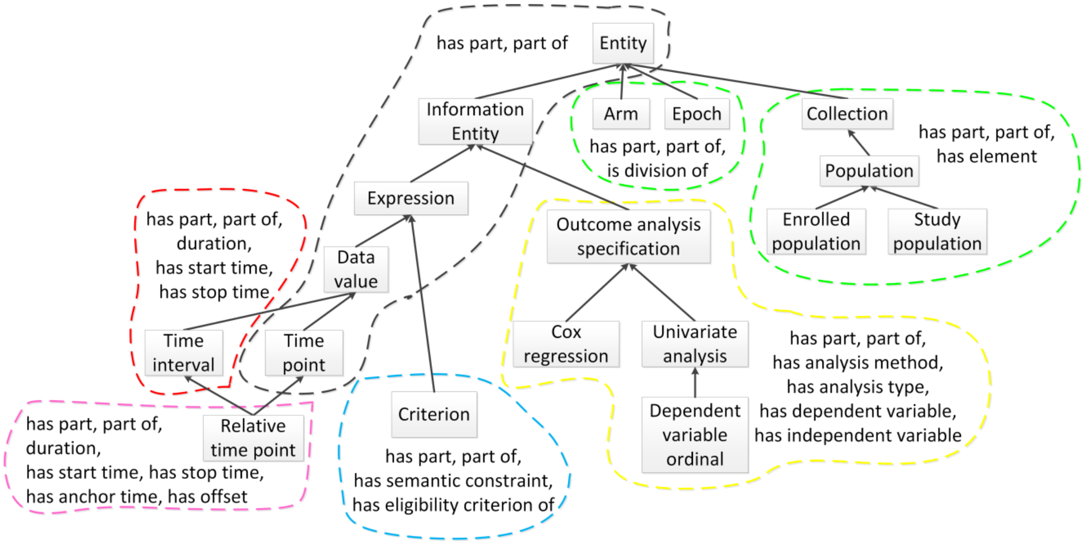

has part, part of
Entity
Information Entity
Arm
Epoch
has part, part of, is division of
Collection
has part, part of, has element
Population
Enrolled population
Study population
Expression
Outcome analysis specification
has part, part of, duration, has start time, has stop time
Data value
Time interval
Time point
Cox regression
Univariate analysis
has part, part of, has analysis method, has analysis type, has dependent variable, has independent variable
Dependent variable ordinal
Relative time point
has part, part of, duration, has start time, has stop time, has anchor time, has offset
Criterion
has part, part of, has semantic constraint, has eligibility criterion of

Figure 3.32 The excerpt of 18 classes from Figure 3.30 shown as a diagram, using bubbles to identify sets of classes with the same object properties.
Source: [121]

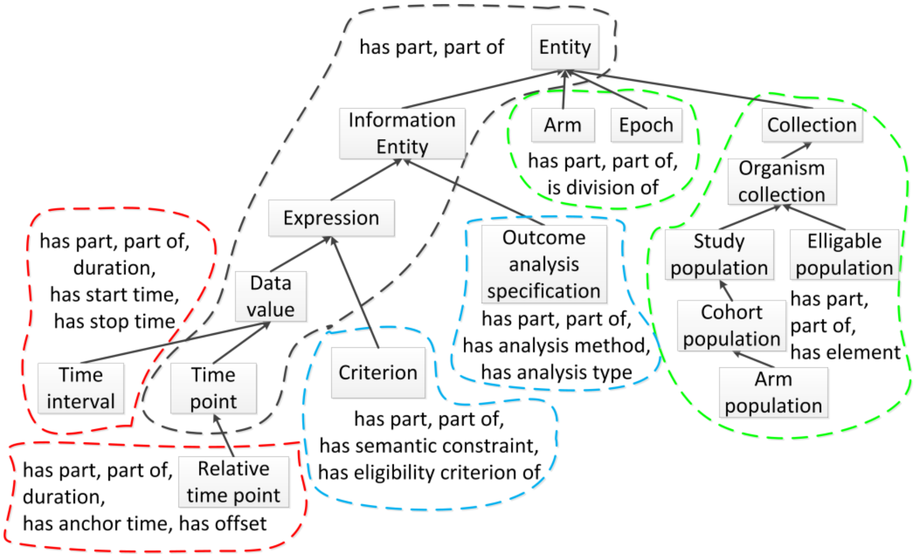

has part, part of
Entity
Information Entity
Arm
Epoch
Collection
has part, part of, is division of
Organism collection
Expression
Outcome analysis specification
Study population
Elligable population
has part, part of, duration, has start time, has stop time
Data value
Cohort population
has part, part of, has element
has part, part of, has analysis method, has analysis type
Time interval
Time point
Criterion
Arm population
has part, part of, has semantic constraint, has eligibility criterion of
has part, part of, duration, has anchor time, has offset
Relative time point

Figure 3.33 The excerpt of classes, after corrections (from Release 258) corresponding to the excerpt of Figure 3.31, shown as a diagram.
Source: [121]

Given two releases of an ontology, O_{from} and O_{to}, AT_{from} is defined as the area taxonomy derived for O_{from} and PAT_{from} as the partial-area taxonomy derived for O_{from}. AT_{to} and PAT_{to} are similarly defined for O_{to}.

3.4.1.1 Diff Area Taxonomy (DAT). A Diff Area Taxonomy (DAT) is an AbN that summarizes the structural changes between two different versions of an ontology (i.e., additions, deletions, and modifications to sets of classes with the same set of object properties). The input of a DAT consists of two ontologies O_{from} and O_{to} and the output consists of a compact, visual summary of the structural changes that occurred between O_{from} to O_{to}.

DAT derivation starts with identifying the set of object properties (both introduced and inherited) used to define each class in O_{from} and O_{to}. Classes and object properties that are added or removed between O_{from} and O_{to} are also identified. The sets of object properties used to define each class in O_{from} and O_{to} are then compared. This process is equivalent to comparing the areas in AT_{from} to the areas in AT_{to}. Four kinds of *Diff Areas* are created based on the identified differences, as follows. These diff areas are used to summarize the structural changes that occurred between O_{from} and O_{to}.

(a) An *Introduced Area* is defined as an area that exists in AT_{to} but does not exist in AT_{from}. An introduced area indicates a set of object properties for which there exists a set of one or more classes in O_{to} but no such class exists in O_{from}. The classes summarized by an introduced area display a new object property structure in the ontology. An introduced area may summarize a set of classes that were previously summarized by different area(s) in AT_{from}, or they are newly added classes, or both.

(b) A *Removed Area* is an area that exists in AT_{from} but does not exist in AT_{to}. A removed area indicates a particular set of object properties for which a non-empty set of classes exists in O_{from} but no such class exists in O_{to}. The classes that were previously summarized by a removed area are now either summarized by a different area in AT_{to} or were removed from the ontology.

(c) The third kind is a *Modified Area*. Such an area *A* does exist in both AT_{from} and AT_{to}, meaning in both versions of the ontology there is a set of object properties for which a set of classes exists (though the set is not the same and one set is not necessarily a subset of the other). If the set of classes summarized by the area *A* in AT_{from} is different from the set of classes summarized by *A* in AT_{to}, then *A* is said to be a modified area. Classes that were originally summarized by *A* in AT_{from} may be summarized by different areas in AT_{to} if their object property sets changed, or the classes may have been removed from the ontology entirely. Similarly, a class may become summarized by *A* in AT_{to} if its object property set changed to match that of *A* or if a class was added to the ontology with *A*'s object property set.

(d) If the set of classes summarized by an area *A* is the same in AT_{from} and AT_{to} then *A* is an *Unmodified Area*. This indicates that no changes occurred to the object property set for the classes in *A* between AT_{from} and AT_{to}.

In regards to the *child-of* links between areas that summarize the class hierarchy, a *child-of* is called an *introduced child-of* if it exists between two areas in AT_{to} but not in AT_{from}. Similarly a *child-of* is called a *removed child-of* if it exists between two areas in AT_{from} but not in AT_{to}. A *child-of* is an *unmodified child-of* if it exists between the same two areas in AT_{from} and in AT_{to}. Additionally, the following rules are defined: (1) All of the *child-of* links sourced at an introduced area are *introduced child-of*s; (2) All of the *child-of* links sourced from a removed area are *removed child-ofs*. It is noted that modified areas *and* unmodified areas may have *introduced*, *removed*, or *unmodified child-of*s. Note that *child-of* links cannot be *modified* because a *child-of* link either existed or did not exist in AT_{from}.

A DAT is represented as a compact network of diff area nodes connected by *child-of* links based on the subclass hierarchies in O_{from} and O_{to}. In a DAT, all areas are shown, including removed areas which summarize no classes in AT_{to}. In a DAT visualization diff areas are shown with differed colored borders to indicate the type of diff area. Modified areas are drawn with a yellow border, introduced areas with a green border, and removed areas with a red border. Unmodified areas are shown with no border. *Child-of* links are colored red if they were removed, green if they were introduced, or black if they are unmodified. (*Child-of* links cannot be modified.)

If the number of classes summarized by an area changes between O_{from} and O_{to}, e.g. the area {*has part, part of, has element*} summarizes four classes in AT_{from} but six in AT_{to}, then the change is noted using an arrow from the old number to the new number (i.e., 4 Classes → 6 Classes). A brief textual summary of the modifications to the area is shown under the number of classes summarized by the diff area. For example, the diff area {*has part, part of, has element*} indicates that one class was removed from the ontology ("-1 Class Removed") and three classes were added ("+3 New Classes") (see right green box in Figure 3.34).

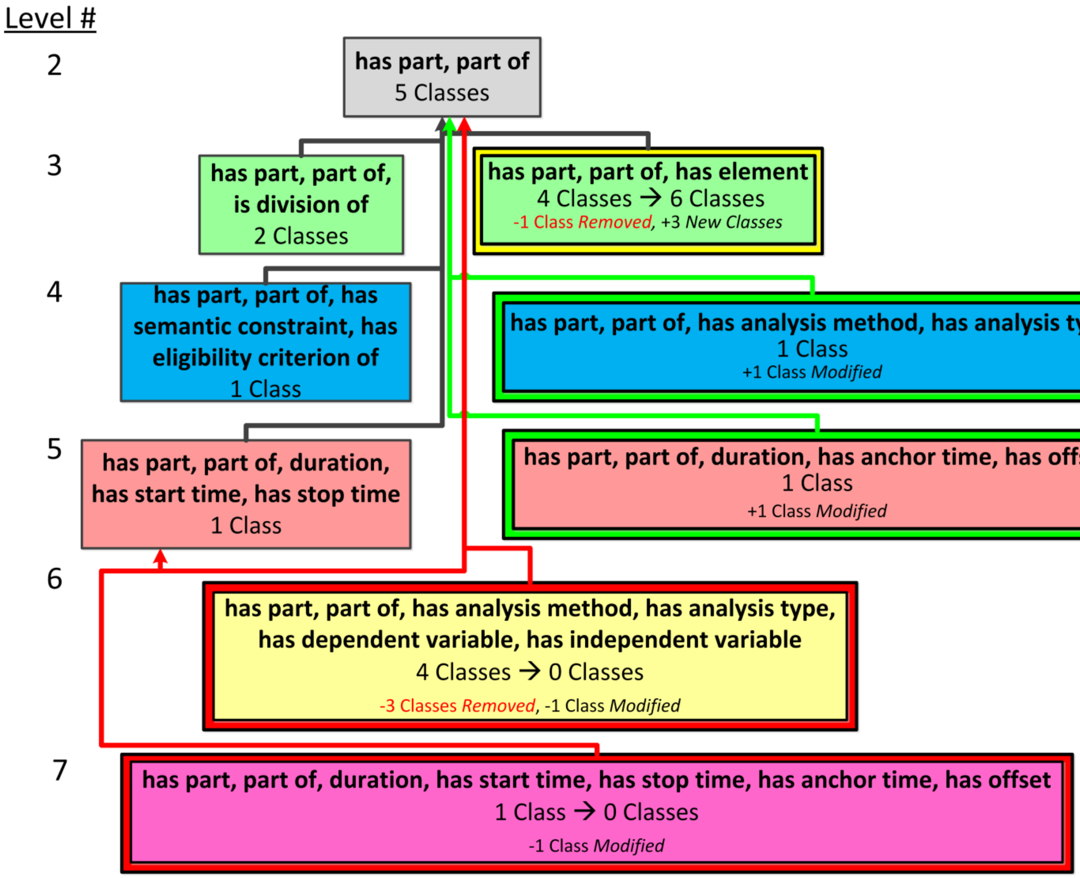

Level #
2
3
4
5
6
7
has part, part of
5 Classes
has part, part of, is division of
2 Classes
has part, part of, has element
4 Classes → 6 Classes
-1 Class Removed, +3 New Classes
has part, part of, has semantic constraint, has eligibility criterion of
1 Class
has part, part of, has analysis method, has analysis type
1 Class
+1 Class Modified
has part, part of, duration, has start time, has stop time
1 Class
has part, part of, duration, has anchor time, has offset
1 Class
+1 Class Modified
has part, part of, has analysis method, has analysis type, has dependent variable, has independent variable
4 Classes → 0 Classes
-3 Classes Removed, -1 Class Modified
has part, part of, duration, has start time, has stop time, has anchor time, has offset
1 Class → 0 Classes
-1 Class Modified

Figure 3.34 The visualization of the diff area taxonomy between the ontology excerpts in Figure 3.32 and Figure 3.33. The diff areas are organized into color coded levels according to the number of their object properties. The level numbers appear at the left edge of the figure.
Source: [121]

Ontology editing operations have various effects. For example, removing the superclass axiom (E4) between *Relative time point* and *Time interval* resulted in *Relative time point* being summarized by a different area, {*has part, part of, duration, has anchor time, has offset*} (Level 4) in AT_{to}. The OCRe DAT, shown in Figure 3.34, captures the structural changes from the ontology excerpt of Figure 3.32 to the excerpt in Figure 3.33.

The diff areas {*has part, part of, has analysis method, has analysis type*} (Level 4) and {*has part, part of, duration, has anchor time, has offset*} (Level 5) are introduced areas marked with a green border; they exist in AT_{to} but did not exist in AT_{from}. In this example, the introduced areas in Figure 3.34 summarize classes that were summarized by different areas in AT_{from}. This indicates a change in the object property structure of these classes (due to E1and E4, respectively) and they are now defined differently.

In Figure 3.34, the single diff area on Level 6 and the single diff area on Level 7 are removed areas, as indicated by their red borders; these areas existed in AT_{from} but no longer exist in AT_{to}. It is important to display the removed areas in the DAT figure, even though these areas no longer exist in AT_{to}, to capture the important change(s) that resulted in their removal. For example, several editing operations led to the yellow Level 6 area being removed: three classes (*e.g., Cox regression*) were removed from the *Entity* hierarchy (E3) and the class *Outcome analysis specification* is summarized by a different area, {*has part, part of, has analysis method, has analysis type*} (Level 4), in AT_{to} (E1).

In Figure 3.34 {*has part, part of, has element*} is a modified area (with a yellow border), because the class *Population* was removed from the ontology (E3) and three new classes, *Organism collection*, *Cohort population*, and *Arm population,* with the modified area's object property set, were added to the ontology (also E3). The new classes inherited their object property set, because they are descendants of *Collection* and they introduce no new object properties to the subhierarchy. The unmodified areas are {*has part, part of*}, {*has part, part of, is division of*}, {*has part, part of, has semantic constraint, has eligibility criterion*}, and {*has part, part of, duration, has start time, has stop time*}.

3.4.1.2 Diff Partial-area Taxonomy (DPAT). In previous studies the partial-area taxonomy has been used to support QA of ontologies [25, 55, 120, 139]. A Diff Partial-area Taxonomy (DPAT) summarizes the changes to the subhierarchies of classes in each

DAT area. Just as a partial-area taxonomy is a refinement of an area taxonomy into partial-areas (i.e., semantically similar subgroups within the structurally similar area groups), a DPAT refines a DAT by summarizing subhierarchy changes, represented as changes to the partial-areas in each area.

The derivation of the DPAT starts from the already derived DAT. For each diff area *A* in the DAT, the changes to the subhierarchies of classes in *A*, as named after the roots, are summarized. The set of root classes of *A* in AT_{from} is compared to the set of root classes of *A* in AT_{to}, in cases where *A* exists in both. If the two sets are not equal, this indicates that partial-areas have been added or removed from the area. Based on the identified changes, four kinds of *Diff Partial-areas* are created.

(a) An *Introduced Partial-area* is a partial-area that exists in area *A* in PAT_{to} but did not exist in *A* in PAT_{from}. A partial-area is introduced to an area *A* whenever a root class is added to *A*. Partial-areas can be introduced to any diff area that is not a removed area. All partial-areas in an introduced area are by definition introduced partial-areas.

(b) A *Removed Partial-area* is a partial-area that exists in area *A* in PAT_{from} but not in *A* in PAT_{to}. A partial-area is removed from an area whenever a root class is removed from *A*. Partial-areas can be removed from any diff area that is not an introduced area. All partial-areas in a removed area are by definition removed partial-areas.

(c) If area *A* has one or more of the same root classes in both PAT_{from} and PAT_{to} then the subhierarchies of classes from both versions are compared. A *Modified Partial-area* is a partial-area that exists in *A* in both PAT_{from} and PAT_{to} and summarizes a different set of classes in PAT_{to} than in PAT_{from}.

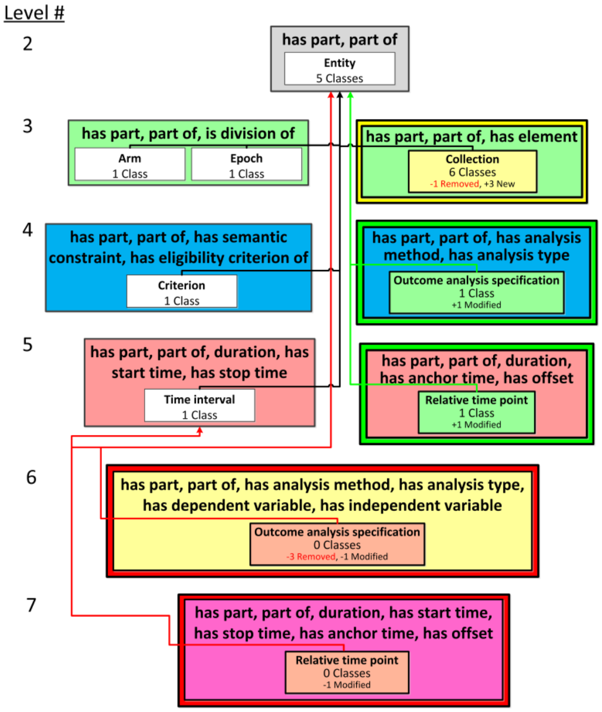
(d) An *Unmodified Partial-area* is a partial-area that summarizes the same set of classes in area *A* in PAT_{from} and in area *A* in PAT_{to}.

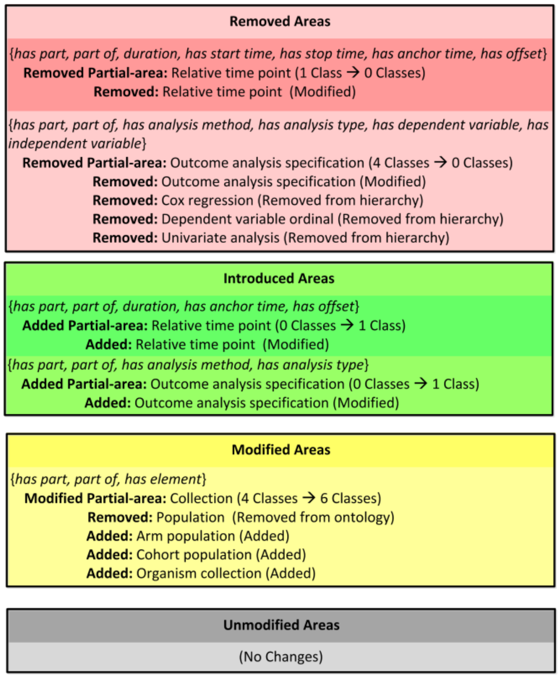
It is noted that an unmodified area can contain modified, introduced, and removed partial-areas. This occurs when the set of classes summarized by the unmodified area remains the same between O_{from} and O_{to} but the subhierarchies of classes change within the diff area. For example, if a descendant of a root class in *A* is made a sibling of the root class then a partial-area is introduced within the unmodified area. Similarly, if a class is summarized by two partial-areas in PAT_{from} (which are, thus, not disjoint) but only one partial-area in PAT_{to}, the diff area can still be unmodified. The definition of *child-ofs* between diff partial-areas follows that of the *child-ofs* between diff areas.

Like the DAT, the DPAT consists of a visualization and a textual list of differences. The visualization of a DPAT is composed of a refined DAT visualization where the DPAT partial-areas are shown within their respective DAT areas. Modified partial-areas are shown with a light yellow background, introduced partial-areas with a light green background, and removed partial-areas with a light red background. A summary of changes is shown below the number of classes summarized by each partial-area. Unmodified partial-areas are shown with a white background. *Child-of* links between partial-areas are drawn red if they were removed, green if introduced, and black if unmodified. Figure 3.35 shows the visualization of the DPAT capturing the changes from the ontology version shown in Figure 3.32 to the new version in Figure 3.33.

The text output of a DPAT is composed of changes grouped by area change type (e.g., removed or modified area). Within each type, the list of affected areas is shown. Indented under each area is a list of modifications to the partial-areas within the area.

The modifications to the set of classes summarized by each partial-area are listed indented under the partial-area root (which is its name). Figure 3.36 shows a colored example of text-based output for the DPAT between the ontologies in Figures 3.32 and 3.33. The background color alternates between brighter and darker shades, in order to visually separate different areas.


Level #
2
3
4
5
6
7
has part, part of
Entity
5 Classes
has part, part of, is division of
Arm
1 Class
Epoch
1 Class
has part, part of, has element
Collection
6 Classes
-1 Removed, +3 New
has part, part of, has semantic constraint, has eligibility criterion of
Criterion
1 Class
has part, part of, has analysis method, has analysis type
Outcome analysis specification
1 Class
+1 Modified
has part, part of, duration, has start time, has stop time
Time interval
1 Class
has part, part of, duration, has anchor time, has offset
Relative time point
1 Class
+1 Modified
has part, part of, has analysis method, has analysis type, has dependent variable, has independent variable
Outcome analysis specification
0 Classes
-3 Removed, -1 Modified
has part, part of, duration, has start time, has stop time, has anchor time, has offset
Relative time point
0 Classes
-1 Modified


Removed Areas

{has part, part of, duration, has start time, has stop time, has anchor time, has offset}
Removed Partial-area: Relative time point (1 Class → 0 Classes)
Removed: Relative time point (Modified)

{has part, part of, has analysis method, has analysis type, has dependent variable, has independent variable}
Removed Partial-area: Outcome analysis specification (4 Classes → 0 Classes)
Removed: Outcome analysis specification (Modified)
Removed: Cox regression (Removed from hierarchy)
Removed: Dependent variable ordinal (Removed from hierarchy)
Removed: Univariate analysis (Removed from hierarchy)

Introduced Areas

{has part, part of, duration, has anchor time, has offset}
Added Partial-area: Relative time point (0 Classes → 1 Class)
Added: Relative time point (Modified)
{has part, part of, has analysis method, has analysis type}
Added Partial-area: Outcome analysis specification (0 Classes → 1 Class)
Added: Outcome analysis specification (Modified)

Modified Areas

{has part, part of, has element}
Modified Partial-area: Collection (4 Classes → 6 Classes)
Removed: Population (Removed from ontology)
Added: Arm population (Added)
Added: Cohort population (Added)
Added: Organism collection (Added)

Unmodified Areas

(No Changes)

Figure 3.35 The visualization of the DPAT between Figure 3.32 and Figure 3.33.
Source: [121]

Figure 3.36 Color-coded text output for the DPAT between Figure 3.32 and Figure 3.33.
Source: [121]

The text-based output of the DPAT is designed to be used in conjunction with the DPAT visualization, enabling a curator to see more details about the modification of each affected taxonomic element.

In Figure 3.35, the introduced partial-area *Outcome analysis specification* appears in the area {*has part, part of, has analysis method, has analysis type*} (Level 4) and the introduced partial-area *Relative time point* in the area {*has part, part of, duration, has anchor time, has offset*} (Level 5). Both of these diff areas are introduced areas, as indicated by their green borders. Note that the green, red, and yellow colors of the areas in levels 3, 5, and 6, respectively, do not communicate changes to the areas, but are the colors of the different levels. At the same time, *Outcome analysis specification* and *Relative time point* are removed partial-areas in the removed areas of Levels 6 and 7.

Occurrences of identically named introduced and removed partial-areas reflect the changes in object properties of the root class in the DPAT. In both of these cases the classes were removed from the domains of the object properties as a result of the errors discovered by Ochs et al. [25].

Collection in {*has part, part of, has element*} is a modified partial-area because one class was removed from the ontology and three new classes were added to the ontology as descendants of *Collection*. *Entity, Arm, Epoch, Criterion, and Time interval* are unmodified partial-areas.

3.4.2 DPAT-based Quality Assurance Methodology

While the diff partial-area taxonomy does not automatically identify erroneous modeling, it does highlight groups of classes that should be reviewed after one or more editing operations were applied to the ontology. For example, if the domain of an object property is changed, then an ontology curator should review the classes in the added and removed partial-areas to ensure they have the correct sets of object properties. Similarly, if a subclass relationship is established or removed between two classes, then the curator should review all of the diff partial-areas that contain the descendants of the modified class to ensure that the inheritance of object properties is still correct.

It is expected that different kinds of DPATs will appear for different ontology development stages. For example, if an ontology is going through a phase of expansion, i.e., new knowledge is being added, then the DPAT will likely contain many introduced and modified areas and partial-areas. When an ontology is going through a QA phase, there may be relatively more removed areas and removed partial-areas than in an expansion phase.

3.4.3 Application of Diff Partial-area Taxonomies

To test the Diff AbN approach to QA, diff partial-area taxonomies were derived for three ontologies: the Ontology of Clinical Research (OCRe) [51], the Sleep Domain Ontology (SDO) [45], and the eagle-I Research Resource Ontology (ERO) [140]. The information provided by the DPAT was compared to a standard ontology diff created using the "Compare Ontologies" feature in Protégé [56], which is based on the OWL Difference Engine [117].

OCRe and SDO were chosen because of the previously performed QA reviews of their content [25, 120]. Several errors and inconsistencies were confirmed and corrected during these QA reviews. In both cases, taxonomies were derived before and after QA reviews [25, 120] and were manually compared. For OCRe and SDO, DPATs were derived using the ontology release before the QA review and the ontology release immediately after the errors uncovered during the QA review were corrected. Details of the errors found are described by Ochs et al. [25, 120]. The goal was to determine if, using the diff partial-area taxonomies of OCRe and of SDO, whether these ontologies were corrected as expected, or whether some unintended and erroneous changes were introduced.

ERO was chosen because it was recently merged [141] with the VIVO Ontology for Researcher Discovery (VIVO) [142]. A DPAT was derived using the August 2013 ERO version available on the NCBO BioPortal (before the merge), and the version after the merge was completed.

