ABSTRACT

ENRICHING AND DESIGNING METASCHEMAS FOR THE UMLS SEMANTIC NETWORK

by
Li Zhang

The disparate terminologies used by various biomedical applications or professionals make the communication between them more difficult. The Unified Medical Language System (UMLS) of the National Library of Medicine (NLM) is an attempt to integrate different medical terminologies into a unified representation framework to improve decision making and the quality of patient care as well as research in the health-care field. Metathesaurus (META) and Semantic Network (SN) are two main components of the UMLS system, where the SN provides a high-level abstract of the concepts in the META.

This dissertation addresses three problems of the SN. First, the SN's two-tree structure is restrictive because it does not allow a semantic type to be a specialization of several other semantic types. This restriction leads to the omission of some subsumption knowledge in the SN. Secondly, the SN is large and complex for comprehension purposes and it does not come with a pictorial representation for users. As a partial solution for this problem, several metaschemas were previously built as higher-level abstractions for the SN to help users' orientation. Third, there is no efficient method to evaluate each metaschema. There is no technique to obtain a consolidated metaschema acceptable for a majority of the UMLS's users.

In this dissertation work the author attacked the described problems by using the following approaches. (1) The SN was expanded into the Enriched Semantic Network

(ESN), a multiple subsumption structure with a directed acyclic graph (DAG) IS-A hierarchy, allowing a semantic type to have multiple parents. New viable IS-A links were added as warranted. Two methodologies were presented to identify and add new viable IS-A links. The ESN serves as an extended high-level abstract of the META. (2) The ESN's semantic relationship distribution and concept configuration were studied. Rules were defined to derive the ESN's semantic relationship distribution from the current SN's semantic relationship distribution. A mapping function was defined to map the SN's concept configuration to the ESN's concept configuration, avoiding redundant classifications in the ESN's concept configuration. (3) Several new metaschemas for the SN and the ESN were built and evaluated based on several different partitioning techniques. Each of these metaschema can serve as a higher-level abstraction of the SN (or the ESN).

ENRICHING AND DESIGNING METASCHEMAS FOR THE UMLS SEMANTIC NETWORK

by

Li Zhang

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

Department of Computer Science

May 2004

APPROVAL PAGE

ENRICHING AND DESIGNING METASCHEMAS FOR THE UMLS SEMANTIC NETWORK

Li Zhang

Yehoshua Perl, Ph.D., Dissertation Advisor — Date
Professor, Computer Science Department, NJIT

James Geller, Ph.D., Committee Member — Date
Professor, Computer Science Department, NJIT

James McHugh, Ph.D., Committee Member — Date
Professor, Computer Science Department, NJIT

James J. Cimino, M.D., Committee Member — Date
Professor, Medical Informatics and Medicine, Columbia University

Michael Halper, Ph.D., Committee Member — Date
Professor, Computer Science Department, Kean University

Helen Gu, Ph.D., Committee Member — Date
Assistant Professor, Health Informatics Department, UMDNJ

BIOGRAPHICAL SKETCH

Author: Li Zhang

Degree: Doctor of Philosophy

Date: May 2004

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2004
- Master of Science in Computer Science,
  Harbin Engineering University, Harbin, China, 1999
- Bachelor of Science in Computer Science,
  Harbin Engineering University, Harbin, China, 1996

Major: Computer Science

Presentations and Publications:

Zhang L., Perl Y., Geller J., Halper M., and Cimino J.J., "Enriching the structure of the UMLS Semantic Network," *Proc. of the 2002 AMIA Annual Symposium*, pp. 939-943, San Antonio, TX, Nov. 2002.

Zhang L., Perl Y., Geller J., Halper M., and Cimino J.J., "An Enriched UMLS Semantic Network with a Multiple Inheritance Hierarchy," *Journal of the American Medical Informatics Association*, 2004, to appear.

Zhang L., Perl Y., Halper M., and Geller J., "Designing metaschemas for the UMLS Enriched Semantic Network," *Journal of Biomedical Informatics*, 36(6), pp. 433–449, Dec. 2003.

Zhang L., Perl Y., Halper M., and Geller J., and Hripcsak G. "A Lexical Metaschema for the UMLS Semantic Network," *Artificial Intelligence of Medicine*, 2003, submitted to journal publication.

Perl Y., Chen Z., Halper M., Geller J., Zhang L., Peng Y., "The cohesive metaschema: a higher-level abstraction of the UMLS Semantic Network," *Journal of Biomedical Informatics*, vol. 35(3), pp. 194–212, 2002.

Gu H., Perl Y., Elhanan G., Min H., Zhang L., Peng Y., "Auditing Concept Categorizations in the UMLS," *Artificial Intelligence of Medicine*, 2004, to appear.

Gu H., Min H., Peng Y., Zhang L., "Using the metaschema to audit UMLS classification errors," *Proc. of the 2002 AMIA Annual Symposium*, pp. 310-314, San Antonio, TX, Nov. 2002.

This work is dedicated to my beloved family

ACKNOWLEDGMENT

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

Special thanks to Dr. James J. Cimino for graciously helping me understand medical terminology and providing invaluable advice.

I would also like to thank Dr. Helen Gu and Dr. James McHugh for serving as members of my committee.

I thank Hua Min for her great help and support throughout the four years of my research and for the joyful days in the Medical Informatics Lab.

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

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

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF TABLES
(Continued)

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

1.1 Motivation

Biomedical professionals have diverse perspectives and approaches for solving contemporary problems. The disparate terminologies used by various biomedical applications or professionals make the communication between them more difficult. Controlled vocabularies have proven to be extremely useful in facilitating such communications. The Unified Medical Language