3.4.3.1 Ontology of Clinical Research DPAT. The quality assurance review of OCRe's inferred *Entity* hierarchy identified several modeling errors [25]. Two examples include the erroneous inclusion of 33 statistical classes due to incorrect domains for the object properties *has dependent variable* and *has independent variable* and an erroneous subclass relationship between *Relative time point* and *Time interval*. The DPAT in Figure 3.37 captures the structural changes that occurred due to the corrections implemented by OCRe's curator, Samson Tu. The complete DPAT has two modified partial-areas, three deleted partial-areas, and three added partial-areas, summarizing the changes to 32 classes (see diff areas with yellow, red, and green borders). Eighteen partial-areas are unmodified.

Following the methodology described in Section 3.4.2, one should review the added and removed areas and partial-areas in the DPAT to determine if their classes have the correct sets of object properties. In Figure 3.37, one finds the introduced partial-area *Relative time point* in the introduced area {*has part, part of, duration, has anchor time, has offset*}. This diff area and this diff partial-area were introduced due to the removal of an incorrect subclass relationship to *Time interval* [25], which corrected the erroneous inheritance of two object properties (*has start time, has stop time*) by *Relative time point*.

After reviewing this introduced area in the DPAT, it was found that *Relative time point* had another incorrect object property: *duration*, since a time point has no duration. Indeed, this object property was determined to be redundant with *has offset*. When correcting the *Relative time point* class, the domain of *duration* was changed from only *Time interval* to *Time interval* **or** *Relative time point*, due to the removal of the subclass relationship between *Relative time point* and *Time interval*. Hence, the *duration* object property was no longer inherited by the class *Relative time point*.

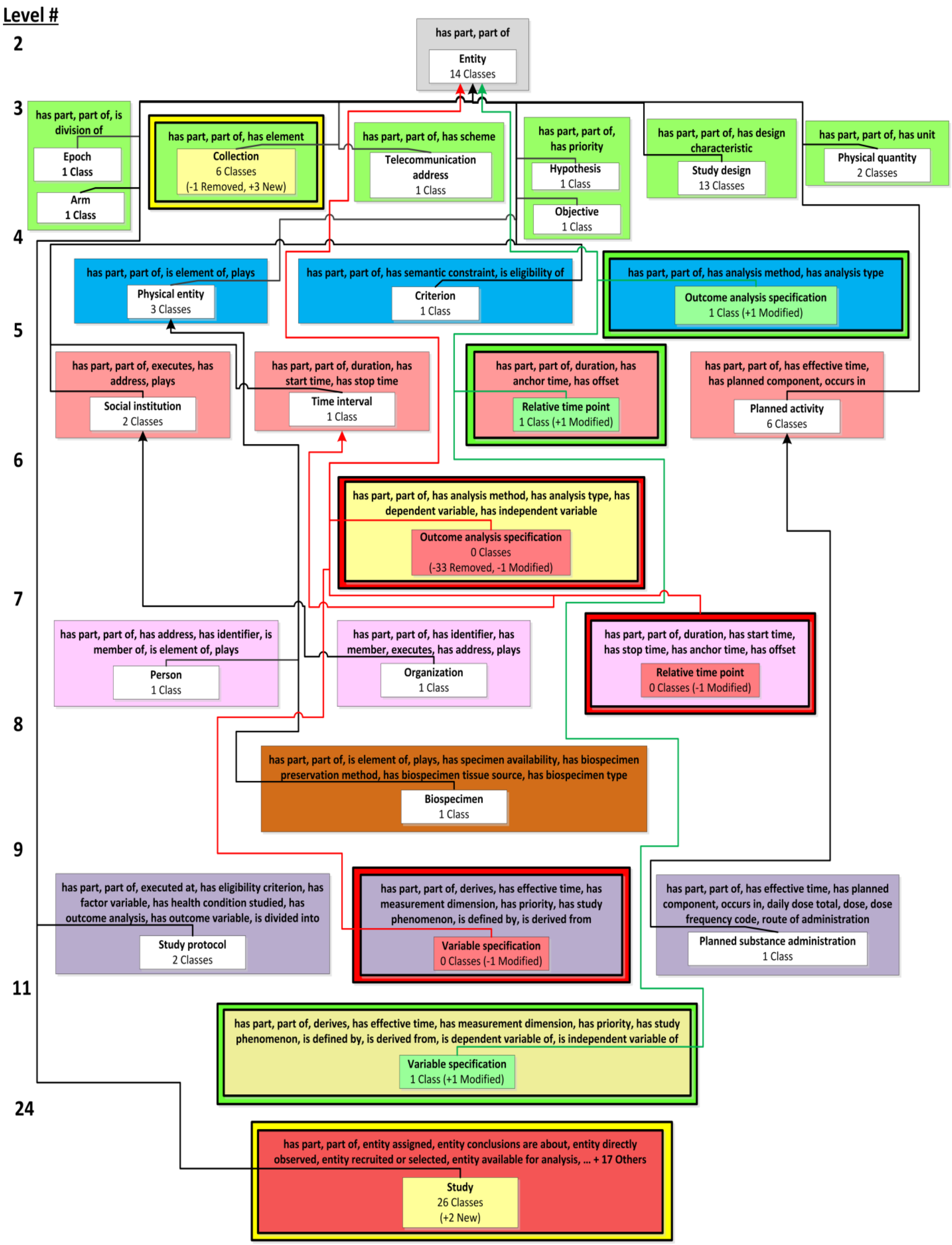

Level #
2
has part, part of
Entity
14 Classes
3
has part, part of, is division of
Epoch
1 Class
Arm
1 Class
has part, part of, has element
Collection
6 Classes
(-1 Removed, +3 New)
has part, part of, has scheme
Telecommunication address
1 Class
has part, part of, has priority
Hypothesis
1 Class
Objective
1 Class
has part, part of, has design characteristic
Study design
13 Classes
has part, part of, has unit
Physical quantity
2 Classes
4
has part, part of, is element of, plays
Physical entity
3 Classes
has part, part of, has semantic constraint, is eligibility of
Criterion
1 Class
has part, part of, has analysis method, has analysis type
Outcome analysis specification
1 Class (+1 Modified)
5
has part, part of, executes, has address, plays
Social institution
2 Classes
has part, part of, duration, has start time, has stop time
Time interval
1 Class
has part, part of, duration, has anchor time, has offset
Relative time point
1 Class (+1 Modified)
has part, part of, has effective time, has planned component, occurs in
Planned activity
6 Classes
6
has part, part of, has analysis method, has analysis type, has dependent variable, has independent variable
Outcome analysis specification
0 Classes
(-33 Removed, -1 Modified)
7
has part, part of, has address, has identifier, is member of, is element of, plays
Person
1 Class
has part, part of, has identifier, has member, executes, has address, plays
Organization
1 Class
has part, part of, duration, has start time, has stop time, has anchor time, has offset
Relative time point
0 Classes (-1 Modified)
8
has part, part of, is element of, plays, has specimen availability, has biospecimen preservation method, has biospecimen tissue source, has biospecimen type
Biospecimen
1 Class
9
has part, part of, executed at, has eligibility criterion, has factor variable, has health condition studied, has outcome analysis, has outcome variable, is divided into
Study protocol
2 Classes
has part, part of, derives, has effective time, has measurement dimension, has priority, has study phenomenon, is defined by, is derived from
Variable specification
0 Classes (-1 Modified)
has part, part of, has effective time, has planned component, occurs in, daily dose total, dose, dose frequency code, route of administration
Planned substance administration
1 Class
11
has part, part of, derives, has effective time, has measurement dimension, has priority, has study phenomenon, is defined by, is derived from, is dependent variable of, is independent variable of
Variable specification
1 Class (+1 Modified)
24
has part, part of, entity assigned, entity conclusions are about, entity directly observed, entity recruited or selected, entity available for analysis, … + 17 Others
Study
26 Classes
(+2 New)

Figure 3.37 The complete diff partial-area taxonomy for OCRe.

Source: [121]

Upon investigation, it was found that *Duration* was previously used to express offsets for relative time points but this should have changed when the object property *has offset* was introduced to the ontology. Samson Tu confirmed the error and *Relative time point* was removed from the domain of the *duration* object property.

3.4.3.2 Sleep Domain Ontology DPAT. In Ochs et al. [120] a preliminary QA review of the Sleep Domain Ontology (SDO)'s *Entity* hierarchy was performed, together with the curator of the SDO, Sivaram Arabandi. The partial-area taxonomy for the hierarchy was reviewed and several modeling errors were identified, e.g., duplicate classes and incorrectly assigned object property domains. Correcting the errors led to significant structural changes in the SDO. While a relatively small number of axioms were edited to fix the errors, hundreds of classes were implicitly modified due to these changes. Sivaram Arabandi was surprised at the extent of modifications to the partial-area taxonomy and could not obtain an adequate display, focusing on those changes, by using the diff view provided in Protégé [56]**.**

During the audit of the SDO [120], two pairs of duplicate classes were identified, two *clinical finding* classes (both imported, one from OGMS [47] and the other from BioTop [48]) and the classes *clinical diagnosis* and *diagnosis*. To remove the duplicate classes, equivalence was established between the classes of each pair. This resulted in many classes' object property sets changing, as captured by the 25 removed areas, 25 introduced areas, and four modified areas (along with all of their diff partial-areas) in the SDO DPAT in Figure 3.38.

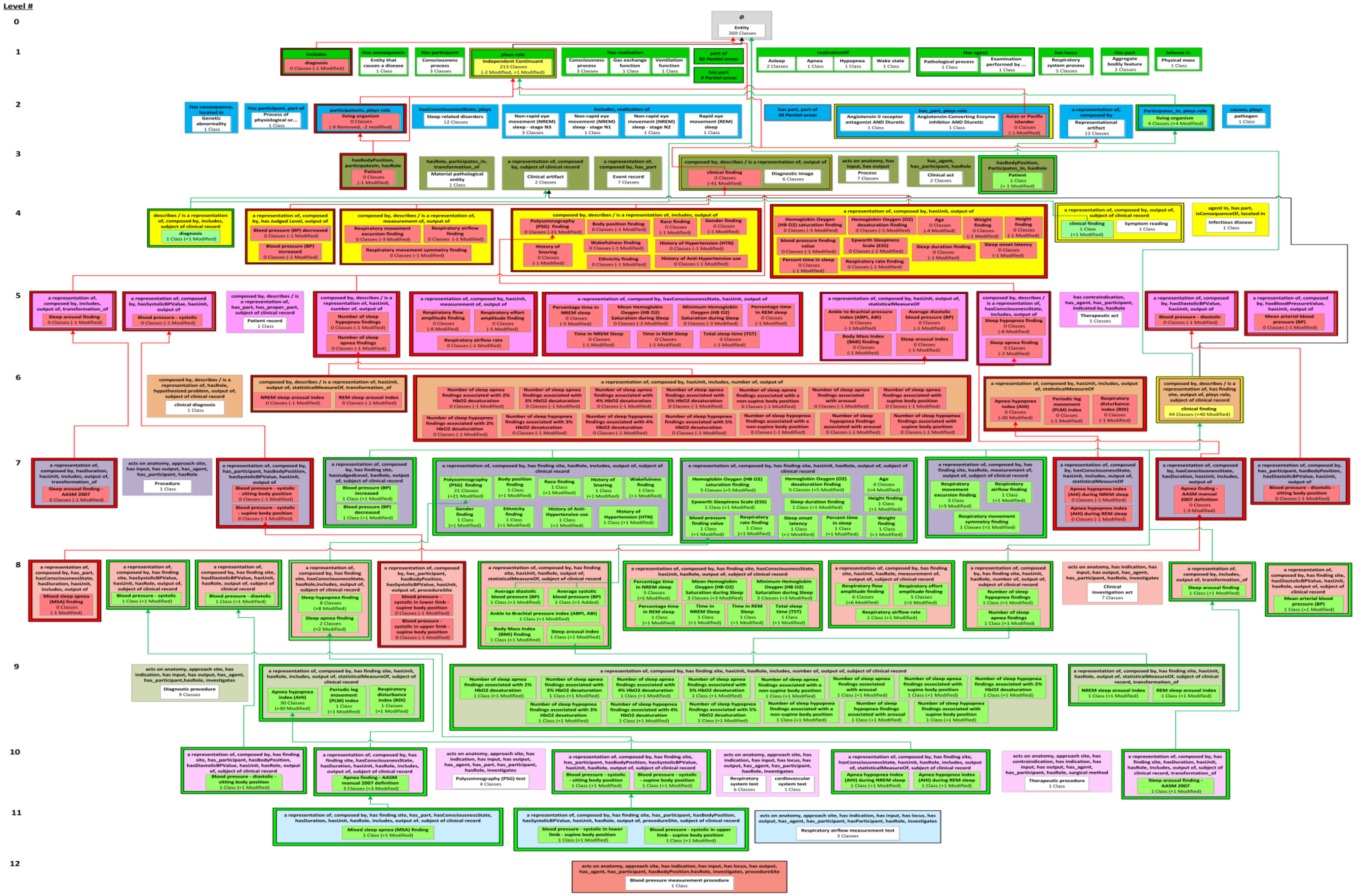


Figure 3.38 The Sleep Domain Ontology's diff partial-area taxonomy. An excerpt of the *child-of* links between added/removed diff partial-areas is shown. *Source: [121]*

When there is this much structural change, in terms of the sets of object properties used to define an ontology's classes, between two releases of an ontology, there is a greater chance of a class being assigned an incorrect object property set.

By reviewing the introduced partial-areas in the SDO's DPAT, several problems with the object properties for the equivalent classes were identified. Even though the *clinical finding* partial-area on Level 3 was (correctly) removed and 42 of its classes are now summarized by the *clinical finding* modified partial-area in the Level 6 modified area {*a representation of*, *composed by*, ***has finding site***, ***hasRole***, *output of*, *subject of clinical record*}, an introduced partial-area *clinical finding* (with one class) on Level 4 in the modified area {*a representation of*, *composed by*, *output of*, *subject of clinical record*} was found. Similarly, *diagnosis* is introduced at Level 4 (and removed from Level 1) in {*composed by*, *describes / is a representation of*, ***includes***, *subject of clinical record*} (the object properties in bold are extra).

However, the equivalent class *clinical diagnosis* is in {*composed by*, *describes / is a representation of*, ***hasRole, hypothesized problem, output of***, *subject of clinical record*}. The object properties for equivalent classes should be equivalent. However, as shown in bold, they are not. For *diagnosis*, one is not even a subset of the other. By reviewing the added and removed partial-areas that contain the classes that were edited, several inconsistencies were identified. Both equivalent classes should have the union of the two sets of object properties, as confirmed by Sivaram Arabandi

3.4.3.3 eagle-I Research Resource Ontology DPAT. The eagle-i Research Resource Ontology (ERO) [143] was developed as part of the eagle-i project [144], which enables biomedical researchers to discover scientific resources via a searchable network of resource repositories. These repositories are curated by over 20 different research institutions [144]. Like the SDO, ERO imports the content of several external ontologies, including BFO and OCRe. However, ERO differs from OCRe and SDO in that it is used to drive applications for data entry and search. ERO is composed of several modules. Notably, the representation of research resource data is in a separate module from the representation of application specific data used to control the appearance and behavior of the user interface. Many of ERO's classes and properties in the application module were designed to drive eagle-i''s user interface and the various data collection tools used in the eagle-i project. ERO is composed of several modules.

Unlike OCRe and SDO, which had a relatively small number of local editing operations applied to correct modeling errors uncovered during quality assurance reviews, ERO underwent a significantly more complex sequence of editing operations. ERO was recently merged [141] with the VIVO ontology [145] which covers the orthogonal but overlapping domain of researcher interests, activities and accomplishments. A DPAT was derived for the version of ERO before the merge (August 2013 release on BioPortal) and the version after the merge (available at [146]), with the goal of summarizing the major structural changes that occurred due to the merge.

The ERO DPAT, which has 26 levels, is shown in Figure 3.39 and 3.40. *Child-of* links from diff partial-areas in Figure 3.40 that have a parent diff partial-area in Figure 3.39 are not shown. The structural changes resulting from the merge are summarized by the 57 introduced areas, 48 removed areas, and one modified area (the root area) of ERO's DPAT. Like OCRe, most of ERO's areas are singly-rooted, meaning there is only one partial-area in most areas.

The most significant structural change highlighted by ERO's DPAT is the highly desirable overall reduction in complexity, in terms of number of object properties used to define ERO's classes. This change is reflected in the large number of removed areas at the bottom of the DPAT (Figure 3.40) (since the levels are listed in increasing order according to the number of object properties of the areas) and the large number of introduced areas at the top (Figure 3.39). For example, before the merge, the class c*ell line* (Level 24 in Figure 3.40) and its six descendants were in the domain of 24 object properties (16 by inheritance from *reagent* (1), eight introduced explicitly at *cell line*). After the merge, *cell line* (the rightmost introduced partial-area in the top level in Figure 10) and its descendants are in the domain of only 12 object properties (seven inherited, five introduced), as reflected by the introduced partial-area *cell line* (7), the rightmost partial-area in Level 12 of Figure 3.40.

A combination of changes led to this reduction in complexity. First, the ontology's set of object properties was significantly changed. A total of 36 object properties were removed and 91 object properties were added. This affected many classes. For example, *cell line* was implicitly in the domain of the object property *agent in*, whose domain was defined as *continuant.* The object property *agent in* was removed from the ontology. Some of these removals happened because a newer version of the Relations Ontology (RO) [147] was imported.

Prior to the merge, eight object properties imported from RO had *continuant* assigned as a domain. All of these object properties are no longer in the ontology after the merge. Thus, the many descendants of *continuant* are no longer implicitly in their domains.

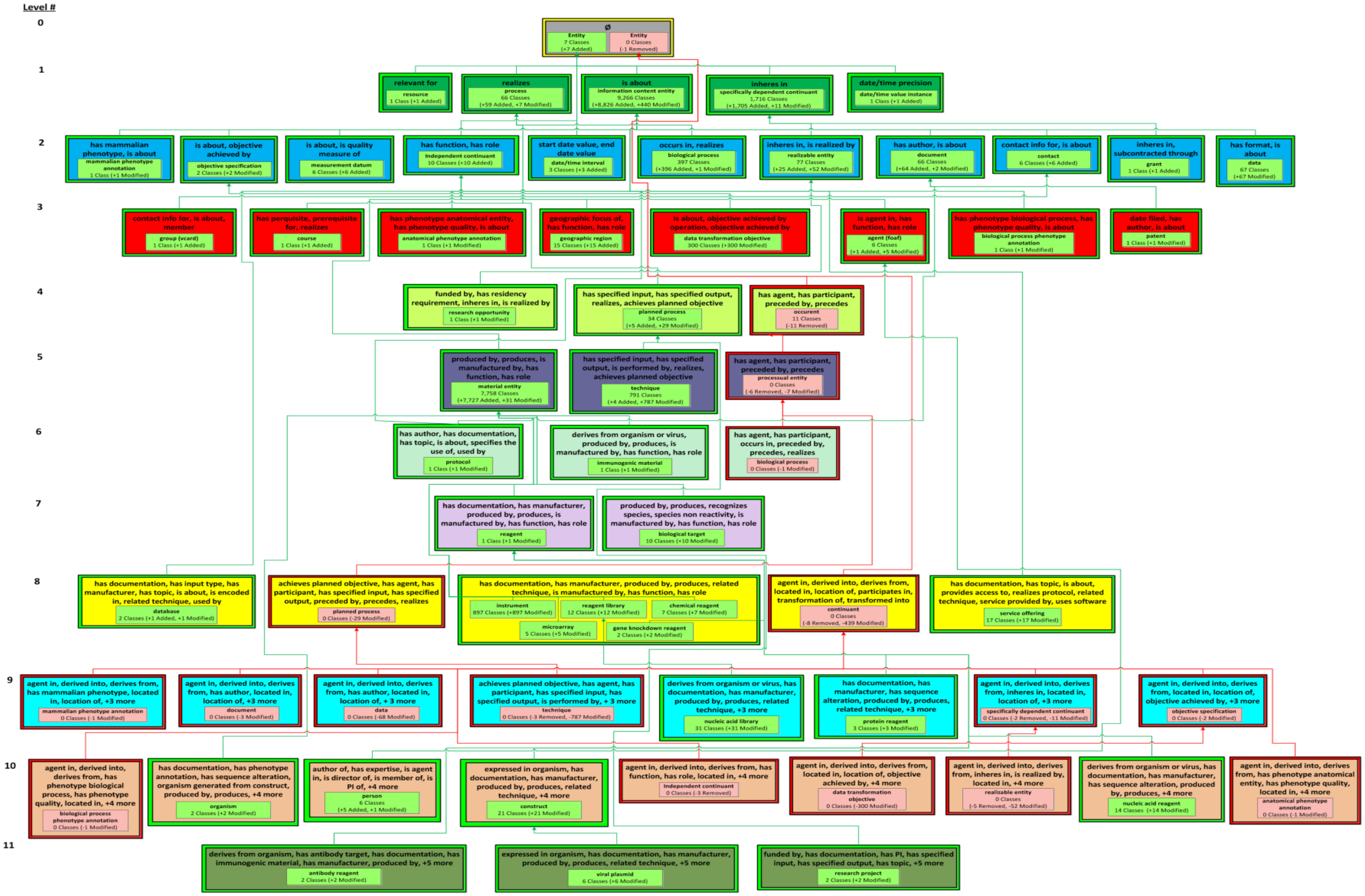

Level #
0
1
2
3
4
5
6
7
8
9
10
11
ø
Entity
7 Classes
(+7 Added)
Entity
0 Classes
(-1 Removed)
relevant for
resource
1 Class (+1 Added)
realizes
process
66 Classes
(+59 Added, +7 Modified)
is about
information content entity
9,266 Classes
(+8,826 Added, +440 Modified)
inheres in
specifically dependent continuant
1,716 Classes
(+1,705 Added, +11 Modified)
date/time precision
date/time value instance
1 Class (+1 Added)
has mammalian phenotype, is about
mammalian phenotype annotation
1 Class (+1 Modified)
is about, objective achieved by
objective specification
2 Classes (+2 Modified)
is about, is quality measure of
measurement datum
6 Classes (+6 Added)
has function, has role
Independent continuant
10 Classes (+10 Added)
start date value, end date value
date/time interval
3 Classes (+3 Added)
occurs in, realizes
biological process
397 Classes
(+396 Added, +1 Modified)
inheres in, is realized by
realizable entity
77 Classes
(+25 Added, +52 Modified)
has author, is about
document
66 Classes
(+64 Added, +2 Modified)
contact info for, is about
contact
6 Classes (+6 Added)
inheres in, subcontracted through
grant
1 Class (+1 Added)
has format, is about
data
67 Classes
(+67 Modified)
contact info for, is about, member
group (vcard)
1 Class (+1 Added)
has perquisite, prerequisite for, realizes
course
1 Class (+1 Added)
has phenotype anatomical entity, has phenotype quality, is about
anatomical phenotype annotation
1 Class (+1 Modified)
geographic focus of, has function, has role
geographic region
15 Classes (+15 Added)
is about, objective achieved by operation, objective achieved by
data transformation objective
300 Classes (+300 Modified)
is agent in, has function, has role
agent (foaf)
6 Classes
(+1 Added, +5 Modified)
has phenotype biological process, has phenotype quality, is about
biological process phenotype annotation
1 Class (+1 Modified)
date filed, has author, is about
patent
1 Class (+1 Modified)
funded by, has residency requirement, inheres in, is realized by
research opportunity
1 Class (+1 Modified)
has specified input, has specified output, realizes, achieves planned objective
planned process
34 Classes
(+5 Added, +29 Modified)
has agent, has participant, preceded by, precedes
occurent
11 Classes
(-11 Removed)
produced by, produces, is manufactured by, has function, has role
material entity
7,758 Classes
(+7,727 Added, +31 Modified)
has specified input, has specified output, is performed by, realizes, achieves planned objective
technique
791 Classes
(+4 Added, +787 Modified)
has agent, has participant, preceded by, precedes
processual entity
0 Classes
(-6 Removed, -7 Modified)
has author, has documentation, has topic, is about, specifies the use of, used by
protocol
1 Class (+1 Modified)
derives from organism or virus, produced by, produces, is manufactured by, has function, has role
immunogenic material
1 Class (+1 Modified)
has agent, has participant, occurs in, preceded by, precedes, realizes
biological process
0 Classes (-1 Modified)
has documentation, has manufacturer, produced by, produces, is manufactured by, has function, has role
reagent
1 Class (+1 Modified)
produced by, produces, recognizes species, species non reactivity, is manufactured by, has function, has role
biological target
10 Classes (+10 Modified)
has documentation, has input type, has manufacturer, has topic, is about, is encoded in, related technique, used by
database
2 Classes (+1 Added, +1 Modified)
achieves planned objective, has agent, has participant, has specified input, has specified output, preceded by, precedes, realizes
planned process
0 Classes (-29 Modified)
has documentation, has manufacturer, produced by, produces, related technique, is manufactured by, has function, has role
instrument
897 Classes (+897 Modified)
reagent library
12 Classes (+12 Modified)
chemical reagent
7 Classes (+7 Modified)
microarray
5 Classes (+5 Modified)
gene knockdown reagent
2 Classes (+2 Modified)
agent in, derived into, derives from, located in, location of, participates in, transformation of, transformed into
continuant
0 Classes
(-8 Removed, -439 Modified)
has documentation, has topic, is about, provides access to, realizes protocol, related technique, service provided by, uses software
service offering
17 Classes (+17 Modified)
agent in, derived into, derives from, has mammalian phenotype, located in, location of, +3 more
mammalian phenotype annotation
0 Classes (-1 Modified)
agent in, derived into, derives from, has author, located in, location of, +3 more
document
0 Classes (-3 Modified)
agent in, derived into, derives from, has author, located in, location of, + 3 more
data
0 Classes (-68 Modified)
achieves planned objective, has agent, has participant, has specified input, has specified output, is performed by, + 3 more
technique
0 Classes (-3 Removed, -787 Modified)
derives from organism or virus, has documentation, has manufacturer, produced by, produces, related technique, +3 more
nucleic acid library
31 Classes (+31 Modified)
has documentation, has manufacturer, has sequence alteration, produced by, produces, related technique, +3 more
protein reagent
3 Classes (+3 Modified)
agent in, derived into, derives from, inheres in, located in, location of, +3 more
specifically dependent continuant
0 Classes (-2 Removed, -11 Modified)
agent in, derived into, derives from, located in, location of, objective achieved by, +3 more
objective specification
0 Classes (-2 Modified)
agent in, derived into, derives from, has phenotype biological process, has phenotype quality, located in, +4 more
biological process phenotype annotation
0 Classes (-1 Modified)
has documentation, has phenotype annotation, has sequence alteration, organism generated from construct, produced by, produces, +4 more
organism
2 Classes (+2 Modified)
author of, has expertise, is agent in, is director of, is member of, is PI of, +4 more
person
6 Classes
(+5 Added, +1 Modified)
expressed in organism, has documentation, has manufacturer, produced by, produces, related technique, +4 more
construct
21 Classes (+21 Modified)
agent in, derived into, derives from, has function, has role, located in, +4 more
Independent continuant
0 Classes (-3 Removed)
agent in, derived into, derives from, located in, location of, objective achieved by, +4 more
data transformation objective
0 Classes (-300 Modified)
agent in, derived into, derives from, inheres in, is realized by, located in, +4 more
realizable entity
0 Classes
(-5 Removed, -52 Modified)
derives from organism or virus, has documentation, has manufacturer, has sequence alteration, produced by, produces, +4 more
nucleic acid reagent
14 Classes (+14 Modified)
agent in, derived into, derives from, has phenotype anatomical entity, has phenotype quality, located in, +4 more
anatomical phenotype annotation
0 Classes (-1 Modified)
derives from organism, has antibody target, has documentation, has immunogenic material, has manufacturer, produced by, +5 more
antibody reagent
2 Classes (+2 Modified)
expressed in organism, has documentation, has manufacturer, produced by, produces, related technique, +5 more
viral plasmid
6 Classes (+6 Modified)
funded by, has documentation, has PI, has specified input, has specified output, has topic, +5 more
research project
2 Classes (+2 Modified)

Figure 3.39 The top portion of the ERO diff partial-area taxonomy, summarizing all classes with 0–11 object properties.
Source: [121]

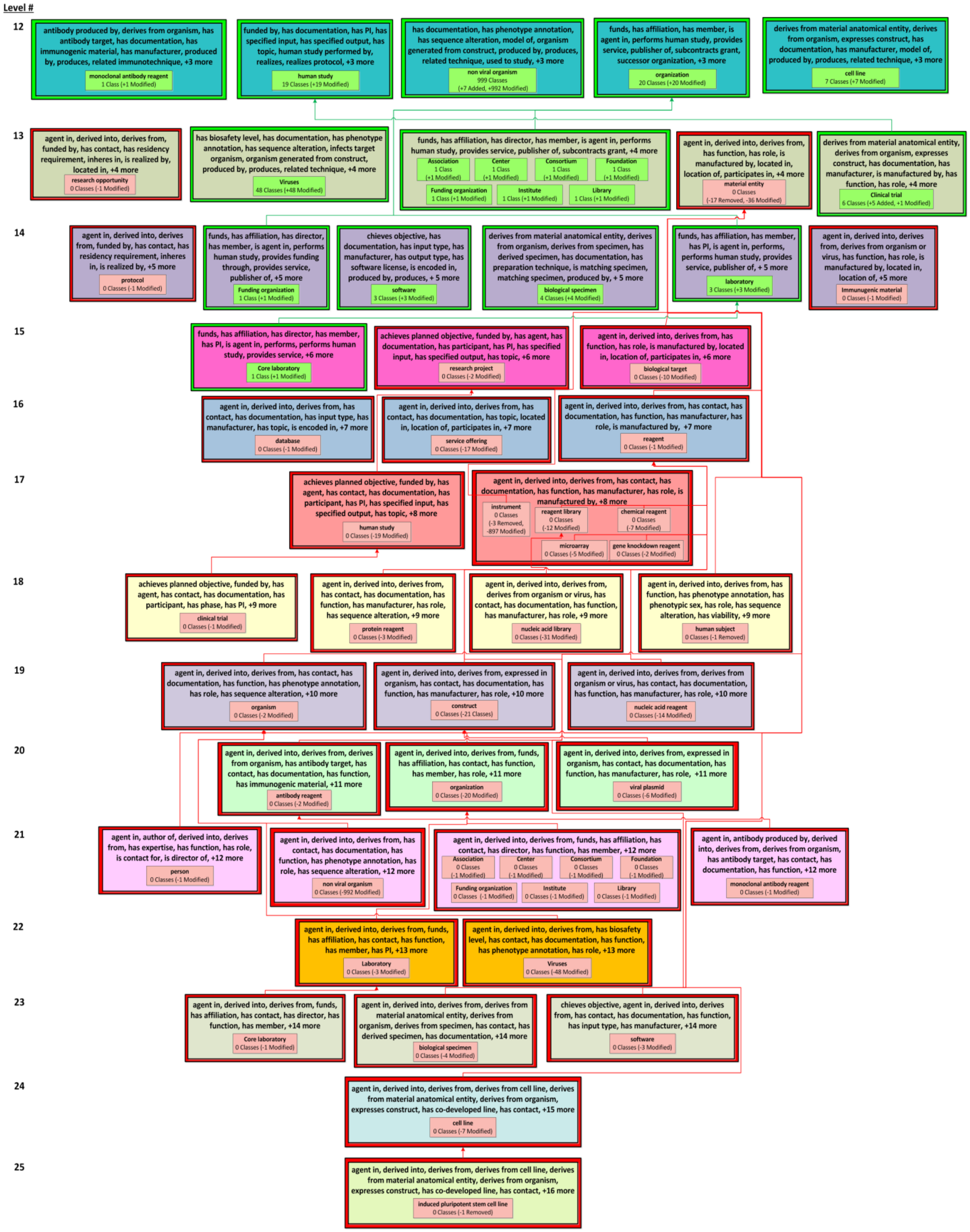

Level #
12
antibody produced by, derives from organism, has antibody target, has documentation, has immunogenic material, has manufacturer, produced by, produces, related immunotechnique, +3 more
monoclonal antibody reagent
1 Class (+1 Modified)
funded by, has documentation, has PI, has specified input, has specified output, has topic, human study performed by, realizes, realizes protocol, +3 more
human study
19 Classes (+19 Modified)
has documentation, has phenotype annotation, has sequence alteration, model of, organism generated from construct, produced by, produces, related technique, used to study, +3 more
non viral organism
999 Classes
(+7 Added, +992 Modified)
funds, has affiliation, has member, is agent in, performs human study, provides service, publisher of, subcontracts grant, successor organization, +3 more
organization
20 Classes (+20 Modified)
derives from material anatomical entity, derives from organism, expresses construct, has documentation, has manufacturer, model of, produced by, produces, related technique, +3 more
cell line
7 Classes (+7 Modified)
13
agent in, derived into, derives from, funded by, has contact, has residency requirement, inheres in, is realized by, located in, +4 more
research opportunity
0 Classes (-1 Modified)
has biosafety level, has documentation, has phenotype annotation, has sequence alteration, infects target organism, organism generated from construct, produced by, produces, related technique, +4 more
Viruses
48 Classes (+48 Modified)
funds, has affiliation, has director, has member, is agent in, performs human study, provides service, publisher of, subcontracts grant, +4 more
Association
1 Class
(+1 Modified)
Center
1 Class
(+1 Modified)
Consortium
1 Class
(+1 Modified)
Foundation
1 Class
(+1 Modified)
Funding organization
1 Class (+1 Modified)
Institute
1 Class (+1 Modified)
Library
1 Class (+1 Modified)
agent in, derived into, derives from, has function, has role, is manufactured by, located in, location of, participates in, +4 more
material entity
0 Classes
(-17 Removed, -36 Modified)
derives from material anatomical entity, derives from organism, expresses construct, has documentation, has manufacturer, is manufactured by, has function, has role, +4 more
Clinical trial
6 Classes (+5 Added, +1 Modified)
14
agent in, derived into, derives from, funded by, has contact, has residency requirement, inheres in, is realized by, +5 more
protocol
0 Classes (-1 Modified)
funds, has affiliation, has director, has member, is agent in, performs human study, provides funding through, provides service, publisher of, +5 more
Funding organization
1 Class (+1 Modified)
chieves objective, has documentation, has input type, has manufacturer, has output type, has software license, is encoded in, produced by, produces, + 5 more
software
3 Classes (+3 Modified)
derives from material anatomical entity, derives from organism, derives from specimen, has derived specimen, has documentation, has preparation technique, is matching specimen, matching specimen, produced by, + 5 more
biological specimen
4 Classes (+4 Modified)
funds, has affiliation, has member, has PI, is agent in, performs, performs human study, provides service, publisher of, + 5 more
laboratory
3 Classes (+3 Modified)
agent in, derived into, derives from, derives from organism or virus, has function, has role, is manufactured by, located in, location of, +5 more
Immunugenic material
0 Classes (-1 Modified)
15
funds, has affiliation, has director, has member, has PI, is agent in, performs, performs human study, provides service, +6 more
Core laboratory
1 Class (+1 Modified)
achieves planned objective, funded by, has agent, has documentation, has participant, has PI, has specified input, has specified output, has topic, +6 more
research project
0 Classes (-2 Modified)
agent in, derived into, derives from, has function, has role, is manufactured by, located in, location of, participates in, +6 more
biological target
0 Classes (-10 Modified)
16
agent in, derived into, derives from, has contact, has documentation, has input type, has manufacturer, has topic, is encoded in, +7 more
database
0 Classes (-1 Modified)
agent in, derived into, derives from, has contact, has documentation, has topic, located in, location of, participates in, +7 more
service offering
0 Classes (-17 Modified)
agent in, derived into, derives from, has contact, has documentation, has function, has manufacturer, has role, is manufactured by, +7 more
reagent
0 Classes (-1 Modified)
17
achieves planned objective, funded by, has agent, has contact, has documentation, has participant, has PI, has specified input, has specified output, has topic, +8 more
human study
0 Classes (-19 Modified)
agent in, derived into, derives from, has contact, has documentation, has function, has manufacturer, has role, is manufactured by, +8 more
instrument
0 Classes
(-3 Removed,
-897 Modified)
reagent library
0 Classes
(-12 Modified)
chemical reagent
0 Classes
(-7 Modified)
microarray
0 Classes (-5 Modified)
gene knockdown reagent
0 Classes (-2 Modified)
18
achieves planned objective, funded by, has agent, has contact, has documentation, has participant, has phase, has PI, +9 more
clinical trial
0 Classes (-1 Modified)
agent in, derived into, derives from, has contact, has documentation, has function, has manufacturer, has role, has sequence alteration, +9 more
protein reagent
0 Classes (-3 Modified)
agent in, derived into, derives from, derives from organism or virus, has contact, has documentation, has function, has manufacturer, has role, +9 more
nucleic acid library
0 Classes (-31 Modified)
agent in, derived into, derives from, has function, has phenotype annotation, has phenotypic sex, has role, has sequence alteration, has viability, +9 more
human subject
0 Classes (-1 Removed)
19
agent in, derived into, derives from, has contact, has documentation, has function, has phenotype annotation, has role, has sequence alteration, +10 more
organism
0 Classes (-2 Modified)
agent in, derived into, derives from, expressed in organism, has contact, has documentation, has function, has manufacturer, has role, +10 more
construct
0 Classes (-21 Classes)
agent in, derived into, derives from, derives from organism or virus, has contact, has documentation, has function, has manufacturer, has role, +10 more
nucleic acid reagent
0 Classes (-14 Modified)
20
agent in, derived into, derives from, derives from organism, has antibody target, has contact, has documentation, has function, has immunogenic material, +11 more
antibody reagent
0 Classes (-2 Modified)
agent in, derived into, derives from, funds, has affiliation, has contact, has function, has member, has role, +11 more
organization
0 Classes (-20 Modified)
agent in, derived into, derives from, expressed in organism, has contact, has documentation, has function, has manufacturer, has role, +11 more
viral plasmid
0 Classes (-6 Modified)
21
agent in, author of, derived into, derives from, has expertise, has function, has role, is contact for, is director of, +12 more
person
0 Classes (-1 Modified)
agent in, derived into, derives from, has contact, has documentation, has function, has phenotype annotation, has role, has sequence alteration, +12 more
non viral organism
0 Classes (-992 Modified)
agent in, derived into, derives from, funds, has affiliation, has contact, has director, has function, has member, +12 more
Association
0 Classes
(-1 Modified)
Center
0 Classes
(-1 Modified)
Consortium
0 Classes
(-1 Modified)
Foundation
0 Classes
(-1 Modified)
Funding organization
0 Classes (-1 Modified)
Institute
0 Classes (-1 Modified)
Library
0 Classes (-1 Modified)
agent in, antibody produced by, derived into, derives from, derives from organism, has antibody target, has contact, has documentation, has function, +12 more
monoclonal antibody reagent
0 Classes (-1 Modified)
22
agent in, derived into, derives from, funds, has affiliation, has contact, has function, has member, has PI, +13 more
Laboratory
0 Classes (-3 Modified)
agent in, derived into, derives from, has biosafety level, has contact, has documentation, has function, has phenotype annotation, has role, +13 more
Viruses
0 Classes (-48 Modified)
23
agent in, derived into, derives from, funds, has affiliation, has contact, has director, has function, has member, +14 more
Core laboratory
0 Classes (-1 Modified)
agent in, derived into, derives from, derives from material anatomical entity, derives from organism, derives from specimen, has contact, has derived specimen, has documentation, +14 more
biological specimen
0 Classes (-4 Modified)
chieves objective, agent in, derived into, derives from, has contact, has documentation, has function, has input type, has manufacturer, +14 more
software
0 Classes (-3 Modified)
24
agent in, derived into, derives from, derives from cell line, derives from material anatomical entity, derives from organism, expresses construct, has co-developed line, has contact, +15 more
cell line
0 Classes (-7 Modified)
25
agent in, derived into, derives from, derives from cell line, derives from material anatomical entity, derives from organism, expresses construct, has co-developed line, has contact, +16 more
induced pluripotent stem cell line
0 Classes (-1 Removed)

Figure 3.40 The bottom portion of the ERO diff partial-area taxonomy. Most of the diff areas at lower levels are removed areas due to the reduction in ERO's complexity.
Source: [121]

Several of these object properties were replaced with object properties from a newer version of RO that had no domains (e.g., *has participant*). However, their sub properties (e.g., *has specified input*), retained the same domains after the merge.

In regards to the 91 newly added object properties, from the DPAT one can see that these newly introduced properties have domains that are mostly disjoint, since there are few classes that are in the domain of many object properties. There are a total of 13 introduced areas in the top four levels of Figure 3.40 but 26 removed areas in levels 16-25. After the merge, the most complex class is *core laboratory*, with 15 object properties, as compared to the most complex class before the merge, *induced pluripotent stem cell line*, with 25 object properties.

The DPAT view shows that by adding more object properties than were removed, ERO became richer in terms of types of properties used to define classes, but also simpler in its model, as the number of object properties per class was reduced.

The second kind of change that led to a reduction in complexity was the modification of various object property domains. For example, before the merge, the domain of the object property *has sequence alteration* was (c*ell line* **or** *protein reagent* **or** *nucleic acid reagent* **or** *human subject* **or** *organism*). After the merge, the introduced partial-area *cell line* is no longer in the object property's domain.

By comparing the object properties of the removed partial-area *cell line* to the introduced partial-area *cell line*, it was found that no new object properties were added to *cell line*, five removed object properties were removed from RO (*agent in, derived into, derives from, located in*, and *location of*), one RO object property's domain was modified (*participates in*), three ERO object properties were removed (*derives from cell line, has* *co-developed line,* and *has contact*), and one ERO object property was modified (*has sequence alteration*). Matthew Brush, an ERO curator, reviewed a sample of added/removed diff area pairs that contained classes defined by ERO (e.g., *Document, Organization*, *Person,* and *Technique*) and confirmed that their classes had the correct object properties.

Another major structural change for ERO is evident from the very large introduced partial-areas in the DPAT, e.g., the *information content entity* introduced partial-area in the introduced area {*is about*} summarizes 9,266 classes. Over 8,800 of these classes are new to ERO. Most were imported from other ontologies, e.g., the Mammalian Phenotype Ontology (MP) [148] and the Software Ontology (SWO) [149]. Similarly, most of the 7,727 added classes in the introduced partial-area *material entity* (7758) are imported from UBERON [150]. From the DPAT, one can see the property structure of the classes imported from these ontologies.

3.4.4 Comparison to Traditional Ontology Diff Output

The DPATs of OCRe, SDO, and ERO were compared to the output of Protégé's "compare ontologies" tool ("Protégé diff" for short, see Figure 2.13), derived using the same before and after ontology versions of the respective ontologies. When applied to OCRe, the diff identified 27 modified entities (classes, properties, etc.). However, since OCRe underwent additional development outside of the QA review [25], eleven of these entities did not have any structural changes (only changes to annotations, e.g., class labels). Four modified classes had restrictions removed, which is not captured by the DPAT shown in Figure 3.37, but is captured by a DPAT derived using restrictions (like SDO's). The remaining 12 modified entities relate to the addition and removal of classes and changes to object property domains, e.g., *duration*. Without a DPAT, identifying the 12 structural changes requires a user to manually review each change in the standard diff. Furthermore, the standard diff does not provide a view that shows the definition of the classes impacted by the change, e.g., *Relative time point*, which already had the *has offset* object property.

In comparison with the Protégé diff, the DPAT provided a more accurate and concise view of the implicit structural changes that occurred. The Protégé diff did not explicitly identify the removal of the 33 statistical classes from the hierarchy, which was a major change. The only differences identified that were related to this change were the modifications to the domains of object properties *has dependent variable* and *has independent variable*. The removal of the classes from the hierarchy is only apparent after applying a reasoner (e.g., Hermit [67]) to the ontology and performing a manual comparison of the output and the input.

The Protégé diff for the SDO identified one added class, ten removed classes, and seven other structural changes (e.g., the equivalences described in the Results section). Unlike OCRe, which underwent development unrelated to QA, the SDO only changed due to the error corrections described by Ochs et al. [120]. However, the Protégé diff did not provide a complete picture of the changes that occurred, particularly in regards to inheritance of properties. For example, while the Protégé diff identified the added equivalence axioms between the two *clinical finding* classes, it did not capture how this change affected their many descendent classes. Furthermore, the Protégé diff did not provide a way of directly comparing the properties for the classes that were declared equivalent. Additionally, it did not uncover that the object property sets for these equivalent classes were not equivalent, as found in the DPAT.

The Protégé diff of ERO was several orders of magnitude larger than the Protégé diffs for OCRe and SDO. A total of 19,256 entities (mostly classes) were identified as created, 159 were deleted, 27 were renamed, and 609 were modified. Reviewing each of these changes (20,051 in total) is impractical. In comparison with the DPAT, the Protégé diff is overwhelming.

In addition to comparing Protégé diff outputs with DPATs, Samson Tu, Sivaram Arabandi, and Melissa Haendel, the curators of OCRe, SDO, and ERO, respectively, to comment on how they used structural diff tools during the previously described development phases [25, 120, 141]. After correcting the errors found by Ochs et al. [25], Samson Tu did not use any diff tools to compare the before and after versions due to the small number of relatively simple changes. In general, he uses OWLDiff [91] when there is a specific need to compare the axioms of two ontology versions. When initially designing the SDO Sivaram Arabandi also occasionally used OWL Diff. However, due to the limited benefits he derived from using it, he did not use it to compare the two releases of the SDO reported in previous work [120]. In contrast Sivaram Arabandi found the DPAT very helpful due to the visualization that compactly summarizes changes. In comparison, OWL Diff presents changes in a text-based indented hierarchy, which can be overwhelming in length, making it difficult to find an important change.

During the merge of ERO and VIVO, the ERO development team used an in-house diff tool [141], that integrates spreadsheet-based information, e.g., class equivalences, with Protégé. Their diff tool highlights different classes based on various modeling decisions. They did not use any third-party diff tools, e.g., OWLDiff or Protégé’s Compare Ontologies tool, due to the various needs and levels of experience on the team responsible for the merge. Melissa Haendel confirmed that by combining the visualization of the DPAT with an explanation of why the different DPAT elements changed (e.g., as done for the *cell line* diff partial-areas), the Diff AbN approach would be helpful when developing and merging ontologies.

3.4.5 Discussion

The development of Diff AbNs addresses the need for methods of summarizing, and visualizing structural changes between two ontology releases. A curator can inspect the change summary provided by the DAT and DPAT to review global changes, as well as determine if the changes have any unintended side-effects (e.g., incorrectly assigned or inferred object properties). In particular, due to the summary, the curator could quickly determine if the classes in the various areas and partial-areas have the intended object properties.

Such a detection of unintended consequences is less likely if the curator needs to review an OWL-based structural diff between ontologies [91, 114, 115] since the amount of information would be overwhelming, as detailed for the SDO audit [120]. Furthermore, unintended and erroneous changes may be identified by reviewing a Diff AbN for nonconsecutive releases, since some unintended changes may not be detected for consecutive pairs of releases, but may be detected between releases that are farther apart, due to the cumulative impact of the changes made between consecutive releases.

In comparison with standard ontology diff approaches, which generally only identify individual changes per-entity (e.g., class or property), the Diff AbN approach shows the global impact of an editing operation. With the Diff AbN a user does not have to manually scan through potentially hundreds or thousands of entries to identify important structural changes. Furthermore, the Diff AbN approach shows the implicit changes that occur due to inheritance of properties within an ontology, e.g., for the many descendants of *Clinical finding* in the SDO.

Notably, even when comparing diagrams of the complete before and after taxonomies of two releases it is difficult for a curator to notice the differences between them. She would have to manually compare the classes and object properties of these two taxonomies and detect the changes. This task is overwhelming for the curator. Thus, the DPAT was introduced to summarize the changes.

The Diff AbNs described in Section 3.4.1 are based on object properties. Several kinds of Diff AbNs can be derived, based on the structural features that are used for the derivation, e.g., data properties can be used instead of object properties. The same general approach for Diff AbN derivation described in this paper can be adapted accordingly. Future research will investigate Diff AbNs based on data properties, equivalence axioms, etc., and their use in uncovering unintended changes.

Another potential use of Diff AbNs is to compare the stated and inferred versions of an ontology to determine if the inferred axioms are correct or have unintended consequences. An error may not be easily detectable in the stated view but may become apparent after a reasoner has been applied. The DAT and the DPAT would show the structural differences between these two views. By creating a DPAT between the stated version of OCRe and the inferred version of OCRe (before the QA review), it would be easier to identify the incorrect object property domains and the erroneous inclusion of 33 statistical classes into OCRe's *Entity* hierarchy.

One potential issue with the DAT and the DPAT is that they produce diagrams that are larger than the taxonomy diagrams of O_{before} and of O_{after}. A DPAT shows all of the areas and partial-areas of the "before partial-area taxonomy" and the "after partial-area taxonomy." For example, in the DPATs of SDO (Figure 3.38) and ERO (Figures 3.39 and 3.40), there are many pairs of added/removed areas and partial-areas. One way of simplifying the DAT and DPAT is to define various views that only show certain types of Diff AbN elements. For example, if a curator is only interested in what has changed, then she can hide unmodified areas and unmodified partial-areas. Alternatively, the curator can view only introduced areas and partial-areas, etc.

One potential drawback of the DPAT is that internal changes within diff partial-areas (e.g., changing the subclass hierarchy within an unmodified partial-area) are not identified. For such a case, a structural diff excerpt for the changed classes within the partial-area could be reviewed, thus producing a targeted partial ontology diff that does not overwhelm a user.

In conclusion, the Diff Abstraction Network approach can support a compact global view of structural changes. Furthermore, it can support the detection of unintended and erroneous changes resulting from a QA effort, as illustrated using examples from OCRe and SDO. Additionally, it can identify the structural changes that occurred when two ontologies were merged, as demonstrated with ERO.

3.5 Abstraction Network Tools

Authoring, maintaining, and browsing ontologies requires the use of software tools, such as Protégé [56, 151] for OWL ontologies, OBO-edit [94] for OBO ontologies, the Neighborhood Auditing Tool [98, 99] for the UMLS, the NLM UTS browsers for the UMLS and SNOMED CT [58], and the IHTSDO Workbench [95], CliniClue browser [57], and Snow Owl [97] for SNOMED CT.

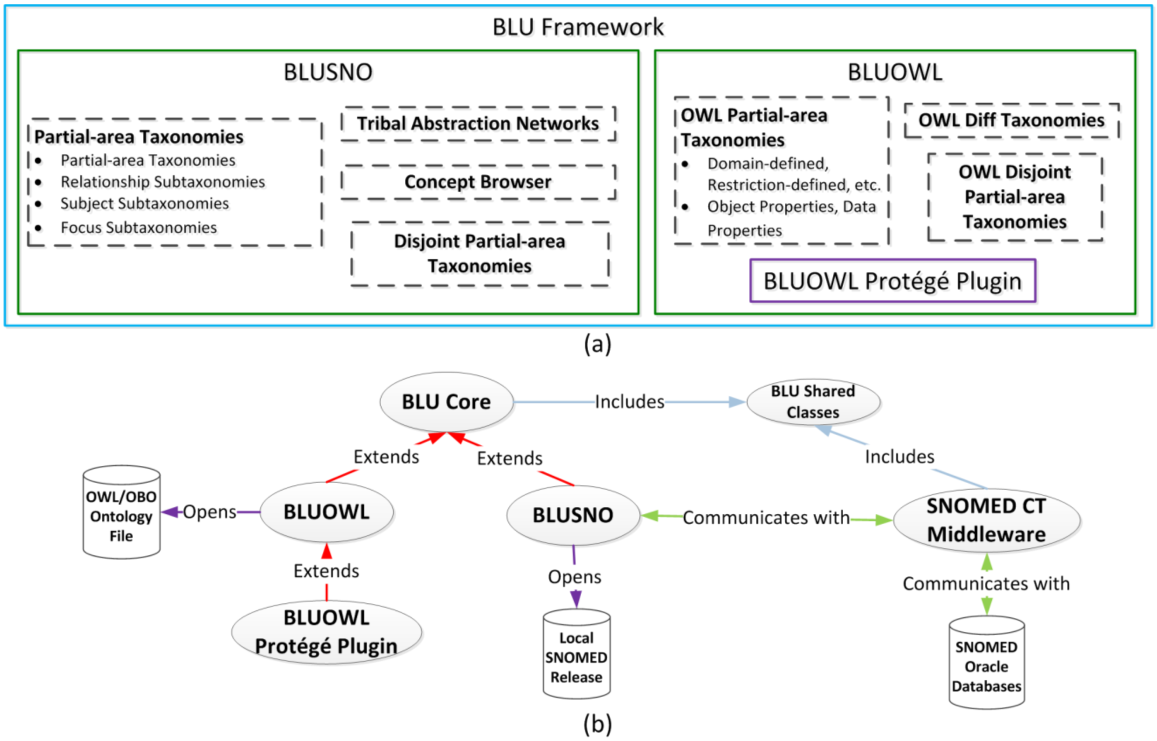


Figure 3.41 **(a**) The subcomponents of the BLUSNO and BLUOWL. **(b)** The structure of the major components of the BLU Framework.

Similarly, software tools are required for creating and exploring abstraction networks. Utilities for automatically creating and visualizing abstraction networks can support abstraction network research and improve the usability of abstraction networks. Such tools need to provide useful information about both the abstraction network and the underlying ontology.

The Biomedical Layout Utility Framework ("BLU Framework") is a collection of software tools for deriving, visualizing, and exploring various kinds of abstraction networks. The BLU Framework is comprised of two major components: the Biomedical Layout Utility for SNOMED CT (BLUSNO) [118] and the Biomedical Layout Utility for the Web Ontology Language (BLUOWL).

Table 3.26 Summary of the BLU Framework's Subcomponents

Component	**Description**
BLU Shared Classes	High level, generic code for representing common ontology elements (e.g., concepts) and abstraction network elements (e.g., nodes).
BLU Core	Generic functionality for deriving and visualizing various kinds of abstraction networks, including taxonomies and tribal abstraction networks.
BLUSNO	Software for deriving SNOMED CT abstraction networks. Also includes a concept browser for viewing SNOMED's content in a traditional concept-centric view. BLUSNO works using either locally stored SNOMED releases or through a web-based middleware.
SNOMED CT Middleware	Web-based middleware for accessing SNOMED CT releases, and associated partial-area taxonomies, which are stored in Oracle databases. Used by BLUSNO when no local release is available on a user's computer.
BLUOWL	Software for deriving abstraction networks for OWL ontologies, including partial-area taxonomies derived using various structural features and diff taxonomies. BLUOWL enables a user to open multiple ontologies that are represented using either OWL or OBO format.
BLUOWL Protégé Plugin	An extension of BLUOWL that is accessible within Protégé [56]. Includes functionality that integrates the BLUOWL user interface into Protégé and vice-versa.

All of the BLU Framework subcomponents are based on a core library that provides generic functionality for creating, representing, and displaying abstraction networks. The core library also provides generic user interface functionality for displaying and searching for abstraction network element information. BLUSNO and BLUOWL extend on this core library by implementing SNOMED-specific and OWL-specific functionality, respectively.

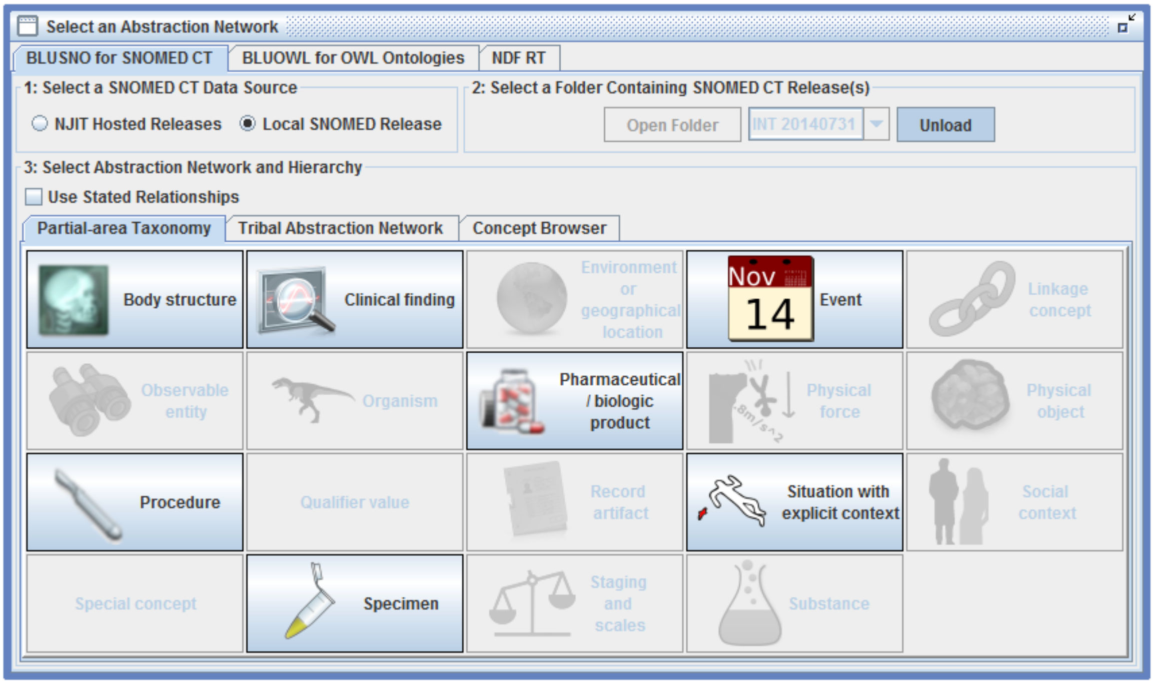
BLUSNO and BLUOWL both have several major subcomponents. Figure 3.41(a) illustrates these subcomponents and Figure 3.41(b) illustrates their dependencies and data sources. Each component is briefly described in Table 3.26 and the major components, BLUSNO and BLUOWL, will be explained in detail throughout the following sections. BLUSNO will be used to illustrate the abstraction network visualization and exploration functionality that is available in each component.

3.5.1 Biomedical Layout Utility for SNOMED CT

The Biomedical Layout Utility for SNOMED CT (BLUSNO) dynamically generates interactive visualizations of SNOMED CT abstraction networks, including area taxonomies, partial-area taxonomies, disjoint partial-area taxonomies, and tribal abstraction networks. BLUSNO also provides functionality for deriving the various subtaxonomies described in Section 3.1. Additionally, BLUSNO includes an innovative concept browser that integrates a traditional view of an ontology with abstraction network information.

When a user starts BLUSNO they are provided with two options for selecting a data source (i.e., a version of SNOMED). First, the user can choose to open a locally stored SNOMED CT release, e.g., one obtained from the NLM UTS [58]. This option enables all of the functionality of BLUSNO. Alternatively, if the user does not have access to a local SNOMED CT release he or she may choose to use a version that is hosted on a web server. This option enables a subset of the available functionality of the

BLUSNO tool (i.e., a user can only derive partial-area taxonomies and the built-in concept browser has limited functionality). When working in this alternate mode, BLUSNO communicates through a web-based middleware API that accesses and processes data stored in Oracle databases. The descriptions throughout this section assume the user has selected a local SNOMED release.



Select an Abstraction Network

BLUSNO for SNOMED CT | BLUOWL for OWL Ontologies | NDF RT

1: Select a SNOMED CT Data Source

○ NJIT Hosted Releases ● Local SNOMED Release

2: Select a Folder Containing SNOMED CT Release(s)

Open Folder | INT 20140731 | Unload

3: Select Abstraction Network and Hierarchy

□ Use Stated Relationships

Partial-area Taxonomy | Tribal Abstraction Network | Concept Browser

Figure 3.42 The BLUSNO abstraction network derivation user interface.

After a SNOMED release has been selected, the user chooses which type of abstraction network they want to derive. To derive a partial-area taxonomy or tribal abstraction network, a user simply selects which hierarchy they want to summarize. The user can further choose between deriving their chosen abstraction network for the inferred or stated view of SNOMED. The user interface for this process is shown in Figure 3.42. Disjoint partial-area taxonomies are derived in the context of their partial-area taxonomy (see Section 2.5.1.3). Subtaxonomies are derivable at appropriate locations, e.g., a relationship subtaxonomy can be derived by selecting an area in taxonomy and choosing the area's relationships as *R*'.

Prior to BLUSNO, creating SNOMED CT abstraction networks was accomplished using text-based reports generated by a variety of small, disconnected software utilities. Analyzing the data required extensive time and effort. There was no way to automatically visualize the abstraction networks; figures were created manually using a graphics editor. The task of creating these figures often required three or four days of work. Due to these limitations, visualization of large SNOMED CT abstraction networks, such as the area and partial-area taxonomies for the *Procedure* or *Clinical finding* hierarchies, was essentially impossible. Likewise, it was impractical to create disjoint partial-area taxonomy abstraction networks for areas which have hundreds, or even thousands, of overlapping concepts.

BLUSNO provides a user with a view where all of the information for a specific abstraction network is contained in a single window. The user can view multiple abstraction networks at the same time. Each window is a self-contained unit with options and functionality tied to the given abstraction network. Within each window the user can switch between interfaces for exploring and editing the associated taxonomy.

The main functionality of BLUSNO (and the BLU Framework, in general) is defined by its graphical (diagram) interface, where each abstraction network element (e.g., partial-areas, areas, or clusters) is selectable and provides information about the underlying terminology and the structure of the abstraction network.

3.5.1.1 Abstraction Network Visualization in BLUSNO. The graphical (diagrammatic) interface of BLUSNO (shown in Figure 3.43) produces interactive displays that are modeled after the static abstraction network diagrams constructed in Wang et. al [24], Ochs et al. [27], etc.

In BLUSNO, users have the ability to move, pan, and zoom throughout an abstraction network, quickly perceiving how the knowledge is structured. To limit visual complexity, *child-of* connections are only shown on request and typically are limited to connections between a small number of abstraction network nodes.

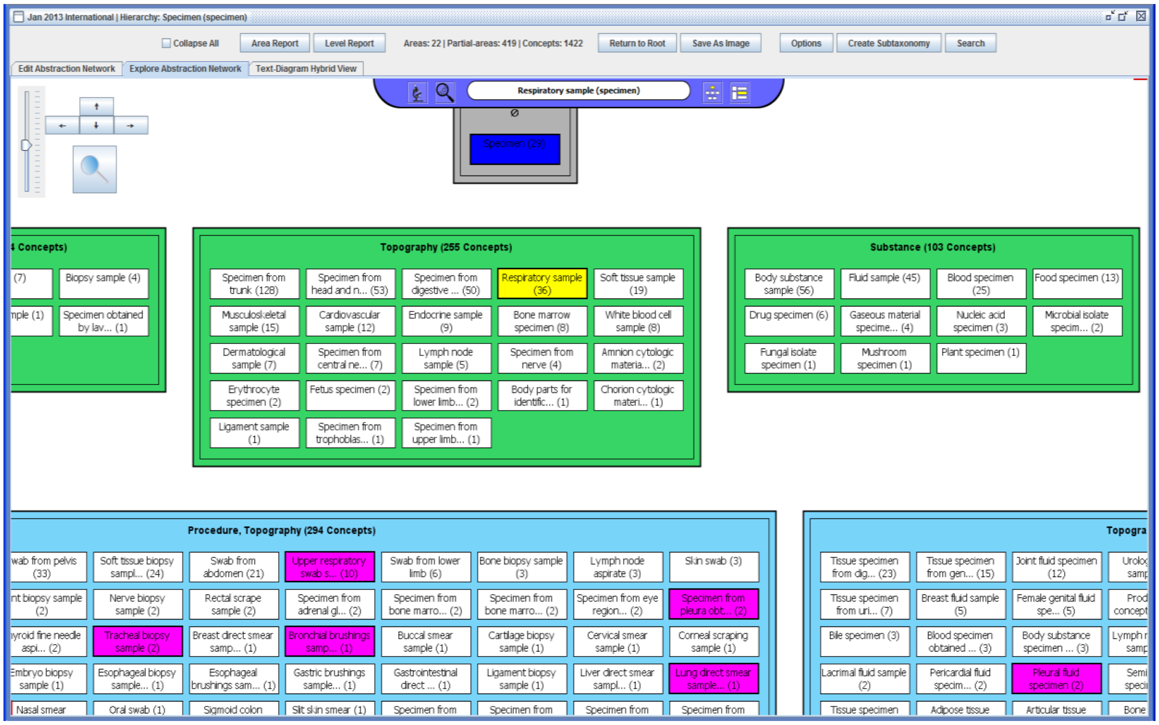
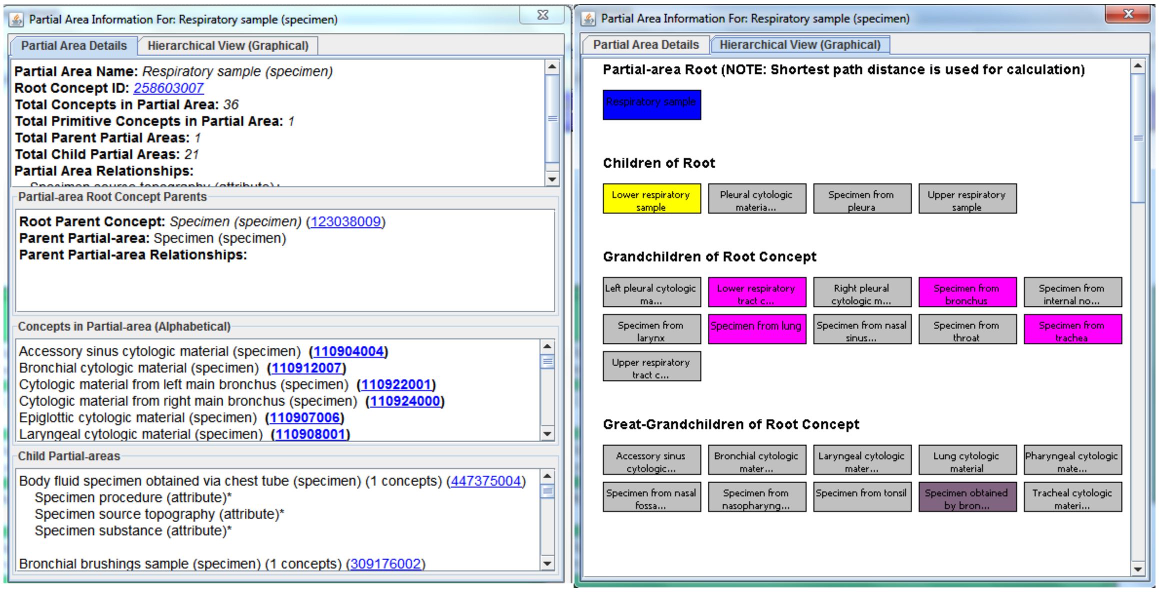

Jan 2013 International | Hierarchy: Specimen (specimen)
☐ Collapse All
Area Report
Level Report
Areas: 22 | Partial-areas: 419 | Concepts: 1422
Return to Root
Save As Image
Options
Create Subtaxonomy
Search
Edit Abstraction Network
Explore Abstraction Network
Text-Diagram Hybrid View
Respiratory sample (specimen)
∅
Specimen (29)
4 Concepts)
(7)
Biopsy sample (4)
mple (1)
Specimen obtained by lav... (1)
Topography (255 Concepts)
Specimen from trunk (128)
Specimen from head and n... (53)
Specimen from digestive ... (50)
Respiratory sample (36)
Soft tissue sample (19)
Musculoskeletal sample (15)
Cardiovascular sample (12)
Endocrine sample (9)
Bone marrow specimen (8)
White blood cell sample (8)
Dermatological sample (7)
Specimen from central ne... (7)
Lymph node sample (5)
Specimen from nerve (4)
Amnion cytologic materia... (2)
Erythrocyte specimen (2)
Fetus specimen (2)
Specimen from lower limb... (2)
Body parts for identific... (1)
Chorion cytologic materi... (1)
Ligament sample (1)
Specimen from trophoblas... (1)
Specimen from upper limb... (1)
Substance (103 Concepts)
Body substance sample (56)
Fluid sample (45)
Blood specimen (25)
Food specimen (13)
Drug specimen (6)
Gaseous material specime... (4)
Nucleic acid specimen (3)
Microbial isolate specim... (2)
Fungal isolate specimen (1)
Mushroom specimen (1)
Plant specimen (1)
Procedure, Topography (294 Concepts)
wab from pelvis (33)
Soft tissue biopsy sampl... (24)
Swab from abdomen (21)
Upper respiratory swab s... (10)
Swab from lower limb (6)
Bone biopsy sample (3)
Lymph node aspirate (3)
Skin swab (3)
nt biopsy sample (2)
Nerve biopsy sample (2)
Rectal scrape sample (2)
Specimen from adrenal gl... (2)
Specimen from bone marro... (2)
Specimen from bone marro... (2)
Specimen from eye region... (2)
Specimen from pleura obt... (2)
yroid fine needle aspi... (2)
Tracheal biopsy sample (2)
Breast direct smear samp... (1)
Bronchial brushings samp... (1)
Buccal smear sample (1)
Cartilage biopsy sample (1)
Cervical smear sample (1)
Corneal scraping sample (1)
Embryo biopsy sample (1)
Esophageal biopsy sample... (1)
Esophageal brushings sam... (1)
Gastric brushings sample... (1)
Gastrointestinal direct ... (1)
Ligament biopsy sample (1)
Liver direct smear sampl... (1)
Lung direct smear sample... (1)
Nasal smear
Oral swab (1)
Sigmoid colon
Slit skin smear (1)
Specimen from
Specimen from
Specimen from
Specimen from
Topogra
Tissue specimen from dig... (23)
Tissue specimen from gen... (15)
Joint fluid specimen (12)
Urolog samp
Tissue specimen from uri... (7)
Breast fluid sample (5)
Female genital fluid spe... (5)
Prod concept
Bile specimen (3)
Blood specimen obtained ... (3)
Body substance specimen ... (3)
Lymph samp
Lacrimal fluid sample (2)
Pericardial fluid specim... (2)
Pleural fluid specimen (2)
Semi speci
Tissue specimen
Adipose tissue
Articular tissue
Bone

Figure 3.43 BLUSNO's graphical interface with the *Specimen* hierarchy's partial-area taxonomy shown. The partial-area *Respiratory sample* (36) has been selected by clicking on it (yellow) and its parent (blue) and children (purple) are highlighted.

All of the elements of an abstraction network are interactive in that they provide specific information and features when clicked on by a user. For example, when viewing a partial-area taxonomy, if a user single clicks on a partial-area then its parent and child partial-areas will be highlighted in blue and purple, respectively.



Partial Area Information For: Respiratory sample (specimen)

Partial Area Details | Hierarchical View (Graphical)

Partial Area Name: *Respiratory sample (specimen)*
Root Concept ID: *258603007*
Total Concepts in Partial Area: *36*
Total Primitive Concepts in Partial Area: *1*
Total Parent Partial Areas: *1*
Total Child Partial Areas: *21*
Partial Area Relationships:
Specimen source topography (attribute):

Partial-area Root Concept Parents

Root Parent Concept: *Specimen (specimen)* (123038009)
Parent Partial-area: Specimen (specimen)
Parent Partial-area Relationships:

Concepts in Partial-area (Alphabetical)

Accessory sinus cytologic material (specimen) **(110904004)**
Bronchial cytologic material (specimen) **(110912007)**
Cytologic material from left main bronchus (specimen) **(110922001)**
Cytologic material from right main bronchus (specimen) **(110924000)**
Epiglottic cytologic material (specimen) **(110907006)**
Laryngeal cytologic material (specimen) **(110908001)**

Child Partial-areas

Body fluid specimen obtained via chest tube (specimen) (1 concepts) (447375004)
Specimen procedure (attribute)*
Specimen source topography (attribute)*
Specimen substance (attribute)*

Bronchial brushings sample (specimen) (1 concepts) (309176002)

Partial Area Information For: Respiratory sample (specimen)

Partial Area Details | Hierarchical View (Graphical)

Partial-area Root (NOTE: Shortest path distance is used for calculation)

Respiratory sample

Children of Root

Lower respiratory sample | Pleural cytologic materia... | Specimen from pleura | Upper respiratory sample

Grandchildren of Root Concept

Left pleural cytologic ma... | Lower respiratory tract c... | Right pleural cytologic m... | Specimen from bronchus | Specimen from internal no...
Specimen from larynx | Specimen from lung | Specimen from nasal sinus... | Specimen from throat | Specimen from trachea
Upper respiratory tract c...

Great-Grandchildren of Root Concept

Accessory sinus cytologic... | Bronchial cytologic mater... | Laryngeal cytologic mater... | Lung cytologic material | Pharyngeal cytologic mate...
Specimen from nasal fossa... | Specimen from nasopharyng... | Specimen from tonsil | Specimen obtained by bron... | Tracheal cytologic materi...

Figure 3.44 **(Left)** The Partial-area Summary Dialog for the *Respiratory sample* partial-area. **(Right)** The singly rooted hierarchy of concepts in the *Respiratory sample* partial-area, with concept *Lower respiratory sample* (in yellow) selected.

In Figure 3.43 the partial-area *Respiratory sample* (36) has been selected by clicking on it. Its single parent partial-area, *Specimen* (29), is highlighted in blue. Its child partial-areas, such as *Upper respiratory swab sample* (10), are shown in purple.

When a user selects an abstraction network node, e.g., a partial-area, an options menu appears at the top of the display. An important option is displaying a dialog that provides various metrics and structural information about the selected node.

This dialog lists the parent nodes, child nodes, and the concepts summarized by the chosen node (in alphabetical order). The left side of Figure 3.44 shows this dialog for the partial-area *Respiratory sample*, from the *Specimen* partial-area taxonomy. This dialog also allows a user to visualize the subhierarchy of concepts summarized by the abstraction network node. For example, the right side of Figure 3.44 shows the subhierarchy of concepts summarized by the *Respiratory sample* partial-area. Individual concepts in this visualization can be selected to highlight their parents (blue) and children (purple) within the abstraction network node. On the right side of Figure 3.44 the concept *Lower respiratory sample* has been selected.

Users can search for concepts and specific nodes within an abstraction network by typing a search term into the search box. Clicking on a search result will focus the abstraction network window on the associated element. In addition to searching, several dialogs are available to provide summaries and metrics of the abstraction network's structure.

For example, clicking on the "Level Report" button will display a dialog containing level-by-level metrics that summarize the structure of the abstraction network. Additional buttons, which provide specific information for different kinds of abstraction networks, are available depending on the type of abstraction network derived. For example, in partial-area taxonomies the "Area Report" button will display a list of all the areas in the taxonomy along with associated metrics. Selecting an area from the list will focus the taxonomy window on the selected area, enabling fast navigation.

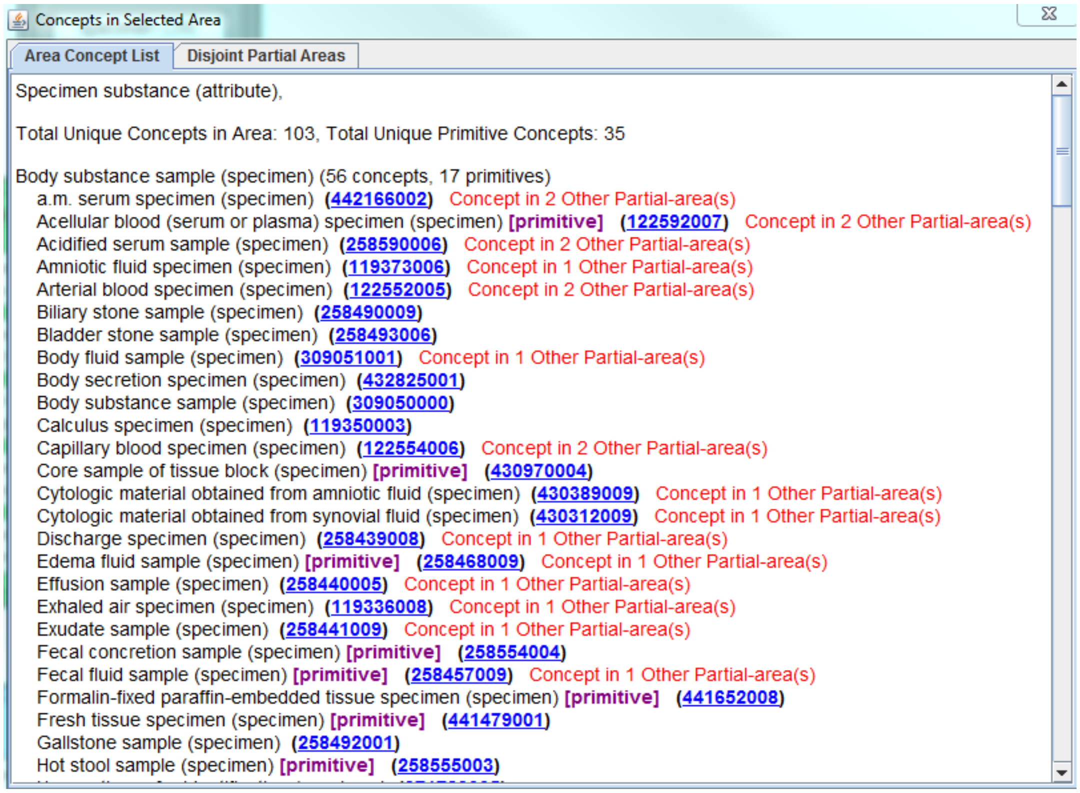
3.5.1.2 Deriving Partial-area Subtaxonomies. The four subtaxonomies described in Section 3.1, relationship subtaxonomies, root subtaxonomies, subject subtaxonomies, and focus subtaxonomies, can be derived in BLUSNO. The option to derive each type of subtaxonomy is made available at appropriate locations in the BLUSNO user interface.

For example, within a taxonomy window a user can create relationship subtaxonomies. Clicking on the "Create Subtaxonomy" button displays a list of attribute relationships defined within the hierarchy. From there, a user can choose which relationship types should be used to derive the relationship subtaxonomy. Alternatively, a user can select an area and then choose to use its set of relationships to define such a subtaxonomy. Root subtaxonomies can be derived by selecting a partial-area in the taxonomy and clicking on the “Derive Root Subtaxonomy” button in the options menu that appears when a partial-area is selected (Figure 3.43). The chosen partial-area will be used as the root partial-area of the root subtaxonomy. Subject subtaxonomies and focus subtaxonomies are derived by selecting a subject concept in BLUSNO’s concept browser (see Section 3.5.1.5).

3.5.1.3 Deriving Disjoint Partial-area Taxonomies. When a partial-area taxonomy is derived in BLUSNO, double clicking on an area in the taxonomy visualization opens the *Area Summary Dialog*, which lists all of the concepts in the selected area according to the area’s partial-areas. This dialog also identifies the total number of unique concepts in the area and highlights overlapping concepts in red. When overlapping partial-areas exist in an area the user can create a disjoint partial-area taxonomy for the area.

Figure 3.45 shows the Area Summary Dialog obtained by double clicking on the {*Substance*} area, located on the right side of Figure 3.43. The relationships of the area are listed first (e.g., *Specimen substance*), followed by the number of unique concepts and how many of the concepts in the area are primitive concepts (103 and 35, respectively).

Partial-areas are listed according to their size (e.g., *Body substance sample* is the largest, thus, it is listed first). This ordering follows the ordering of partial-areas in the graphical interface. Finally, the concepts in each partial-area are listed alphabetically indented underneath each partial-area. The red text “Concept in 2 Other Partial-area(s)” next to the concept *Arterial blood specimen* shows that *Arterial blood specimen* is an overlapping concept.



Specimen substance (attribute),

Total Unique Concepts in Area: 103, Total Unique Primitive Concepts: 35

Body substance sample (specimen) (56 concepts, 17 primitives)
a.m. serum specimen (specimen) (442166002) Concept in 2 Other Partial-area(s)
Acellular blood (serum or plasma) specimen (specimen) [primitive] (122592007) Concept in 2 Other Partial-area(s)
Acidified serum sample (specimen) (258590006) Concept in 2 Other Partial-area(s)
Amniotic fluid specimen (specimen) (119373006) Concept in 1 Other Partial-area(s)
Arterial blood specimen (specimen) (122552005) Concept in 2 Other Partial-area(s)
Biliary stone sample (specimen) (258490009)
Bladder stone sample (specimen) (258493006)
Body fluid sample (specimen) (309051001) Concept in 1 Other Partial-area(s)
Body secretion specimen (specimen) (432825001)
Body substance sample (specimen) (309050000)
Calculus specimen (specimen) (119350003)
Capillary blood specimen (specimen) (122554006) Concept in 2 Other Partial-area(s)
Core sample of tissue block (specimen) [primitive] (430970004)
Cytologic material obtained from amniotic fluid (specimen) (430389009) Concept in 1 Other Partial-area(s)
Cytologic material obtained from synovial fluid (specimen) (430312009) Concept in 1 Other Partial-area(s)
Discharge specimen (specimen) (258439008) Concept in 1 Other Partial-area(s)
Edema fluid sample (specimen) [primitive] (258468009) Concept in 1 Other Partial-area(s)
Effusion sample (specimen) (258440005) Concept in 1 Other Partial-area(s)
Exhaled air specimen (specimen) (119336008) Concept in 1 Other Partial-area(s)
Exudate sample (specimen) (258441009) Concept in 1 Other Partial-area(s)
Fecal concretion sample (specimen) [primitive] (258554004)
Fecal fluid sample (specimen) [primitive] (258457009) Concept in 1 Other Partial-area(s)
Formalin-fixed paraffin-embedded tissue specimen (specimen) [primitive] (441652008)
Fresh tissue specimen (specimen) [primitive] (441479001)
Gallstone sample (specimen) (258492001)
Hot stool sample (specimen) [primitive] (258555003)

Figure 3.45 The Area Summary Dialog for the {*Substance*} area.

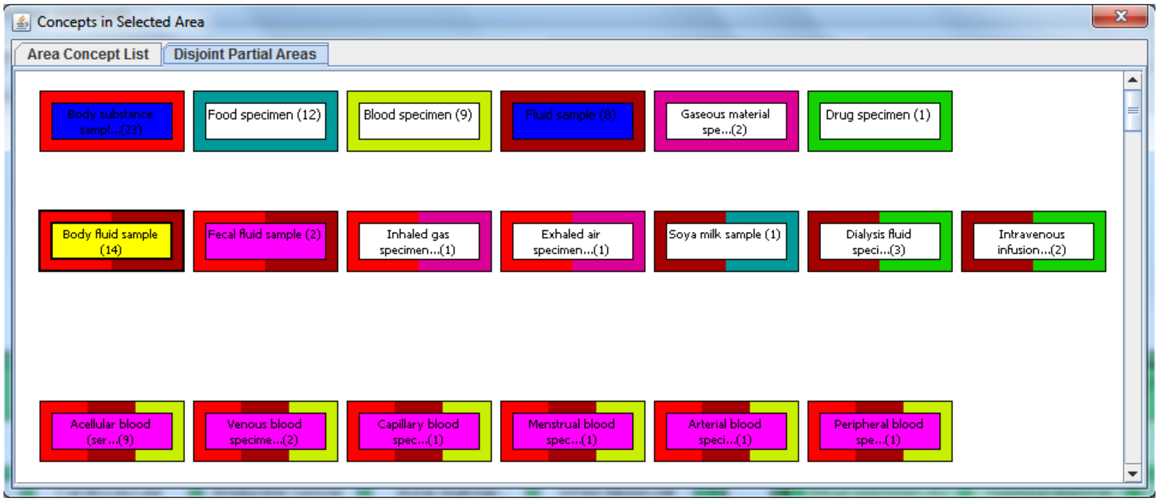

Concepts in Selected Area
Area Concept List
Disjoint Partial Areas
Body substance sampl...(23)
Food specimen (12)
Blood specimen (9)
Fluid sample (8)
Gaseous material spe...(2)
Drug specimen (1)
Body fluid sample (14)
Fecal fluid sample (2)
Inhaled gas specimen...(1)
Exhaled air specimen...(1)
Soya milk sample (1)
Dialysis fluid speci...(3)
Intravenous infusion...(2)
Acellular blood (ser...(9)
Venous blood specime...(2)
Capillary blood spec...(1)
Menstrual blood spec...(1)
Arterial blood speci...(1)
Peripheral blood spe...(1)

Figure 3.46 The disjoint partial-area taxonomy for the area {*Substance*} in the Specimen hierarchy's partial-area taxonomy. The disjoint partial-area *Body fluid sample* (yellow) has been selected and its parents (blue) and children (purple) are highlighted.

The disjoint partial-area taxonomy for an area (shown in Figure 3.46 for {*Substance*}) provides an interactive display for disjoint partial-area taxonomies, modeled after the hand-drawn diagrams designed by Wang et al. in [23] and Figure 2.12. A subset of the functionality available in the overall partial-area taxonomy interface is available in this view. Users can single click on a disjoint partial-area to highlight its parent and child disjoint partial-areas. Likewise, double clicking on a disjoint partial-area displays the *Disjoint Partial-area Summary Dialog*, which includes information similar to that of the Partial-area Summary Dialog.

3.5.1.4 Deriving Tribal Abstraction Networks. Tribal abstraction networks can be derived for complete SNOMED CT hierarchies using the same user interface introduced for deriving partial-area taxonomies (see the "Tribal Abstraction Network" tab in Figure 3.42). However, BLUSNO also enables a user to derive TANs for any singly-rooted subhierarchy of concepts in SNOMED CT. This includes recursively deriving a TAN for a cluster and deriving a TAN for a partial-area in a partial-area taxonomy.

To derive a TAN for a cluster or a partial-area, a user first selects a cluster or partial-area. The user then clicks the "TAN" button that appears in the options menu at the top of the display. This derives a TAN for the subhierarchy of concepts summarized by the cluster or partial-area. Figure 3.47 illustrates a TAN derived for the partial-area *Mass of body structure* in the *Clinical finding* partial-area taxonomy.

Tribal abstraction networks can also be derived for any singly rooted subhierarchy of concepts from the concept browser. In the concept browser the user can choose to derive a TAN rooted at the current focus concept, providing a subject-focused TAN that summarizes the hierarchy of concepts that are specializations of the chosen root.

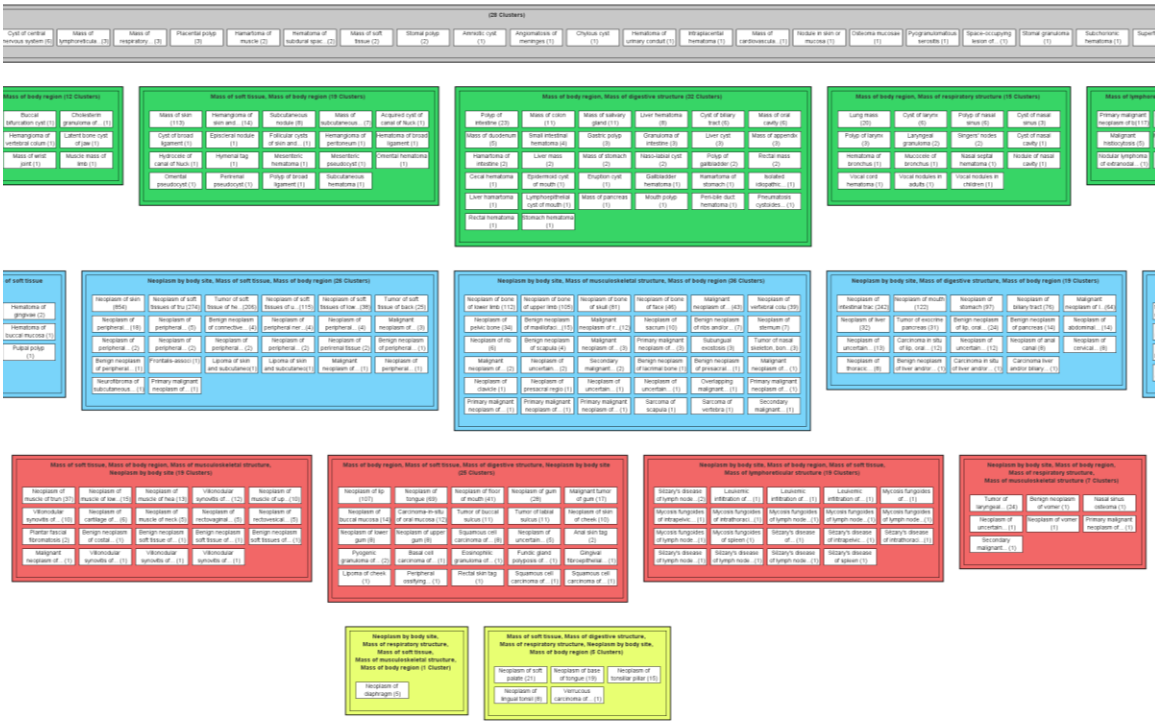
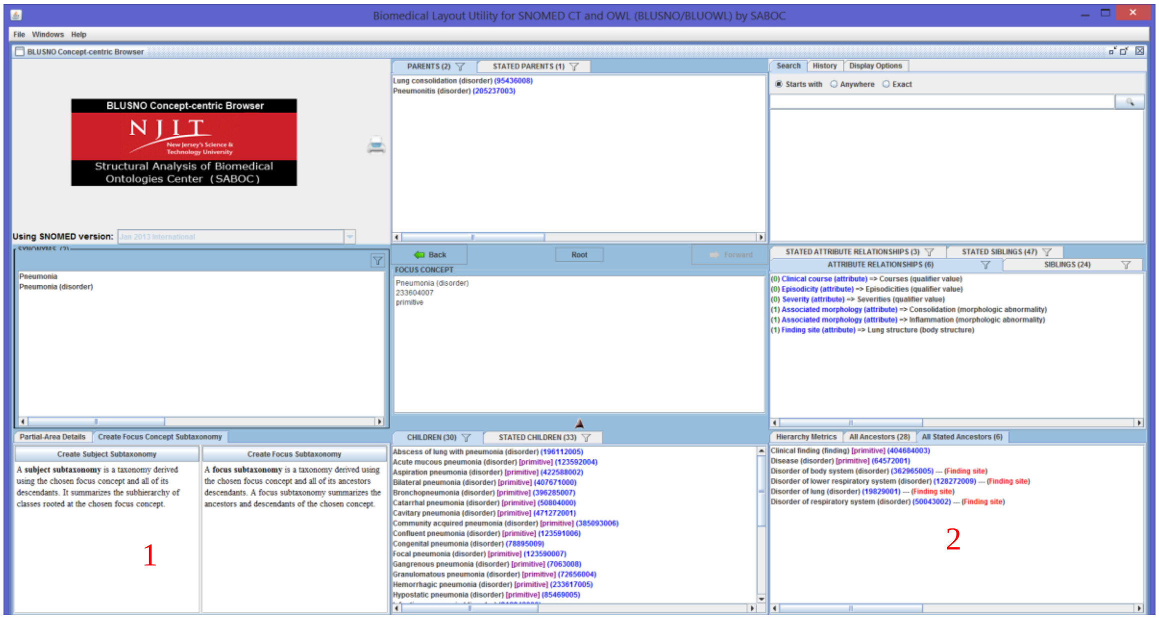

(28 Clusters)
Mass of body region (12 Clusters)
Mass of soft tissue, Mass of body region (19 Clusters)
Mass of body region, Mass of digestive structure (32 Clusters)
Mass of body region, Mass of respiratory structure (15 Clusters)
Neoplasm by body site, Mass of soft tissue, Mass of body region (26 Clusters)
Neoplasm by body site, Mass of musculoskeletal structure, Mass of body region (36 Clusters)
Neoplasm by body site, Mass of digestive structure, Mass of body region (19 Clusters)
Mass of soft tissue, Mass of body region, Mass of musculoskeletal structure, Neoplasm by body site (19 Clusters)
Mass of body region, Mass of soft tissue, Mass of digestive structure, Neoplasm by body site (25 Clusters)
Neoplasm by body site, Mass of body region, Mass of soft tissue, Mass of lymphoreticular structure (19 Clusters)
Neoplasm by body site, Mass of body region, Mass of respiratory structure, Mass of musculoskeletal structure (7 Clusters)
Neoplasm by body site, Mass of respiratory structure, Mass of soft tissue, Mass of musculoskeletal structure, Mass of body region (1 Cluster)
Neoplasm of diaphragm (5)
Mass of soft tissue, Mass of digestive structure, Mass of respiratory structure, Neoplasm by body site, Mass of body region (5 Clusters)
Neoplasm of soft palate (21)
Neoplasm of base of tongue (19)
Neoplasm of tonsillar pillar (15)
Neoplasm of lingual tonsil (8)
Verrucous carcinoma of... (1)

Figure 3.47 Part of a TAN derived for the *Mass of body structure* partial-area in the *Clinical finding* partial-area taxonomy.

3.5.1.5 Hybrid Text-diagram Concept Browser. BLUSNO includes a hybrid text-diagram concept browser based on the previously developed Neighborhood Auditing Tool (NAT) [98] for the UMLS. This browser allows a user to view many details about individual SNOMED CT concepts. BLUSNO's concept browser is unique in that it is directly linked with abstraction network summaries. Additionally, unlike other SNOMED browsers, the BLUSNO browser displays concept information from both the inferred version and stated version of SNOMED, side-by-side.

BLUSNO's concept browser provides a neighborhood view around a selected *focus concept*. Information about the focus concept is displayed relative to where the elements would be in a diagrammatic view, i.e., parents are displayed above the focus concept, children below, siblings and targets of lateral relationships to the side, etc.


Biomedical Layout Utility for SNOMED CT and OWL (BLUSNO/BLUOWL) by SABOC
File
Windows
Help
BLUSNO Concept-centric Browser
BLUSNO Concept-centric Browser
NJIT
New Jersey's Science & Technology University
Structural Analysis of Biomedical Ontologies Center (SABOC)
PARENTS (2)
STATED PARENTS (1)
Search
History
Display Options
Back
Root
Forward
FOCUS CONCEPT
STATED ATTRIBUTE RELATIONSHIPS (3)
STATED SIBLINGS (47)
ATTRIBUTE RELATIONSHIPS (6)
SIBLINGS (24)
SYNONYMS (2)
Partial-Area Details
Create Focus Concept Subtaxonomy
Create Subject Subtaxonomy
Create Focus Subtaxonomy
CHILDREN (30)
STATED CHILDREN (33)
Hierarchy Metrics
All Ancestors (28)
All Stated Ancestors (6)
1
2

Using SNOMED version: Jan 2013 International

● Starts with ○ Anywhere ○ Exact

Lung consolidation (disorder) (95436008)
Pneumonitis (disorder) (205237003)

Pneumonia
Pneumonia (disorder)

Pneumonia (disorder)
233604007
primitive

(0) Clinical course (attribute) => Courses (qualifier value)
(0) Episodicity (attribute) => Episodicities (qualifier value)
(0) Severity (attribute) => Severities (qualifier value)
(1) Associated morphology (attribute) => Consolidation (morphologic abnormality)
(1) Associated morphology (attribute) => Inflammation (morphologic abnormality)
(1) Finding site (attribute) => Lung structure (body structure)

A **subject subtaxonomy** is a taxonomy derived using the chosen focus concept and all of its descendants. It summarizes the subhierarchy of classes rooted at the chosen focus concept.

A **focus subtaxonomy** is a taxonomy derived using the chosen focus concept and all of its ancestors descendants. A focus subtaxonomy summarizes the ancestors and descendants of the chosen concept.

Abscess of lung with pneumonia (disorder) (196112005)
Acute mucous pneumonia (disorder) [primitive] (123592004)
Aspiration pneumonia (disorder) [primitive] (422588002)
Bilateral pneumonia (disorder) [primitive] (407671000)
Bronchopneumonia (disorder) [primitive] (396285007)
Catarrhal pneumonia (disorder) [primitive] (50804000)
Cavitary pneumonia (disorder) [primitive] (471272001)
Community acquired pneumonia (disorder) [primitive] (385093006)
Confluent pneumonia (disorder) [primitive] (123591006)
Congenital pneumonia (disorder) (78895009)
Focal pneumonia (disorder) [primitive] (123590007)
Gangrenous pneumonia (disorder) [primitive] (7063008)
Granulomatous pneumonia (disorder) [primitive] (72656004)
Hemorrhagic pneumonia (disorder) [primitive] (233617005)
Hypostatic pneumonia (disorder) [primitive] (85469005)

Clinical finding (finding) [primitive] (404684003)
Disease (disorder) [primitive] (64572001)
Disorder of body system (disorder) (362965005) --- (Finding site)
Disorder of lower respiratory system (disorder) (128272009) --- (Finding site)
Disorder of lung (disorder) (19829001) --- (Finding site)
Disorder of respiratory system (disorder) (50043002) --- (Finding site)

Figure 3.48 The BLUSNO concept browser with *Pneumonia* as the focus concept.

The user can navigate to different focus concepts by double clicking on any concept in the user interface. Alternatively, the user can search for a new focus concept by its term(s) or its unique concept identifier.

To open a concept browser window, a user can select "Concept Browser" in the main BLUSNO user interface (Figure 3.42). Alternatively, within any BLUSNO dialog a user can click on a concept's unique identifier to view information about the associated concept in a concept browser window. Figure 3.48 shows the concept browser after the concept *Pneumonia* was chosen from the *Clinical finding* hierarchy's taxonomy.

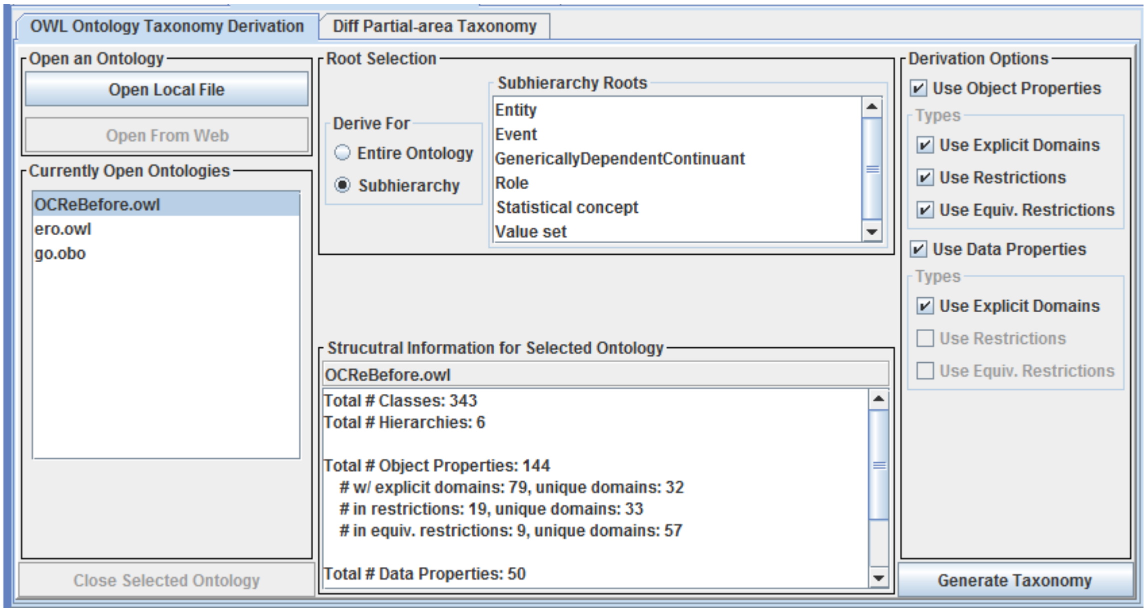
Within the concept browser, the Abstraction Network Panel (labeled 1 in Figure 3.48) identifies which area, partial-area(s), band, and cluster(s) a concept belongs to. From the concept browser the user can choose to view the abstraction network elements associated with the focus concept in an abstraction network window. This functionality allows a user to quickly switch back and forth between viewing a SNOMED CT hierarchy using an abstraction network view, e.g., a taxonomy, and a traditional concept-centric view. Additionally, the Abstraction Network Panel enables the derivation of subject subtaxonomies, focus subtaxonomies, and tribal abstraction networks. The chosen focus concept will be used as the root for each of these abstraction networks.

The BLUSNO concept browser also includes the Hierarchy Metrics Panel (labeled 2 in Figure 3.48), which provides information about the focus concept's position in the overall hierarchy. The "Hierarchy Metrics" tab shows how many ancestors and descendants the focus concept has in both the inferred and stated versions of SNOMED. This panel also includes a list of all of the ancestors of the focus concept, shown in a topological order, for both the inferred and stated version of SNOMED. The introduction point of different attribute relationships is shown in red next to each ancestor. Finally, the hierarchy metrics panel includes a tab that displays all of the focus concepts descendants, listed in topological order.

3.5.2 Biomedical Layout Utility for OWL (BLUOWL)

The Biomedical Layout Utility for the Web Ontology Language (BLUOWL) is a collection of software tools for deriving and visualizing different kinds of abstraction networks for ontologies in Web Ontology Language (OWL) and Open Biological and Biomedical Ontologies (OBO) formats. BLUOWL includes much of the same abstraction network visualization and exploration functionality available in BLUSNO but tailors it to OWL. The major components of BLUOWL will now be described in detail.

3.5.2.1 OWL Taxonomies. BLUOWL is able to derive area taxonomies, partial-area taxonomies, and disjoint partial-area taxonomies using both object properties and data properties. Taxonomies that focus on the different usages of these structural features, e.g., defined domains and restrictions, are can be derived individually or in combinations.



OWL Ontology Taxonomy Derivation | Diff Partial-area Taxonomy

Open an Ontology
Open Local File
Open From Web

Currently Open Ontologies
OCReBefore.owl
ero.owl
go.obo

Close Selected Ontology

Root Selection
Derive For
○ Entire Ontology
● Subhierarchy

Subhierarchy Roots
Entity
Event
GenericallyDependentContinuant
Role
Statistical concept
Value set

Strucutral Information for Selected Ontology
OCReBefore.owl
Total # Classes: 343
Total # Hierarchies: 6

Total # Object Properties: 144
# w/ explicit domains: 79, unique domains: 32
# in restrictions: 19, unique domains: 33
# in equiv. restrictions: 9, unique domains: 57

Total # Data Properties: 50

Derivation Options
☑ Use Object Properties
Types
☑ Use Explicit Domains
☑ Use Restrictions
☑ Use Equiv. Restrictions
☑ Use Data Properties
Types
☑ Use Explicit Domains
□ Use Restrictions
□ Use Equiv. Restrictions

Generate Taxonomy

Figure 3.49 The BLUOWL taxonomy derivation user interface.

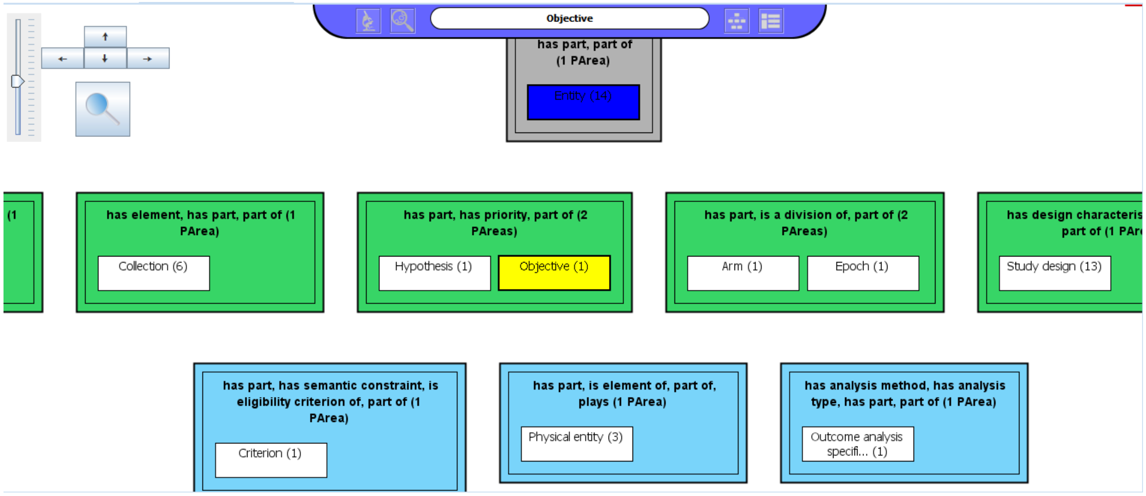

Objective
has part, part of
(1 PArea)
Entity (14)
(1
has element, has part, part of (1 PArea)
Collection (6)
has part, has priority, part of (2 PAreas)
Hypothesis (1)
Objective (1)
has part, is a division of, part of (2 PAreas)
Arm (1)
Epoch (1)
has design characteris
part of (1 PAr
Study design (13)
has part, has semantic constraint, is eligibility criterion of, part of (1 PArea)
Criterion (1)
has part, is element of, part of, plays (1 PArea)
Physical entity (3)
has analysis method, has analysis type, has part, part of (1 PArea)
Outcome analysis specifi... (1)

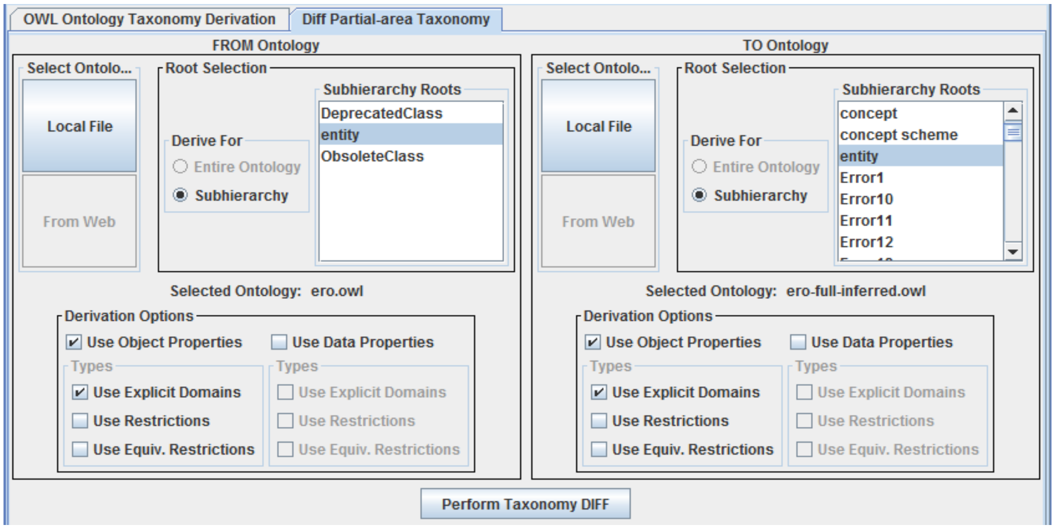
Figure 3.50 A domain-defined partial-area taxonomy derived for OCRe's *Entity* hierarchy in BLUOWL.

Figure 3.49 shows BLUOWL's taxonomy derivation user interface. A user can open one or more ontologies (e.g., OCRe, ERO, and GO in Figure 3.49). Selecting an ontology from the "Currently Open Ontologies" list (left side of Figure 3.49), will display various information about the ontology, including metrics for the different structural features and their usages. OCRe has been selected in Figure 3.49. Derivation options are enabled and disabled based on the structure of the selected ontology, e.g., if the ontology has no data properties it will not be possible to derive a taxonomy using data properties. To derive a taxonomy, the user selects an ontology and chooses a root class or chooses to derive a taxonomy for the entire ontology (implicitly using OWL:Thing as the root). Next, the user selects which structural features to use in the derivation. Finally, the user clicks the "Generate Taxonomy" button, which will derive the taxonomy and display its visualization.

Figure 3.50 shows an example of a taxonomy derived for the Ontology of Clinical Research (OCRe) *Entity* hierarchy. The dynamic visualization provided by BLUOWL includes much of the functionality available in BLUSNO. This includes the derivation of disjoint partial-area taxonomies when there is an area with overlapping partial-areas. The various displays and dialogs will indicate which structural features are associated with a taxonomy element. For example, areas may be defined by a combination of object properties and data properties. Selecting a partial-area within such an area will display the type and usage of each property, as well as if the property is introduced or inherited.

3.5.2.2 Diff Taxonomies. BLUOWL also includes the ability to derive and visualize diff partial-area taxonomies. To derive a diff partial-area taxonomy a user selects the "Diff Partial-area Taxonomy" tab shown in Figure 3.49. Figure 3.51 shows the diff partial-area taxonomy derivation user interface. To derive a diff partial-area taxonomy, a user selects the "from" and "to" ontologies. Next, the root of the ontology for the from and to ontologies is selected and the structural features used to derive the from and to taxonomies are chosen. If a user chooses the same ontology as both the from ontology and to ontology then the diff taxonomy can be used to create a *granularity diff*, which identifies how different ontology elements are summarized by taxonomies derived using different structural features (see Section 4.4). Finally, to view the diff taxonomy the user clicks on "Perform Taxonomy Diff."

Figure 3.52 provides an example of a domain-defined diff partial-area taxonomy for the eagle-I Research Resource Ontology (ERO), derived using the ERO releases described in Section 3.4.3.3. The visualization scheme for diff taxonomy elements follows the one described in Section 3.4.1.1 and 3.4.1.2: red for removed, green for introduced, and yellow for modified. Like the BLUSNO and regular OWL taxonomy visualizations, the diff taxonomy produced by BLUOWL is interactive.



FROM Ontology

Derive For: ○ Entire Ontology ● Subhierarchy

Subhierarchy Roots: DeprecatedClass, entity (selected), ObsoleteClass

Selected Ontology: ero.owl

☑ Use Object Properties — Types: ☑ Use Explicit Domains, ☐ Use Restrictions, ☐ Use Equiv. Restrictions

☐ Use Data Properties — Types: ☐ Use Explicit Domains, ☐ Use Restrictions, ☐ Use Equiv. Restrictions

TO Ontology

Derive For: ○ Entire Ontology ● Subhierarchy

Subhierarchy Roots: concept, concept scheme, entity (selected), Error1, Error10, Error11, Error12

Selected Ontology: ero-full-inferred.owl

☑ Use Object Properties — Types: ☑ Use Explicit Domains, ☐ Use Restrictions, ☐ Use Equiv. Restrictions

☐ Use Data Properties — Types: ☐ Use Explicit Domains, ☐ Use Restrictions, ☐ Use Equiv. Restrictions

Figure 3.51 The diff partial-area taxonomy derivation user interface.

When selected, each diff taxonomy element, e.g., diff areas and diff partial-areas, provides information about the underlying structural changes that are summarized by the diff taxonomy element. For example, double clicking on a diff partial-area will display a dialog with information about how the selected diff partial-area's classes changed between releases.

Figure 3.53 shows the dialog that is displayed when the removed partial-area *processual entity* is selected. This dialog, similar to the Partial-area Summary Dialog shown in Figure 3.44, provides information about why the *processual entity* partial-area was removed between the two releases. By looking at this dialog, a user can quickly determine what happened to the classes that were summarized by *processual entity* and what structural changes lead to the changes.

In the "Diff Details" tab, a summary of the changes that led to this partial-area being removed is provided. Additional structural information, such as the parent partial-areas and child partial-areas, is also displayed.

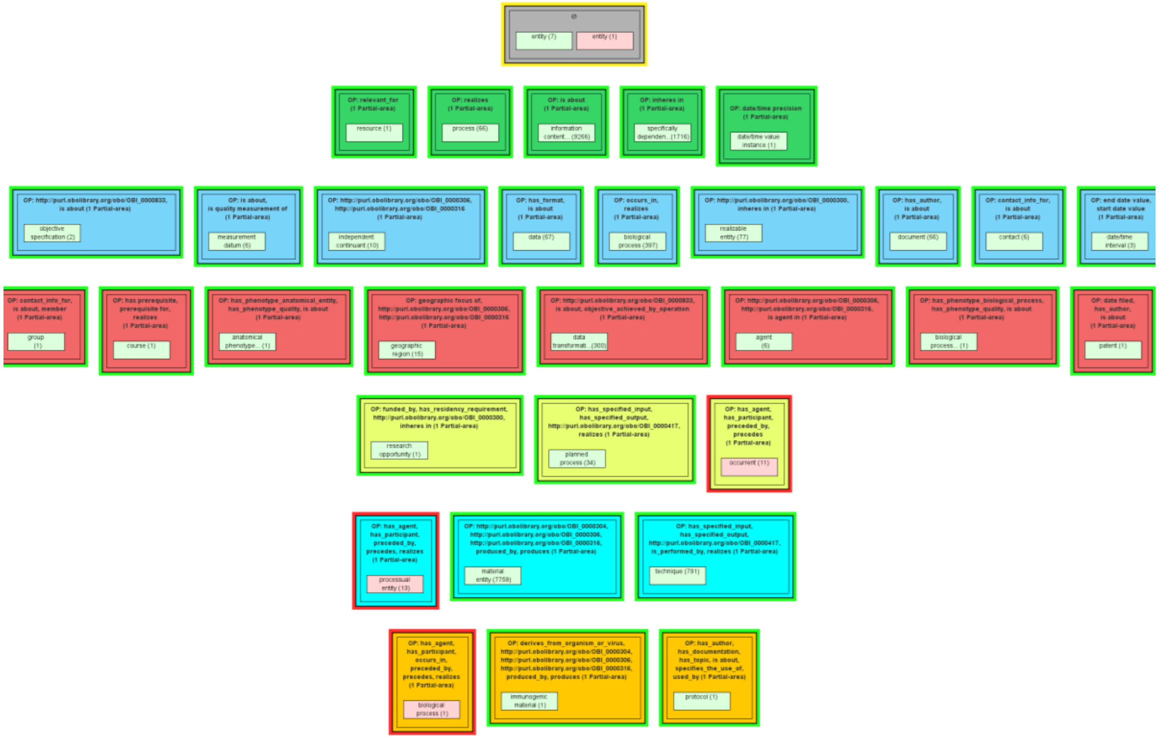
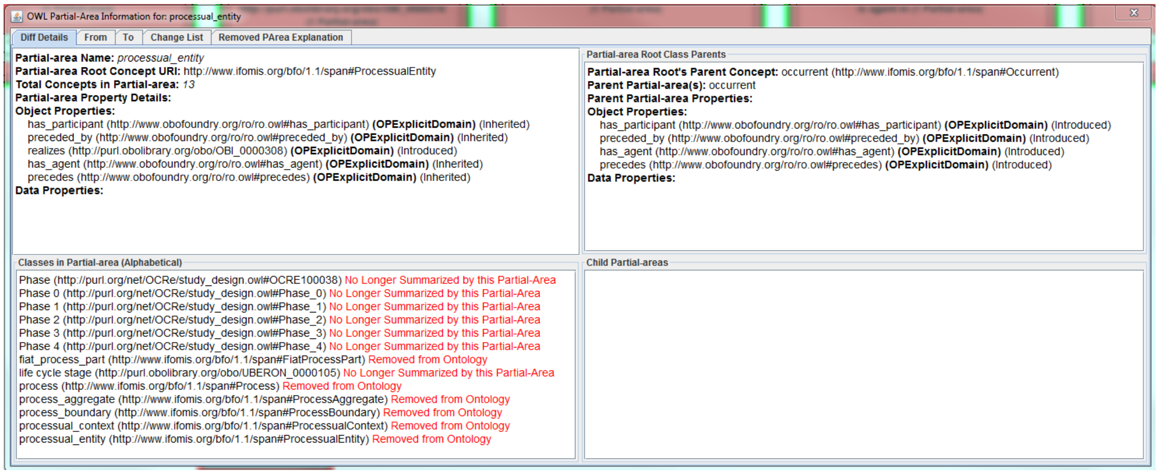

∅
entity (7)
entity (1)
OP: relevant_for (1 Partial-area)
resource (1)
OP: realizes (1 Partial-area)
process (66)
OP: is about (1 Partial-area)
information content... (9266)
OP: inheres in (1 Partial-area)
specifically dependen...(1716)
OP: date/time precision (1 Partial-area)
date/time value instance (1)
OP: http://purl.obolibrary.org/obo/OBI_0000833, is about (1 Partial-area)
objective specification (2)
OP: is about, is quality measurement of (1 Partial-area)
measurement datum (6)
OP: http://purl.obolibrary.org/obo/OBI_0000306, http://purl.obolibrary.org/obo/OBI_0000316 (1 Partial-area)
independent continuant (10)
OP: has_format, is about (1 Partial-area)
data (67)
OP: occurs_in, realizes (1 Partial-area)
biological process (397)
OP: http://purl.obolibrary.org/obo/OBI_0000300, inheres in (1 Partial-area)
realizable entity (77)
OP: has_author, is about (1 Partial-area)
document (66)
OP: contact_info_for, is about (1 Partial-area)
contact (6)
OP: end date value, start date value (1 Partial-area)
date/time interval (3)
OP: contact_info_for, is about, member (1 Partial-area)
group (1)
OP: has prerequisite, prerequisite for, realizes (1 Partial-area)
course (1)
OP: has_phenotype_anatomical_entity, has_phenotype_quality, is about (1 Partial-area)
anatomical phenotype... (1)
OP: geographic focus of, http://purl.obolibrary.org/obo/OBI_0000306, http://purl.obolibrary.org/obo/OBI_0000316 (1 Partial-area)
geographic region (15)
OP: http://purl.obolibrary.org/obo/OBI_0000833, is about, objective_achieved_by_operation (1 Partial-area)
data transformati...(300)
OP: http://purl.obolibrary.org/obo/OBI_0000306, http://purl.obolibrary.org/obo/OBI_0000316, is agent in (1 Partial-area)
agent (5)
OP: has_phenotype_biological_process, has_phenotype_quality, is about (1 Partial-area)
biological process... (1)
OP: date filed, has_author, is about (1 Partial-area)
patent (1)
OP: funded_by, has_residency_requirement, http://purl.obolibrary.org/obo/OBI_0000300, inheres in (1 Partial-area)
research opportunity (1)
OP: has_specified_input, has_specified_output, http://purl.obolibrary.org/obo/OBI_0000417, realizes (1 Partial-area)
planned process (34)
OP: has_agent, has_participant, preceded_by, precedes (1 Partial-area)
occurrent (11)
OP: has_agent, has_participant, preceded_by, precedes, realizes (1 Partial-area)
processual entity (13)
OP: http://purl.obolibrary.org/obo/OBI_0000304, http://purl.obolibrary.org/obo/OBI_0000306, http://purl.obolibrary.org/obo/OBI_0000316, produced_by, produces (1 Partial-area)
material entity (7758)
OP: has_specified_input, has_specified_output, http://purl.obolibrary.org/obo/OBI_0000417, is_performed_by, realizes (1 Partial-area)
technique (791)
OP: has_agent, has_participant, occurs_in, preceded_by, precedes, realizes (1 Partial-area)
biological process (1)
OP: derives_from_organism_or_virus, http://purl.obolibrary.org/obo/OBI_0000304, http://purl.obolibrary.org/obo/OBI_0000306, http://purl.obolibrary.org/obo/OBI_0000316, produced_by, produces (1 Partial-area)
immunogenic material (1)
OP: has_author, has_documentation, has_topic, is about, specifies_the_use_of, used_by (1 Partial-area)
protocol (1)

Figure 3.52 The top seven levels of the diff partial-area taxonomy for the eagle-I Research Resource Ontology (ERO), as shown in BLUOWL.


OWL Partial-Area Information for: processual_entity
Diff Details
From
To
Change List
Removed PArea Explanation
Partial-area Root Class Parents
Classes in Partial-area (Alphabetical)
Child Partial-areas

Partial-area Name: *processual_entity*
Partial-area Root Concept URI: http://www.ifomis.org/bfo/1.1/span#ProcessualEntity
Total Concepts in Partial-area: *13*
Partial-area Property Details:
Object Properties:
has_participant (http://www.obofoundry.org/ro/ro.owl#has_participant) **(OPExplicitDomain)** (Inherited)
preceded_by (http://www.obofoundry.org/ro/ro.owl#preceded_by) **(OPExplicitDomain)** (Inherited)
realizes (http://purl.obolibrary.org/obo/OBI_0000308) **(OPExplicitDomain)** (Introduced)
has_agent (http://www.obofoundry.org/ro/ro.owl#has_agent) **(OPExplicitDomain)** (Inherited)
precedes (http://www.obofoundry.org/ro/ro.owl#precedes) **(OPExplicitDomain)** (Inherited)
Data Properties:

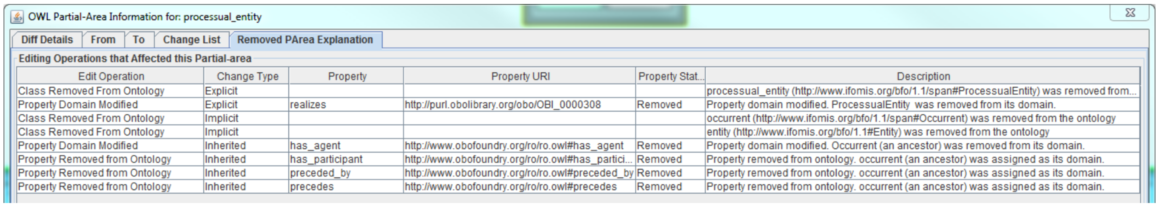
Partial-area Root's Parent Concept: occurrent (http://www.ifomis.org/bfo/1.1/span#Occurrent)
Parent Partial-area(s): occurrent
Parent Partial-area Properties:
Object Properties:
has_participant (http://www.obofoundry.org/ro/ro.owl#has_participant) **(OPExplicitDomain)** (Introduced)
preceded_by (http://www.obofoundry.org/ro/ro.owl#preceded_by) **(OPExplicitDomain)** (Introduced)
has_agent (http://www.obofoundry.org/ro/ro.owl#has_agent) **(OPExplicitDomain)** (Introduced)
precedes (http://www.obofoundry.org/ro/ro.owl#precedes) **(OPExplicitDomain)** (Introduced)
Data Properties:

Phase (http://purl.org/net/OCRe/study_design.owl#OCRE100038) No Longer Summarized by this Partial-Area
Phase 0 (http://purl.org/net/OCRe/study_design.owl#Phase_0) No Longer Summarized by this Partial-Area
Phase 1 (http://purl.org/net/OCRe/study_design.owl#Phase_1) No Longer Summarized by this Partial-Area
Phase 2 (http://purl.org/net/OCRe/study_design.owl#Phase_2) No Longer Summarized by this Partial-Area
Phase 3 (http://purl.org/net/OCRe/study_design.owl#Phase_3) No Longer Summarized by this Partial-Area
Phase 4 (http://purl.org/net/OCRe/study_design.owl#Phase_4) No Longer Summarized by this Partial-Area
fiat_process_part (http://www.ifomis.org/bfo/1.1/span#FiatProcessPart) Removed from Ontology
life cycle stage (http://purl.obolibrary.org/obo/UBERON_0000105) No Longer Summarized by this Partial-Area
process (http://www.ifomis.org/bfo/1.1/span#Process) Removed from Ontology
process_aggregate (http://www.ifomis.org/bfo/1.1/span#ProcessAggregate) Removed from Ontology
process_boundary (http://www.ifomis.org/bfo/1.1/span#ProcessBoundary) Removed from Ontology
processual_context (http://www.ifomis.org/bfo/1.1/span#ProcessualContext) Removed from Ontology
processual_entity (http://www.ifomis.org/bfo/1.1/span#ProcessualEntity) Removed from Ontology

Figure 3.53 The diff partial-area summary dialog for the removed partial-area *processual entity*.

The list of classes formerly summarized by the removed partial-area provides information about what happened to each class between the two releases. For example, the root class of the partial-area, *processual entity*, was removed from the ontology (as indicated by red text next to the class name) and the class *Phase* is no longer summarized by this partial-area in the "to" taxonomy.

The "From" and "To" tabs show the details of the partial-area in the context of the the *from* taxonomy and *to* taxonomy, respectively. The "Change List" tab provides a summary of the changes to the set of classes summarized by the partial-area. It lists the changes by type (e.g., *removed from ontology* and *no longer summarized by this partial-area*) and gives a full explanation of what happened to the class. For example, for the class *Phase* it indicates that it is no longer summarized by *processual entity* in the to taxonomy, but is now summarized by the partial-area *process* in the area {*realizes*}.



OWL Partial-Area Information for: processual_entity
Diff Details
From
To
Change List
Removed PArea Explanation

Editing Operations that Affected this Partial-area

Edit Operation	Change Type	Property	Property URI	Property Stat...	Description
Class Removed From Ontology	Explicit				processual_entity (http://www.ifomis.org/bfo/1.1/span#ProcessualEntity) was removed from...
Property Domain Modified	Explicit	realizes	http://purl.obolibrary.org/obo/OBI_0000308	Removed	Property domain modified. ProcessualEntity was removed from its domain.
Class Removed From Ontology	Implicit				occurrent (http://www.ifomis.org/bfo/1.1/span#Occurrent) was removed from the ontology
Class Removed From Ontology	Implicit				entity (http://www.ifomis.org/bfo/1.1#Entity) was removed from the ontology
Property Domain Modified	Inherited	has_agent	http://www.obofoundry.org/ro/ro.owl#has_agent	Removed	Property domain modified. Occurrent (an ancestor) was removed from its domain.
Property Removed from Ontology	Inherited	has_participant	http://www.obofoundry.org/ro/ro.owl#has_partici...	Removed	Property removed from ontology. occurrent (an ancestor) was assigned as its domain.
Property Removed from Ontology	Inherited	preceded_by	http://www.obofoundry.org/ro/ro.owl#preceded_by	Removed	Property removed from ontology. occurrent (an ancestor) was assigned as its domain.
Property Removed from Ontology	Inherited	precedes	http://www.obofoundry.org/ro/ro.owl#precedes	Removed	Property removed from ontology. occurrent (an ancestor) was assigned as its domain.

Figure 3.54 The diff partial-area explanation tab.

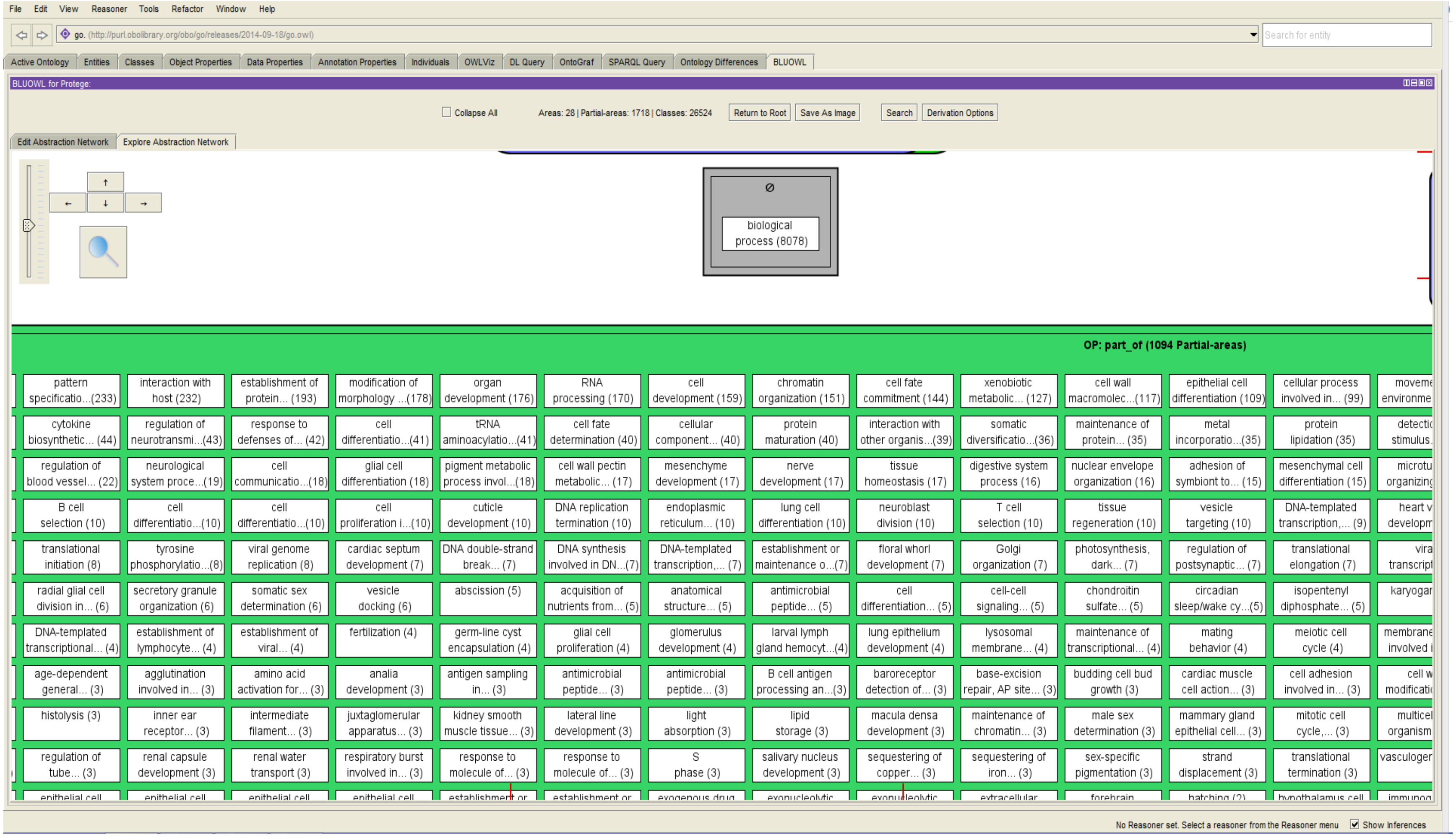
The last tab, "Removed PArea Explanation" (named "Introduced PArea Explanation" when an introduced partial-area is selected, rather than a removed partial-area), provides a list of structural changes (i.e., editing operations) that lead to partial-area being removed, e.g., modifications to classes and properties. The display indicates if each change directly affected the partial-area or if the change implicitly affected the partial-area's classes through inheritance.

Figure 3.54 shows the Removed PArea Explanation tab for the *processual entity* partial-area. A total of eight structural changes lead to *processual entity* being removed. Two of these changes, the removal of the class *processual entity* from the ontology and the modification of the *realizes* object property domain, directly affected the partial-area. The six other changes implicitly affected the partial-area. For example, three object properties were removed from the ontology and their domain (*occurent*) was an ancestor of *processual entity*.

3.5.2.3 Protégé Plugin. Protégé [56], with over 200,000 users, is one of the most widely used ontology development tools. Protégé is designed to be extendable via community-developed plugins [152]. The BLUOWL Protégé Plugin integrates BLUOWL with Protégé, enabling the use of OWL abstraction networks during the ontology development process. The BLUOWL plugin allows a Protégé user to seamlessly transition between the standard Protégé user interface, which includes class and property definitions, and the BLUOWL abstraction network interface, which captures a structural summary of the ontology.

The BLUOWL Protégé Plugin, shown in Figure 3.55 with the Gene Ontology's *Biological process* taxonomy, includes all of the functionality of the standalone BLUOWL tool and includes Protégé-specific features. For example, clicking on classes or properties within BLUOWL's dialogs will show the definition of the selected element in Protégé. Furthermore, if a user selects an entity in Protégé, it is highlighted within the taxonomy.

The Protégé plugin can also derive diff taxonomies using the currently opened ontology and another ontology that the user chooses. Alternatively, the BLUOWL plugin can derive diff taxonomies "on the fly" as a user is editing an ontology. This allows an ontology editor to visualize the global impact of an editing operation as it is applied to the ontology.



go. (http://purl.obolibrary.org/obo/go/releases/2014-09-18/go.owl)

Areas: 28 | Partial-areas: 1718 | Classes: 26524

pattern specificatio...(233)	interaction with host (232)	establishment of protein... (193)	modification of morphology ...(178)	organ development (176)	RNA processing (170)	cell development (159)	chromatin organization (151)	cell fate commitment (144)	xenobiotic metabolic... (127)	cell wall macromolec...(117)	epithelial cell differentiation (109)	cellular process involved in... (99)	moveme environme
cytokine biosynthetic... (44)	regulation of neurotransmi...(43)	response to defenses of... (42)	cell differentiatio...(41)	tRNA aminoacylatio...(41)	cell fate determination (40)	cellular component... (40)	protein maturation (40)	interaction with other organis...(39)	somatic diversificatio...(36)	maintenance of protein... (35)	metal incorporatio...(35)	protein lipidation (35)	detectic stimulus.
regulation of blood vessel... (22)	neurological system proce...(19)	cell communicatio...(18)	glial cell differentiation (18)	pigment metabolic process invol...(18)	cell wall pectin metabolic... (17)	mesenchyme development (17)	nerve development (17)	tissue homeostasis (17)	digestive system process (16)	nuclear envelope organization (16)	adhesion of symbiont to... (15)	mesenchymal cell differentiation (15)	microtu organizing
B cell selection (10)	cell differentiatio...(10)	cell differentiatio...(10)	cell proliferation i...(10)	cuticle development (10)	DNA replication termination (10)	endoplasmic reticulum... (10)	lung cell differentiation (10)	neuroblast division (10)	T cell selection (10)	tissue regeneration (10)	vesicle targeting (10)	DNA-templated transcription,... (9)	heart v developm
translational initiation (8)	tyrosine phosphorylatio...(8)	viral genome replication (8)	cardiac septum development (7)	DNA double-strand break... (7)	DNA synthesis involved in DN...(7)	DNA-templated transcription,... (7)	establishment or maintenance o...(7)	floral whorl development (7)	Golgi organization (7)	photosynthesis, dark... (7)	regulation of postsynaptic... (7)	translational elongation (7)	vira transcript
radial glial cell division in... (6)	secretory granule organization (6)	somatic sex determination (6)	vesicle docking (6)	abscission (5)	acquisition of nutrients from... (5)	anatomical structure... (5)	antimicrobial peptide... (5)	cell differentiation... (5)	cell-cell signaling... (5)	chondroitin sulfate... (5)	circadian sleep/wake cy...(5)	isopentenyl diphosphate... (5)	karyogar
DNA-templated transcriptional... (4)	establishment of lymphocyte... (4)	establishment of viral... (4)	fertilization (4)	germ-line cyst encapsulation (4)	glial cell proliferation (4)	glomerulus development (4)	larval lymph gland hemocyt...(4)	lung epithelium development (4)	lysosomal membrane... (4)	maintenance of transcriptional... (4)	mating behavior (4)	meiotic cell cycle (4)	membrane involved i
age-dependent general... (3)	agglutination involved in... (3)	amino acid activation for... (3)	analia development (3)	antigen sampling in... (3)	antimicrobial peptide... (3)	antimicrobial peptide... (3)	B cell antigen processing an...(3)	baroreceptor detection of... (3)	base-excision repair, AP site... (3)	budding cell bud growth (3)	cardiac muscle cell action... (3)	cell adhesion involved in... (3)	cell w modificatio
histolysis (3)	inner ear receptor... (3)	intermediate filament... (3)	juxtaglomerular apparatus... (3)	kidney smooth muscle tissue... (3)	lateral line development (3)	light absorption (3)	lipid storage (3)	macula densa development (3)	maintenance of chromatin... (3)	male sex determination (3)	mammary gland epithelial cell... (3)	mitotic cell cycle,... (3)	multicel organism
regulation of tube... (3)	renal capsule development (3)	renal water transport (3)	respiratory burst involved in... (3)	response to molecule of... (3)	response to molecule of... (3)	S phase (3)	salivary nucleus development (3)	sequestering of copper... (3)	sequestering of iron... (3)	sex-specific pigmentation (3)	strand displacement (3)	translational termination (3)	vasculoger
epithelial cell	epithelial cell	epithelial cell	epithelial cell	establishment or	establishment or	exogenous drug	exonucleolytic	exonucleolytic	extracellular	forebrain	hatching (2)	hypothalamus cell	immunog

Figure 3.55 A segment of the Gene Ontology's *Biological process* partial-area taxonomy, as viewed in the BLUOWL Protégé Plugin.

CHAPTER 4

FUTURE WORK

4.1 Subtaxonomies for Large SNOMED CT Hierarchies

The development of various kinds of subtaxonomies has further enabled the scalability of taxonomy-based quality assurance. In particular, subject subtaxonomies allow auditors to obtain a summary of the concepts in a specified subject area. For the *Bleeding* subhierarchy, certain groups of concepts within a subject subtaxonomy (i.e., overlapping concepts) were found to be statistically more likely to be erroneous than other groups (i.e., non-overlapping concepts). However, there are several important issues that will be investigated in future studies.

First, additional subject subtaxonomy quality assurance studies will be performed to verify the results of the *Bleeding* study described in Section 3.1.3.2. These studies will investigate the error rates of overlapping concepts in subject subtaxonomies for other important subject areas, e.g., Infectious diseases and Cancer. A related issue that will be investigated is overlapping concepts in a subject subtaxonomy that are not overlapping concepts in a complete taxonomy. This phenomenon occurs when partial-areas exists in a subject subtaxonomy but do not exist in the taxonomy for a complete hierarchy, e.g., the yellow partial-areas of the *Cancer* subtaxonomy (Figure 3.10).

In the *Cancer* subject subtaxonomy there are 2,398 overlapping concepts in {*Associated morphology, Finding site*}. However, only a small number of these concepts are overlapping concepts in the complete *Clinical finding* taxonomy. Most of these concepts are only in the large *Mass of body structure* partial-area in the complete *Clinical finding* taxonomy. This situation complicates the definition of an overlapping concept, thus, the error rates for these concepts will have to be investigated. The error rates of these concepts will be compared to concepts that overlap in the complete taxonomy and concepts that do not overlap in either the complete taxonomy or in the subject subtaxonomy.

Other groups of concepts that have been shown to exhibit higher error rates, e.g., concepts in small partial-areas [26, 28], will be investigated within subject subtaxonomies. To determine the effectiveness of reviewing these groups, versus reviewing overlapping concepts, the error rates of these concepts will be compared to those of overlapping concepts.

Additionally, partial-areas that only exist in a subject subtaxonomy will contain concepts that are summarized by different partial-area(s) in a complete taxonomy (e.g., the concepts in the *Cancer* subtaxonomy's unique partial-areas are mostly summarized by only *Mass of body structure* in the complete *Clinical finding* taxonomy). Thus, in a subject subtaxonomy a concept may be summarized by a "small" partial-area, while in the complete taxonomy it is summarized by a "large" partial-area. A sample of these concepts will be reviewed to determine their error rate.

A significant issue with the taxonomy-based quality assurance methodology was raised by James T. Case, the head of the US Extension of SNOMED CT. All previous SNOMED CT quality assurance studies, e.g., [26, 28, 29, 119], have used the inferred version of SNOMED CT. The inferred version of SNOMED CT is created by applying a reasoner on the stated version of SNOMED CT, which only includes the relationships defined by SNOMED's editors. Often, the domain experts in previous studies would correctly identify that concepts were erroneous, but the cause of the error and suggested solutions were incorrect. This occurred because the domain experts were reviewing and basing corrections off of the inferred version of SNOMED CT, while SNOMED CT's editors make changes to the stated version. The domain experts were generally unfamiliar with the stated version or SNOMED's concept model. Thus, the erroneous concepts had to be investigated further by the US Extension Center.

In a future study, domain experts will be familiarized with the SNOMED concept model and will be provided with the stated release of SNOMED CT. The results of this study will be compared to a study where domain experts are not provided with this information. It is hypothesized that the additional information will enable the identification of more errors and will allow the domain experts to provide better suggested corrections.

One important aspect of taxonomy-based quality assurance that will be investigated is the identification of concepts with a higher likelihood of errors of commission or errors of omission. In general, errors of commission, e.g., an incorrect parent or incorrect relationship, are considered more critical than errors of omission, e.g., a missing parent or missing relationship. In this future study the type of each concept error (omission, commission) will be identified. The goal will be to determine if taxonomy elements (i.e., partial-areas and disjoint partial-areas) with certain properties (i.e., small, overlapping, or both) are more likely to have errors of commission or errors of omission. This error type analysis will need to be based on errors identified in the stated version of SNOMED CT, as opposed to the inferred version. For example, a missing parent in the inferred version may be due to an incorrect relationship in the stated version. If concepts summarized by certain taxonomic elements are more likely to have an error of commission then quality assurance efforts should be focused on those elements.

Finally, the summarization aspect of subject subtaxonomies will be investigated. Several subject subtaxonomies will be reviewed to determine if they provide an accurate and useful summary of a subject area. A group of domain experts and SNOMED users will review each subject subtaxonomy and provide their feedback on how well (in terms of accuracy and utility) the subtaxonomy summarizes the subject area. This information will be used to guide the development of additional types of SNOMED abstraction networks that can be applied to various use cases.

4.2 Tribal Abstraction Network

There are several important open research questions for the Tribal Abstraction Network. One important issue, introduced in Section 3.2.3 is the relatively low number of errors uncovered in the quality assurance review of the *Observable entity* hierarchy. To determine why the error rate was so low TAN-based studies will be conducted for other target hierarchies of SNOMED CT, e.g., *Body structure* and *Substance*. Additionally, TANs will be derived for the root partial-areas of the *Procedure* and *Clinical finding* hierarchies. The error rates of the concepts in these TANs will be compared to the error rates of the *Observable entity* hierarchy and the error rates found based on partial-area taxonomies (e.g., overlapping concepts, small partial-areas).

Another significant issue is the emergence of disproportionately large clusters (“super-large clusters,” for short) that summarize thousands, or tens of thousands, of concepts. These clusters represent an over summarization of a set of concepts. As discussed in Section 3.2, the number of concepts with multiple parents is not as important in deriving a TAN as the locations where the concepts with multiple parents appear in a hierarchy.

The placement of such concepts may lead to the emergence of super-large clusters, such as *Clinical history/examination observable* (4138) and *Function* (1384) in the *Observable entity* TAN's first level (see Figure 3.18), containing a relatively large number of concepts, and together containing 67% (=5522/8231) of the *Observable entity* hierarchy. Following *Function*, the next largest cluster is *Social / personal history observable* (300), an order of magnitude smaller. Such super-large clusters may appear anywhere in a TAN, not necessarily just at Level 1

Super-large clusters represent an over summarization of the terminology's hierarchy and are of too coarse a granularity [120]. To address this problem, two algorithmic methods for summarizing the concepts in super-large clusters will be investigated. The first method is the *Recursive TAN*, which derives a TAN for only the concepts in a chosen super-large cluster. Concepts in any cluster many have multiple parents in the same cluster. Thus, by recursively applying the TAN derivation methodology on a super-large cluster, its content can be summarized. The recursive approach will work by using the root of a super-large cluster as a hierarchy root for a TAN. The children of the cluster root are then defined as tribal patriarchs. The TAN derivation methodology is then recursively applied to the concepts in the super-large cluster. The resulting Recursive TAN would summarize the concepts in the super-large cluster.

For the case of *Observable entity*, the question is whether it is feasible to derive a recursive TAN for the *Clinical history/examination observable* and *Function* clusters. A study will be conducted where a Recursive TAN is derived for these super-large clusters. A review of concepts in the recursively derived TANs will be performed and a methodology for quality assurance that utilizes Recursive TANs will be developed. The hypotheses for this quality assurance methodology mirror those of the complete TAN: concepts in large clusters of a recursively derived TAN will have more errors than concepts in small clusters and concepts at higher numbered levels (towards the bottom) of a Recursive TAN will have more errors than concepts at lower numbered levels.

Another method for summarizing super-large clusters, called the *Expanded TAN*, is limited to Level 1, e.g., super-large clusters containing concepts which are descendants of only one patriarch. One can define a TAN that uses the children of a super-large cluster's root (i.e., a subset of the grandchildren of the hierarchy root) as tribal patriarchs. Using this method, a more refined summary of a hierarchy is obtained. Unlike the Recursive TAN, which summarizes only the concepts in a single super-large cluster, the Expanded TAN summarizes the concepts of a super-large Level 1 cluster in the context of the entire hierarchy. One or more super-large Level 1 clusters could be split into several smaller clusters according to their children.

For example, in Figure 4.1 the *Function* cluster on Level 1 is split into two Level 1 clusters based on its two children: *Breast function* and *Digestive system function*. One potential drawback of the Expanded TAN is it potentially introduces many new patriarchs (dozens, or even hundreds), depending on the number of children of the super-large cluster root(s). This could lead to many small tribal bands at Level 1. In a future study, the benefits and drawbacks of the Recursive TAN and Expanded TAN for super-large clusters will be compared.

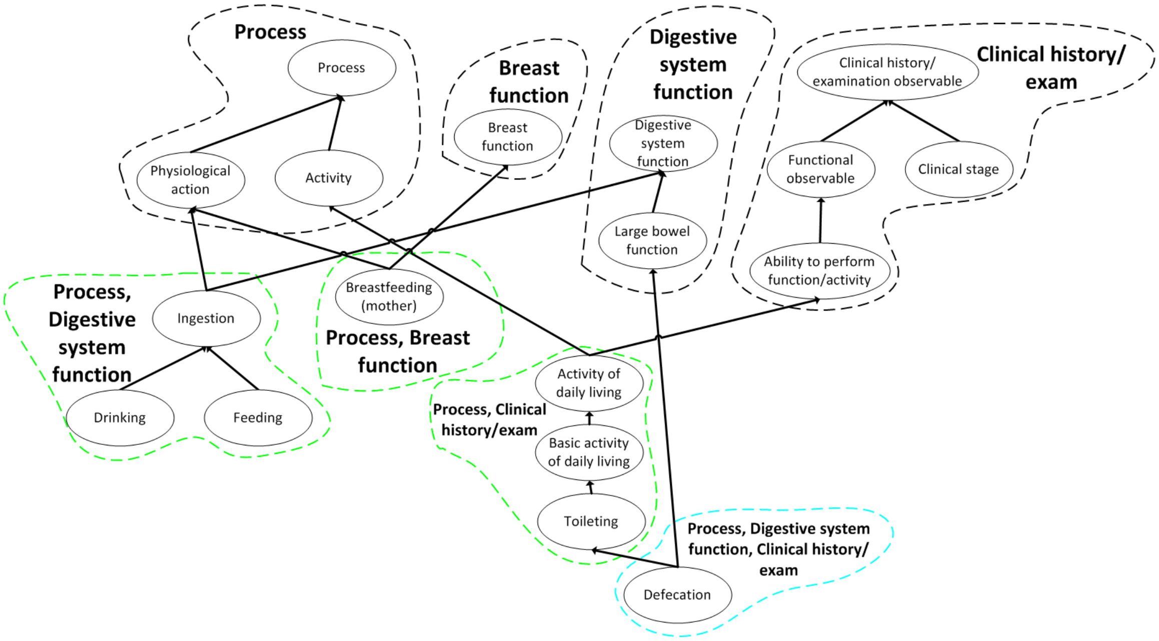

Process
Process
Physiological action
Activity
Breast function
Breast function
Digestive system function
Digestive system function
Large bowel function
Clinical history/ exam
Clinical history/ examination observable
Functional observable
Clinical stage
Ability to perform function/activity
Process, Digestive system function
Ingestion
Drinking
Feeding
Breastfeeding (mother)
Process, Breast function
Activity of daily living
Process, Clinical history/exam
Basic activity of daily living
Toileting
Process, Digestive system function, Clinical history/ exam
Defecation

Figure 4.1 The concepts from Figure 3.14 grouped based on their common tribes. The two children of *Function* in the example, *Breast function* and *Digestive system function* are now defined as patriarchs instead of *Function*.

After creating a Recursive TAN or Expanded TAN, it is possible that there will still be an over-summarization of a hierarchy’s concepts. For example, a Recursive TAN may contain super-large clusters if enough concepts in the original super-large cluster do not belong to multiple recursively-defined tribes. In such a case, the Recursive TAN can be applied until a TAN of desired summarization granularity is obtained. Different approaches for addressing over summarization in Recursive TANs and Expanded TANs will be investigated in a future study.

Another issue which will be investigated is the applicability of TANs to other terminologies and ontologies. The TAN derivation methodology is potentially applicable to any ontology that has concepts with multiple parents (i.e., its concepts are organized as a directed acyclic graph). TANs will be derived for other ontologies, e.g., those from

BioPortal [50], and their properties will be investigated as part of the family-based quality assurance approach introduced by He et al. [55].

4.3 Abstraction Networks for OWL Ontologies

The development of OWL partial-area taxonomy derivation methodologies was an important part of the family-based ontology quality assurance methodology introduced by He et al. [55]. The taxonomy derivation methods, and associated quality assurance reviews, described throughout Section 3.4 showed the feasibility of the family-based approach. However, several important issues will be investigated.

One future study will focus on the development of refined structural classifications that will organize OWL ontologies into refined families. In the preliminary study described in He et al. [55] ontologies were organized into seven disjoint families according to the existence and non-existence of object properties. The decision to organize ontologies into families according to object properties was based on the importance of object properties in taxonomy derivation. However, this initial classification does not accurately represent the structure of many ontologies.

For example, many ontologies have both object properties and data properties. This information was not captured by the preliminary classification process, inhibiting the development of abstraction networks that are applicable to a certain structural family. A new classification process, based on a *structural meta-ontology*, will consider all of the structural features of an ontology, and thus, will enable more accurate classifications.

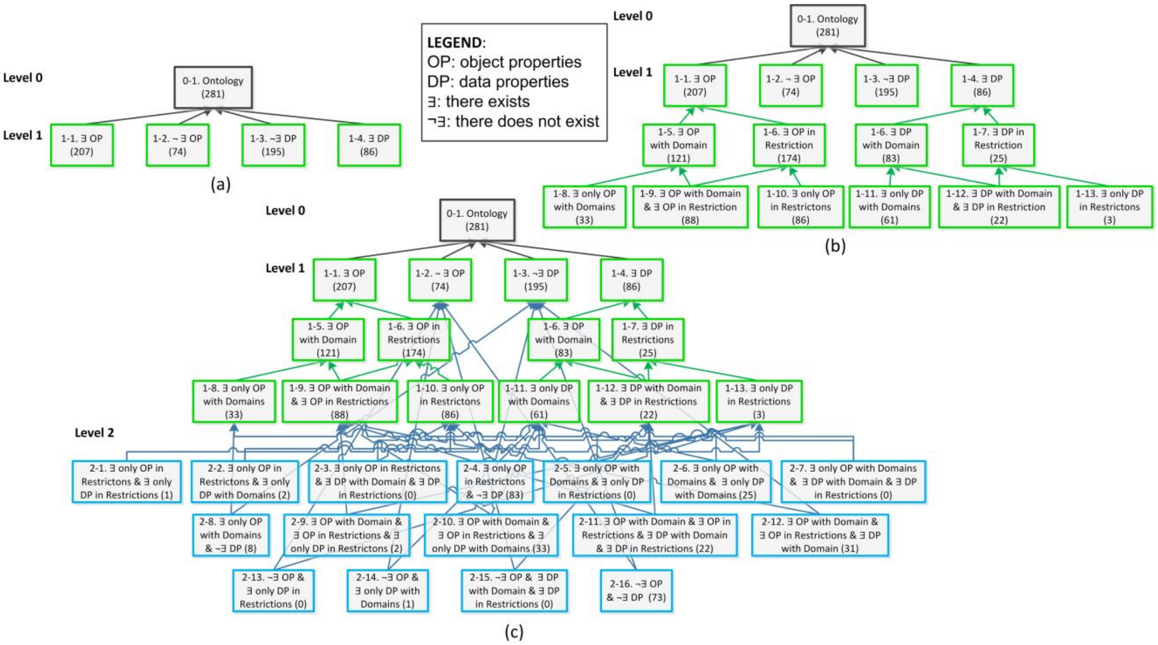

Level 0
0-1. Ontology (281)
Level 1
1-1. ∃ OP (207)
1-2. ¬ ∃ OP (74)
1-3. ¬∃ DP (195)
1-4. ∃ DP (86)
(a)
LEGEND:
OP: object properties
DP: data properties
∃: there exists
¬∃: there does not exist
Level 0
0-1. Ontology (281)
Level 1
1-1. ∃ OP (207)
1-2. ¬ ∃ OP (74)
1-3. ¬∃ DP (195)
1-4. ∃ DP (86)
1-5. ∃ OP with Domain (121)
1-6. ∃ OP in Restriction (174)
1-6. ∃ DP with Domain (83)
1-7. ∃ DP in Restriction (25)
1-8. ∃ only OP with Domains (33)
1-9. ∃ OP with Domain & ∃ OP in Restriction (88)
1-10. ∃ only OP in Restrictons (86)
1-11. ∃ only DP with Domains (61)
1-12. ∃ DP with Domain & ∃ DP in Restriction (22)
1-13. ∃ only DP in Restrictons (3)
(b)
Level 0
0-1. Ontology (281)
Level 1
1-1. ∃ OP (207)
1-2. ¬ ∃ OP (74)
1-3. ¬∃ DP (195)
1-4. ∃ DP (86)
1-5. ∃ OP with Domain (121)
1-6. ∃ OP in Restrictions (174)
1-6. ∃ DP with Domain (83)
1-7. ∃ DP in Restrictions (25)
1-8. ∃ only OP with Domains (33)
1-9. ∃ OP with Domain & ∃ OP in Restrictions (88)
1-10. ∃ only OP in Restrictons (86)
1-11. ∃ only DP with Domains (61)
1-12. ∃ DP with Domain & ∃ DP in Restrictions (22)
1-13. ∃ only DP in Restrictions (3)
Level 2
2-1. ∃ only OP in Restrictons & ∃ only DP in Restrictions (1)
2-2. ∃ only OP in Restrictons & ∃ only DP with Domains (2)
2-3. ∃ only OP in Restrictons & ∃ DP with Domain & ∃ DP in Restrictions (0)
2-4. ∃ only OP in Restrictons & ¬∃ DP (83)
2-5. ∃ only OP with Domains & ∃ only DP in Restrictions (0)
2-6. ∃ only OP with Domains & ∃ only DP with Domains (25)
2-7. ∃ only OP with Domains & ∃ DP with Domain & ∃ DP in Restrictions (0)
2-8. ∃ only OP with Domains & ¬∃ DP (8)
2-9. ∃ OP with Domain & ∃ OP in Restrictions & ∃ only DP in Restrictons (2)
2-10. ∃ OP with Domain & ∃ OP in Restrictions & ∃ only DP with Domains (33)
2-11. ∃ OP with Domain & ∃ OP in Restrictions & ∃ DP with Domain & ∃ DP in Restrictions (22)
2-12. ∃ OP with Domain & ∃ OP in Restrictions & ∃ DP with Domain (31)
2-13. ¬∃ OP & ∃ only DP in Restrictions (0)
2-14. ¬∃ OP & ∃ only DP with Domains (1)
2-15. ¬∃ OP & ∃ DP with Domain & ∃ DP in Restrictions (0)
2-16. ¬∃ OP & ¬∃ DP (73)
(c)

Figure 4.2 **(a)** The first two levels of the structural meta-ontology. **(b)** The classes of Level 1 refined based on their usage. **(c)** The complete structural meta-ontology for 281 BioPortal ontologies with *F* = {*object property*, *data property*}.

A structural meta-ontology is an ontology that classifies a given set of ontologies according to their structure. The structural meta-ontology derivation methodology will utilize combinations of the existence (or non-existence) and usage of a set of structural features to define its classes. The classes of a structural meta-ontology will categorize ontologies into structurally similar families based on their structural feature conditions.

Given a set of ontologies *O* and a set $F = \{f_1, f_2, f_3, \ldots, f_k\}$ of *k* structural features, a structural meta-ontology is organized into *k+1* levels of classes, L_0-L_k, based on the combination of the existence or nonexistence of *i* structural features at the level L_i. At L_0 a single root class named *Ontology* is defined to represent every ontology in *O*. All classes in the structural meta-ontology are descendants of *Ontology*.

Figure 4.2 illustrates the derivation of a structural meta-ontology for 281 BioPortal ontologies using *F* = {*object properties*, *data properties*}. Figure 4.2(a) shows the high level classes that are used to organize ontologies according to the existence of non-existence of object properties and data properties. Figure 4.2(b) refines these classes according to how each structural feature is used. Finally, Figure 4.2(c) shows the complete structural meta-ontology, which captures the existence of usage of object properties and data properties within the 281 BioPortal ontologies.

It is anticipated that the refined classification provided by the structural meta-ontology will lead to improved abstraction network derivation methodologies. For example, if many ontologies have both object properties and data properties, then taxonomies based on both structural features can be derived for all such ontologies. It will also likely be necessary to develop additional types of OWL abstraction networks that are applicable to certain families of ontologies where taxonomies do not provide ideal summarization. For example, the tribal abstraction network may be used for the family of ontologies that do not have either object properties or data properties. To investigate the applicability of different abstraction networks to different families, abstraction networks will be derived for each member of the family and the properties of the abstraction network within the family will be investigated.

Finally, with the development of the BLUOWL Protégé Plugin, it is now feasible to use abstraction networks during the ontology development process. The findings of the various planned family studies will be used to identify characteristics that may lead to problems in the ontology. The BLUOWL Plugin will use this information to alert an ontology curator to these potential problem areas while he or she is editing an ontology.

4.4 Diff Abstraction Networks

The diff taxonomies described in Section 3.4 enable the summarization and visualization of structural changes between two ontology releases. Diff Abstraction Networks based on other structural features, e.g., data properties, class equivalences, and hierarchical relationships, will be developed. Moreover, refined versions of the diff taxonomies that capture different kinds of structural changes will be able to provide further insight into the different types of changes that occur between two ontology releases. However, several significant issues that potentially affect all types of Diff AbNs will be investigated.

First, ideally, errors should be identified and corrected during the development process of an ontology. If an ontology curator can see the global impact of an editing operation before she modifies the ontology then certain kinds of errors can be avoided all together. The development of the BLUOWL Protégé Plugin, which can derive diff taxonomies, will be used to investigate the use of Diff AbNs to enable "what if?" analysis in support of ontology development. As an ontology curator is making changes he or she will be provided with a diff taxonomy that reflects the state of the ontology after a given potential editing operation is applied. If the curator determines this diff taxonomy exposes an anomaly then the potential editing operation would not be applied to the ontology.

Next, a common ontology design pattern, extensively used in biomedical ontologies, is to import and reuse the content of other ontologies, e.g., a top-level ontology like Basic Formal Ontology (BFO) [46] or a top-domain ontology like the Ontology for General Medical Science (OGMS) [47]. Ontology curators are often not interested in changes that happened within imported ontologies. As described in Section 3.4.3.3 with regards to ERO, the current Diff AbN derivation technique considers all structural changes between two versions of an ontology, including those that occurred to content from imported ontologies.

In some situations, this information could be important for detecting errors and inconsistencies in the ontology. Changes in the modeling of the imported ontology could lead to unintended changes to the content added by an ontology curator. However, if an ontology curator is not interested in seeing these changes, she could instead derive a Diff AbN that only captures the changes to her ontology. Methodologies for controlling which content is summarized by a Diff AbN will be investigated.

A visualization issue, which is illustrated by the SDO and ERO DPATs, is the emergence of many removed diff area/introduced diff area pairs and corresponding diff partial-area pairs, summarizing the changes in the object properties for the same set of classes. To address this issue, several refinements of the diff approach and the diff taxonomy visualization will be investigated. For example, when a removed diff area/introduced diff area pair exists, this information can be expressed as a modification to the area’s properties instead of a removed and an introduced area.

Finally, Diff AbNs can be used to compare abstraction networks of different granularity, e.g., [120]. A Diff AbN can be used to compare abstraction networks derived using different structural features. Such a Diff AbN will enable a user to see how different ontology classes are summarized in different abstraction networks. This kind of diff abstraction network will highlight the different granularities of summarization and can be used to determine which abstraction network is best for quality assurance.

4.5 Abstraction Network Tools

The BLU Framework enabled the majority of the research described in this dissertation. Future work will focus on two important areas. First, BLUSNO and BLUOWL will continue to be improved upon and expanded in terms of functionality. User studies are planned to evaluate both tools to improve the user interface and user experience. With the public release of both BLUSNO and the BLUOWL Protégé Plugin, user feedback will also be utilized to improve both tools.

The second major area of future research will be the development of the BLU Framework into a generic system for deriving abstraction networks. In previous phases of development, abstraction networks were implemented in a "one at a time" manner into each BLU Framework component. For example, the software components needed to derive and represent partial-area taxonomies for SNOMED CT and OWL ontologies were disconnected, even though the derivation followed the same general process.

To address this issue, significant portions of the BLU Framework's subcomponents are in the process of being redeveloped into generic systems that can be applied to any ontological system. Thus, when a new abstraction network is developed it can be implemented generically in the BLU Framework. The new abstraction network could then be applied to any ontology system supported by the BLU Framework (e.g., SNOMED CT, OWL ontologies, and OBO ontologies). This generic approach is currently used when deriving partial-area taxonomies and disjoint partial-area taxonomies. When support for a new ontological system is added to the BLU Framework a minimal amount of work will be needed to derive the different kinds of abstraction networks discussed throughout this dissertation (e.g., taxonomies, TANs, etc.).

To test this generic approach, the BLU Framework is currently being expanded to support the derivation of abstraction networks for the National Drug File - Reference Terminology (NDF-RT) [153]. NDF-RT is released in Apelon Distributed Terminology System (DTS) format [154]. The BLU Framework can now be used to derive abstraction networks for any DTS-based terminology. For example, partial-area taxonomies can be derived for NDF-RT. Furthermore, a new kind of abstraction network developed for NDF RT has been implemented generically, enabling it to be derived for SNOMED CT or OWL ontologies

Another potential area of research is determining the summarization needs of a BLU Framework user and providing them with a summary that would best support their intended use case. Currently, the BLUOWL tool performs a basic analysis of the structural of a selected ontology. The number of object properties and data properties, along with the total number of unique domains for each, is computed. Based on this information BLUOWL can suggest a type of partial-area taxonomy (domain defined, restriction defined, etc.) which would provide a reasonable summary. If a given ontology only has, say, only one data property with an explicitly defined domain then the tool does not suggest using a data property defined partial-area taxonomy.

Using a more advanced approach, a user could specify an area of interest, like in a subject subtaxonomy, and specify what structural features of the ontology they are most interested in for their use case. BLUOWL could then determine, based on the user's chosen criteria, which abstraction networks are most relevant for the user's needs.

.

CHAPTER 5

CONCLUSIONS

In conclusion, this dissertation expanded on the applicability of abstraction networks by exploring five important research topics:

1. The scalability of abstraction network quality assurance methodologies to large SNOMED CT hierarchies using various kinds of subtaxonomies
2. The development of an abstraction network for SNOMED CT hierarchies without attribute relationships called the Tribal Abstraction Network
3. Abstraction network derivation methodologies for Web Ontology Language ontologies
4. Diff Abstraction Networks for summarizing and visualizing the structural changes between two ontology releases
5. The development of various software tools to support abstraction network research and utility

First, partial-area taxonomy subsets called subtaxonomies were developed to enable the scalability of taxonomy-based quality assurance methodologies to large SNOMED CT hierarchies. A relationship subtaxonomy was utilized in a quality assurance review of the *Procedure* hierarchy's large partial-area taxonomy and a subject subtaxonomy was used in a quality assurance review of the *Bleeding* subhierarchy from the *Clinical finding* hierarchy. These initial studies showed that, by creating subsets of taxonomies, previously developed taxonomy-based quality assurance methodologies can be utilized to improve the quality of SNOMED's larger hierarchies. Concepts in small partial-areas were found to be more likely to contain errors than concepts in large partial-areas in a relationship subtaxonomy and various characteristics of overlapping concepts were identified as having higher error rates in a subject subtaxonomy.

Next, the Tribal Abstraction Network (TAN) was introduced to enable the summarization and quality assurance of SNOMED CT hierarchies which have no attribute relationships. The TAN was shown to successfully summarize the content and structure of these hierarchies. TANs were also shown to support quality assurance by identifying groups of concepts that were more likely to contain errors. Utilizing a TAN, a quality assurance review of the *Observable entity* hierarchy was performed. The study found that large clusters were more likely to contain erroneous concepts than small clusters. TANs can also be used in several additional settings. For example, TANs can be used to support quality assurance of concepts in large partial-areas and in hierarchies with attribute relationships. Several additional TAN studies were proposed to investigate various issues encountered during the *Observable entity* quality assurance review. For example, two methods for summarizing super-large clusters were introduced.

In regards to the third research topic, the domain-defined partial-area taxonomy and restriction-defined partial-area taxonomy were introduced as abstraction networks that can be applied to many structurally similar ontologies. A domain-defined partial-area taxonomy was derived for the Ontology of Clinical Research (OCRe) and a restriction-defined partial-area taxonomy was derived for the Sleep Domain Ontology (SDO). A study of the SDO's taxonomies investigated the differences in abstraction network granularity for a domain-defined taxonomy and a restriction-defined taxonomy.

A quality assurance review of OCRe showed that erroneous classes could be identified using an abstraction network. The errors were fixed and a new version of OCRe was released. Similarly, a quality assurance review of the SDO using a (domain or restriction)-defined taxonomy identified several errors and inconsistencies. A preliminary review of the Gene Ontology's taxonomy found that overlapping terms were more likely to have errors than non-overlapping terms. Future OWL abstraction network studies will focus on developing improved classification techniques for organizing ontologies into structural families and the investigation of properties for abstraction networks for each family.

For the fourth topic, two kinds of Diff Abstraction Networks, the diff area taxonomy and the diff partial-area taxonomy, were introduced to summarize and visualize the structural changes between two ontology releases. Diff partial-area taxonomies were derived for the Ontology of Clinical Research, Sleep Domain Ontology, and eagle-I Research Resource Ontology. The diff taxonomies were compared to the output provided by a standard ontology diff and each ontology's curator provided feedback and suggestions to improve the utility of the diff taxonomies. In future studies diff taxonomies will be integrated into the ontology development process, enabling ontology curators to view the global impact of their changes as they are being made.

Finally, the various components of the BLU Framework were developed to derive, visualize, and explore abstraction networks. BLUSNO enables the derivation of abstraction networks for SNOMED CT and includes an innovative concept browser that combines a traditional concept neighborhood view with information from abstraction networks. At any time a user can seamlessly transition between both views of SNOMED's content. BLUOWL enables the derivation of partial-area taxonomies using various structural features of OWL ontologies. BLUOWL can also derive Diff Partial-area Taxonomies. The BLUOWL Protégé plugin integrates BLUOWL's abstraction network derivation functionality into Protégé, a software tool commonly used to create ontologies. In future work, the functionality of the BLU Framework will be improved and expanded on. Additionally, generic functionality will be developed to enable the derivation of abstraction networks for different ontology systems (e.g., NDF RT).

REFERENCES

[1] Rector, A. L., Qamar, R., Marley, T. Binding ontologies and coding systems to electronic health records and messages. Applied Ontology. 2009;4(1):51-69.

[2] Eichelberg, M., Aden, T., Riesmeier, J., et al. A survey and analysis of electronic healthcare record standards. ACM Computing Surveys (CSUR). 2005;37(4):277-315.

[3] Giannangelo, K., Berkowitz, L. SNOMED CT helps drive EHR success. Journal of American Health Information Management Association (J AHIMA). 2005;76(4):66-7.

[4] Rubin, D. L., Shah, N. H., Noy, N. F. Biomedical ontologies: a functional perspective. Briefings in Bioinformatics. 2008;9(1):75-90.

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

[6] Fensel, D. Ontology-based knowledge management. Computer, IEEE. 2002;35(11):56-9.

[7] Maedche, A., Motik, B., Stojanovic, L., et al. Ontologies for enterprise knowledge management. Intelligent Systems, IEEE. 2003;18(2):26-33.

[8] Cao, F., Sun, X., Wang, X., et al. Ontology-based knowledge management for personalized adverse drug events detection. Studies in Health Technology and Informatics. 2011;169:699-703.

[9] Kiryakov, A. Ontologies for knowledge management. Semantic Web Technologies: trends and research in ontology-based systems. 2006:115-38.

[10] Lussier, Y. A. Ontologies for natural language processing. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 2005.

[11] Estival, D., Nowak, C., Zschorn, A. Towards ontology-based natural language processing. Proceeedings of the Workshop on NLP and XML (NLPXML-2004). 2004:59-66.

[12] Ovchinnikova, E. Integration of world knowledge for natural language understanding. Vol 3. Atlantis Press: Paris, France. 2012.

[13] Bodenreider, O., Stevens, R. Bio-ontologies: current trends and future directions. Briefings in Bioinformatics. 2006;7(3):256-74.

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

[15] SNOMED CT [10 February 2015]. Available from: http://www.ihtsdo.org/snomed-ct/.

[16] Stearns, M. Q., Price, C., Spackman, K. A., et al. SNOMED clinical terms: overview of the development process and project status. AMIA Annu Symp Proc. 2001:662-6.

[17] Stenzhorn, H., Schulz, S., Boeker, M., et al. Adapting clinical ontologies in real-world environments. Journal of Universal Computer Science (J UCS). 2008;14(22):3767-80.

[18] Schulz, S., Suntisrivaraporn, B., Baader, F., et al. SNOMED reaching its adolescence: ontologists' and logicians' health check. International Journal of Medical Informatics. 2009;78 Suppl 1:S86-94.

[19] Rector, A. L., Brandt, S., Schneider, T. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. J Am Med Inform Assoc. 2011;18(4):432-40.

[20] Gu, H., Perl, Y., Geller, J., et al. Modeling the UMLS using an OODB. AMIA Annu Symp Proc. 1999:82-6.

[21] Gu, H., Cimino, J. J., Halper, M., et al. Utilizing OODB schema modeling for vocabulary management. Proc AMIA Annu Symp. 1996:274-8.

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

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

[24] Wang, Y., Halper, M., Min, H., et al. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007;40(5):561-81.

[25] Ochs, C., Agrawal, A., Perl, Y., et al. Deriving an abstraction network to support quality assurance in OCRe. AMIA Annu Symp Proc. 2012:681-9.

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

[27] Ochs, C., Geller, J., Perl, Y., et al. A tribal abstraction network for SNOMED CT hierarchies without attribute relationships. J Am Med Inform Assoc. 2014. doi: 10.1136/amiajnl-2014-003173

[28] Ochs, C., Perl, Y., Geller, J., et al. Scalability of abstraction-network-based quality assurance to large SNOMED hierarchies. AMIA Annu Symp Proc. 2013:1071-80.

[29] Wang, Y., Halper, M., Wei, D., et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. J Biomed Inform. 2012;45(1):1-14.

[30] The Description Logic Handbook: Theory, implementation, and applications: Cambridge University Press: Cambridge, United Kingdom. 2003.

[31] Cimino, J. J., Zhu, X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform. 2006:124-35.

[32] Yu, A. C. Methods in biomedical ontology. J Biomed Inform. 2006;39(3):252-66.

[33] Members: International Health Terminology Standards Development Organization; 2013 [10 Februrary 2015]. Available from: http://www.ihtsdo.org/members/.

[34] American Recovery and Reinvestment Act, United States Congress, 111 Sess. (2009).

[35] World Health Organization. ICD-10: International statistical classification of diseases and related health problems. American Psychiatric Publishing: Arlington, VA, United States. 2004.

[36] The CORE Problem List Subset of SNOMED CT [10 Februrary 2015]. Available from: http://www.nlm.nih.gov/research/umls/Snomed/core_subset.html.

[37] UMLS enhanced VA/KP Problem List Subset of SNOMED [10 Februrary 2015]. Available from: http://www.nlm.nih.gov/research/umls/licensedcontent/vakpproblemlist.html.

[38] SNOMED CT User Guide [4 April 2015]. Available from: http://ihtsdo.org/fileadmin/user_upload/doc/

[39] Baader, F., Brandt, S., Lutz, C. Pushing the EL envelope. Interntional Joint Conference on Artificial Intelligence (IJCAI). 2005;5:364-9.

[40] Motik, B., Patel-Schneider, P. F., Parsia, B. OWL 2 Web Ontology Language structural specification and functional style syntax. W3C -- World Wide Web Consortium, 2009.

[41] de Coronado, S., Haber, M. W., Sioutos, N., et al. NCI Thesaurus: using science-based terminology to integrate cancer research results. Studies in Health Technology and Informatics. 2004;107(Pt 1):33-7.

[42] Ashburner, M., Ball, C. A., Blake, J. A., et al. Gene Ontology: tool for the unification of biology. Nature Genetics. 2000;25(1):25-9.

[43] Brinkman, R. R., Courtot, M., Derom, D., et al. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010;1(Suppl 1):S7.

[44] Horridge, M., Drummond, N., Goodwin, J., et al. The Manchester OWL Syntax. OWLed. 2006;216.

[45] Arabandi, S., Ogbuji, C., Redline, S., et al. Developing a sleep domain ontology. AMIA Clinical Research Informatics Summit. 2010.

[46] Grenon, P., Smith, B., Goldberg, L. Biodynamic Ontology: Applying BFO in the biomedical domain. In: Pisanelli DM, editor. Ontologies in Medicine: IOS Press: Amsterdam, The Netherlands. 2004. p. 20-38.

[47] Goldfain, A. Ontology for General Medical Science (OGMS) [10 February 2015]. Available from: http://code.google.com/p/ogms/.

[48] Beisswanger, E., Schulz, S., Stenzhorn, H., et al. BioTop: An upper domain ontology for the life sciences. A description of its current structure, contents and interfaces to OBO ontologies. Applied Ontology. 2008;3(4):205-12.

[49] Goldfain, A., Smith, B., Arabandi, S., et al. Vital sign ontology. Proceedings of the Workshop on Bio-Ontologies. 2011:71-4.

[50] Whetzel, P. L., Noy, N. F., Sham, N. H., 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 (NAR). 2011;39:W541-5.

[51] Sim, I., Carini, S., Tu, S., et al. The human studies database project: federating human studies design data using the ontology of clinical research. AMIA Summits Transl Sci Proc. 2010:51-5.

[52] Smith, B., Ashburner, M., Rosse, C., et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology. 2007;25(11):1251-5.

[53] RDF/XML Syntax Specification: World Wide Web Consortium; 2004 [10 February 2015]. Available from: http://www.w3.org/TR/REC-rdf-syntax/.

[54] Mortensen, J. M., Horridge, M., Musen, M. A., et al. Applications of ontology design patterns in biomedical ontologies. AMIA Annu Symp Proc. 2012:643–52.

[55] He, Z., Ochs, C., Agrawal, A., et al. A family-based framework for supporting quality assurance of biomedical ontologies in BioPortal. AMIA Annu Symp Proc. 2013:581-90.

[56] Noy, N. F., Crubézy, M., Fergerson, R. W., et al. Protege-2000: an open-source ontology-development and knowledge-acquisition environment. AMIA Annu Symp Proc. 2003:953.

[57] CliniClue Xplore [10 February 2015]. Available from: http://www.cliniclue.com/software.

[58] UMLS Terminology Services [10 February 2015]. Available from: https://uts.nlm.nih.gov/home.html.

[59] Gu, H., Perl, Y., Geller, J., et al. Representing the UMLS as an object-oriented database: modeling issues and advantages. J Am Med Inform Assoc. 2000;7(1):66-80.

[60] Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(Database issue):D267-70.

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

[62] Cimino, J. J., Clayton, P. D., Hripcsak, G., et al. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994;1(1):35-50.

[63] Geller, J., Gu, H., Perl, Y., et al. Semantic refinement and error correction in large terminological knowledge bases. Data & Knowledge Engineering. 2003;45(1):1-32.

[64] 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.

[65] Gu, H., Perl, Y., Elhanan, G., et al. Auditing concept categorizations in the UMLS. Artif Intell Med. 2004;31(1):29-44.

[66] 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):1070-4.

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

[68] Pellet [10 February 2015]. Available from: http://clarkparsia.com/pellet.

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

[70] Li, N., Motta, E., d'Aquin, M. Ontology summarization: an analysis and an evaluation. The International Workshop on Evaluation of Semantic Technologies (IWEST 2010). 2010.

[71] Peroni, S., Motta, E., d'Aquin, M. Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological measures. The Semantic Web. 2008:242-56.

[72] Dzbor, M., Peroni, S., Motta, E., et al. NeOn Toolkit pug-in for visualization and navigation in ontologies and ontology networks based on concept summarization and categorizing. NeOn Project Deliverable. 2009. [10 February 2015]. Available from: http://www.neon-project.org/deliverables/WP4/NeOn_2009_D454.pdf.

[73] Haase, P., Lewen, H., Studer, R., et al. The neon ontology engineering toolkit. WWW Developers Track. 2008.

[74] Zhang, X., Cheng, G., Qu, Y. Ontology summarization based on rdf sentence graph. Proceedings of the 16th International Conference on World Wide Web. 2007:707-16.

[75] Page, L., Brin, S., Motwan, R., et al. The PageRank citation ranking: bringing order to the web. 1999. [10 February 2015]. Available from: http://ilpubs.stanford.edu:8090/422.

[76] Queiroz-Sousa, P. O., Salgado, A. C., Pires, C. E. A method for building personalized ontology summaries. Journal of Information and Data Management. 2013;4(3):236-250.

[77] Zhu, X., Fan, J.-W., Baorto, D. M., et al. A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomed Inform. 2009;42(3):413-25.

[78] de Coronado, S., Wright, L. W., Fragoso, G., et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform. 2009;42(3):530-9.

[79] Verspoor, K., Dvorkin, D., Cohen, K. B., et al. Ontology quality assurance through analysis of term transformations. Bioinformatics 2009;25(12):i77-i84.

[80] Agrawal, A., Perl, Y., Elhanan, G. Identifying problematic concepts in SNOMED CT using a lexical approach. Studies in Health Technology and Informatics. 2013:773-7.

[81] Agrawal, A., Perl, Y., Chen, Y., et al. Identifying Inconsistencies in SNOMED CT Problem Lists using Structural Indicators. AMIA Annu Symp Proc. 2013:17–26.

[82] Agrawal, A., Elhanan, G., Halper, M. Dissimilarities in the Logical Modeling of Apparently Similar Concepts in SNOMED CT. AMIA Annu Symp Proc. 2010:212-6.

[83] Ceusters, W., Smith, B., Kumar, A., et al. Ontology-based error detection in SNOMED-CT. Studies in Health Technology and Informatics. 2004;107(Pt 1):482-6.

[84] Ceusters, W., Martens, P., Dhaen, C., et al. LinkFactory: an advanced formal ontology management system. Proceedings of KCAP-2001 Interactive Tools for Knowledge Capture Workshop. 2001;175-204.

[85] Rector, A. L., Iannone, L. Lexically suggest, logically define: quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT. J Biomed Inform. 2011:199-209.

[86] Schulz, S., Hahn, U., Rogers, J. Semantic clarification of the representation of procedures and diseases in SNOMED CT. Studies in Health Technology and Informatics. 2005;116:773-8.

[87] Schulz, S., Hanser, S., Hahn, U., et al. The semantics of procedures and diseases in SNOMED CT. Methods Inf Med. 2006;45(4):354-8.

[88] Mortensen, J. M., Musen, M. A., Noy, N. F. Crowdsourcing the verification of relationships in biomedical ontologies. AMIA Annu Symp Proc. 2013:1020-9.

[89] Kittur, A., Chi, E. H., Suh, B. Crowdsourcing user studies with Mechanical Turk. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2008:453-6.

[90] Mortensen, J. M., Minty, E. P., Januszyk, M., et al. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. J Am Med Inform Assoc. 2014. doi: 10.1136/amiajnl-2014-002901

[91] Kremen, P., Smid, M., Kouba, Z. OWLDiff: A Practical Tool for Comparison and Merge of OWL ontologies. 22nd International Workshop on Database and Expert Systems Applications. 2011:229-33.

[92] Tudorache, T., Vendetti, J., Noy, N. F. Web-Protege: a lightweight OWL ontology editor for the web. OWLED. 2008;432.

[93] Kalyanpur, A., Parsia, B., Sirin, E., et al. Swoop: A web ontology editing browser. Web Semantics: Science, Services and Agents on the World Wide Web 2006;4(2):144-53.

[94] Day-Richter, J., Harris, M. A., Haendel, M., et al. OBO-Edit—an ontology editor for biologists. Bioinformatics 2007;23(16): 2198-200.

[95] International Health Terminology Standards Development Organization - Tooling 2014 [18 March 2014]. Available from: http://www.ihtsdo.org/develop/tooling/.

[96] Rogers, J., Bodenreider, O. SNOMED CT: Browsing the browsers. KR-MED 2008 Representing and sharing knowledge using SNOMED2008.

[97] B2i Healthcare - Snow Owl 2013 [10 February 2015]. Available from: http://www.b2international.com/portal/snow-owl.

[98] Morrey, C. P., Geller, J., Halper, M., et al. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. J Biomed Inform. 2009;42(3):468-89.

[99] Ochs, C., Geller, J., Perl, Y. A Relationship-centric Hybrid Interface for Browsing and Auditing the UMLS. Journal of Integrated Design and Process Science. 2011;15.4:3-25.

[100] Binns, D., Dimmer, E., Huntley, R., et al. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 2009;25(22):3045-6.

[101] The Neuroscience Lexicon: Gene Ontology Tools [23 March 2014]. Available from: http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools.

[102] Katifori, A., Halatsis, C., Lepouras, G., et al. Ontology visualization methods—a survey. ACM Computing Surveys (CSUR). 2007;39(4):10.

[103] Katifori, A., Torou, E., Halatsis, C., et al. A comparative study of four ontology visualization techniques in protege: Experiment setup and preliminary results. Tenth International Conference on Information Visualization. 2006;417-23.

[104] Storey, M.-A., Lintern, R., Ernst, N. A., et al. Visualization and protégé. 7th International Protégé Conference. 2004.

[105] Lanzenberger, M., Sampson, J., Rester, M. Visualization in ontology tools. International Conference on Complex, Intelligent and Software Intensive Systems. 2009:705-11.

[106] Fu, B., Noy, N. F., Storey, M.-A. Indented Tree or Graph? A Usability Study of Ontology Visualization Techniques in the Context of Class Mapping Evaluation. The Semantic Web–ISWC 2013. 2013:117-34.

[107] Ellson, J., Gansner, E., Koutsofios, L., et al. Graphviz—open source graph drawing tools. 9th International Symposium on Graph Drawing. 2002:483-4.

[108] Storey, M.-A., Musen, M., Silva, J., et al. Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé. Workshop on Interactive Tools for Knowledge Capture. 2001.

[109] Bosca, A., Bonino, D., Pellegrino, P. OntoSphere: more than a 3D ontology visualization tool. SWAP. 2005.

[110] Falconer, S. M., Callendar, C., Storey, M.-A. FLEXVIZ: Visualizing biomedical ontologies on the web. International Conference on Biomedical Ontology (ICBO). 2009.

[111] Fu, B., Grammel, L., Storey, M.-A. D. BioMixer: A web-based collaborative ontology visualization tool. International Conference on Biomedical Ontology (ICBO). 2012.

[112] Hunt, J. W., McIlroy, M. D. An algorithm for differential file comparison. Bell Laboratories. 1976 [4 April 2015]. Available from: http://cm.bell-labs.com/cm/cs/cstr/41.pdf

[113] Noy, N. F., Kunnatur, S., Klein, M., et al. Tracking changes during ontology evolution. The Semantic Web–ISWC. 2004:259-73.

[114] Noy, N. F., Musen, M. Promptdiff: A fixed-point algorithm for comparing ontology versions. Association for the Advancement of Aritificial Intelligence (AAAI/IAAI). 2002:744-50.

[115] Jiménez-Ruiz, E., Grau, B. C., Horrocks, I., et al. Building ontologies collaboratively using ContentCVS. 22nd International Workshop on Description Logics. 2009;447.

[116] Gonçalves, R. S., Parsia, B., Sattler, U. Ecco: A Hybrid Diff Tool for OWL 2 ontologies. OWLED. 2012.

[117] Redmond, T., Noy, N. Computing the Changes Between Ontologies. Joint Workshop on Knowledge Evolution and Ontology Dynamics. 2011:1-14.

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

[119] Ochs, C., Geller, J., Perl, Y., et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. J Am Med Inform Assoc. 2014. doi: 10.1136/amiajnl-2014-003151

[120] Ochs, C., He, Z., Perl, Y., et al. Choosing the granularity of abstraction networks for orientation and quality assurance of the Sleep Domain Ontology. International Conference on Biomedical Ontology (ICBO). 2013:84-9.'

[121] Ochs, C., Perl, Y., Geller, J., et al. Summarizing and visualizing structural changes during the evolution of biomedical ontologies using a a diff abstraction network. J Biomed Inform. 2015. (Submitted for review).

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

[123] Fisher, R. A. Statistical Methods for Research Workers. 14 ed: Macmillan Pub Co; 1970 June 1970. 378 p.

[124] Cormen, T. H., Leiserson, C. E., Rivest, R. L., et al. Introduction to Algorithms: MIT Press and McGraw-Hill: Cambridge, MA, United States. 2001.

[125] Deaths and Mortality: Centers for Disease Control and Prevention (CDC); 2014 [9 September 2014]. Available from: http://www.cdc.gov/nchs/fastats/deaths.htm.

[126] Gu, H., Elhanan, G., Perl, Y., et al. A study of terminology auditors' performance for UMLS semantic type assignments. J Biomed Inform. 2012;45(6):1042-8.

[127] U.S. SNOMED CT Content Request System (USCRS): National Library of Medicine; 2012 [updated 20129 July 2014]. Available from: https://uscrs.nlm.nih.gov/.

[128] Agrawal, A., He, Z., Perl, Y., et al. The readiness of SNOMED problem list concepts for meaningful use of EHRs. Artif Intel Med. 2013;58(2):73-80.

[129] Tirmizi, S. H., Aitken, S., Moreira, D. A., et al. Mapping between the OBO and OWL ontology languages. Journal of Biomedical Semantics. 2011;2(Suppl 1):S3.

[130] OCRe in BioPortal BioPortal [10 February 2015]. Available from: http://bioportal.bioontology.org/ontologies/45657.

[131] Cook, D. L., Mejino, J. L., Rosse, C. The foundational model of anatomy: a template for the symbolic representation of multi-scale physiological functions. Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference. 2004;7:5415-8.

[132] Ochs, C., Perl, Y. Structural Analysis of Biomedical Ontologies Center (SABOC) - Taxonomy Figure Repository 2013 [1 October 2014]. Available from: http://cs.njit.edu/~oohvr/SABOC/figures.php.

[133] Charlet, J. Meneles Project Top-Level Ontology on BioPortal 2013 [10 February 2015]. Available from: http://bioportal.bioontology.org/ontologies/TOP-MENELAS.

[134] Winston, M. E., Chaffin, R., Herrmann, D. A taxonomy of part-whole relations. Cognitive Science. 1987;11(4):417-44.

[135] Consortium, T. G. O. Gene Ontology annotations and resources. Nucleic Acids Research. 2013(41):D530-D5.

[136] Ochs, C., Perl, Y., Halper, M., et al. Gene Ontology Summarization to Support Visualization and Quality Assurance. International Conference on Bioinformatics and Computational Biology (BICoB). 2015.

[137] Mungall, C. J., Bada, M., Berardini, T. Z., et al. Cross-product extensions of the Gene Ontology. J Biomed Inform. 2011;44(1):80-6.

[138] Degtyarenko, K., Matos, P. D., Ennis, M., et al. ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Research. 2008;36(1):D344-D50.

[139] He, Z., Ochs, C., Soldatova, L., et al. Auditing Redundant Import in Reuse of a Top Level Ontology for the Drug Discovery Investigations Ontology. Vaccines and Drug Ontology Studies (VDOS). 2013.

[140] Torniai, C., Brush, M. H., Vasilevsky, N., et al. Developing an Application Ontology for Biomedical Resource Annotation and Retrieval: Challenges and Lessons Learned. International Conference on Biomedical Ontology (ICBO). 2011:101-8.

[141] Torniai, C., Essaid, S., Lowe, B., et al. Finding common ground: integrating the eagle-i and VIVO ontologies. International Conference on Biomedical Ontology (ICBO). 2013:46-9.

[142] Krafft, D. B., Cappadona, N. A., Caruso, B., et al. Vivo: Enabling national networking of scientists. Proceedings of the Web Science Conference. 2010:1310-3.

[143] Ahmed, M., Chen, S., Ding, Y., et al. Aligning research resource and researcher representation: the eagle-i and VIVO use case. International Conference on Biomedical Ontology (ICBO). 2013:260-2.

[144] Vasilevsky, N., Johnson, T., Corday, K., et al. Research resources: curating the new eagle-i discovery system. Database. 2012;bar067.

[145] Mitchell, S., Chen, S., Ahmed, M., et al. The VIVO ontology: enabling networking of scientists. ACM WebScience Conference. 2011:14-7.

[146] connect-isf: The ontology developed in the context of CTSAConnect to represent agents, resources and grants 2014 [24 September 2014]. Available from: https://code.google.com/p/connect-isf/.

[147] obo-relations: The OBO Relations Ontology 2014 [18 September 2014]. Available from: https://code.google.com/p/obo-relations/.

[148] Smith, C. L., Goldsmith, C.-A. W., Eppig, J. T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. Genome Biology. 2004;6(1):R7.

[149] Malone, J., Brown, A., Lister, A. L., et al. The Software Ontology (SWO): a resource for reproducibility in biomedical data analysis, curation and digital preservation. Journal of Biomedical Semantics. 2014;5(1).

[150] Mungall, C. J., Torniai, C., Gkoutos, G. V., et al. Uberon, an integrative multi-species anatomy ontology. Genome Biology. 2012;13(1):R5.

[151] Protégé [20 December 2014]. Available from: http://protege.stanford.edu/.

[152] Protege Plugin Library - Protege Wiki 2014 [20 December 2014]. Available from: http://protegewiki.stanford.edu/wiki/Protege_Plugin_Library.

[153] Nelson, S. J., Brown, S. H., Erlbaum, M. S., et al. A semantic normal form for clinical drugs in the UMLS: early experiences with the VANDF. AMIA Annu Symp Proc. 2002:557-61.

[154] Apelon Distributed Terminology System (DTS) [9 January 2015]. Available from: http://www.apelondts.org/.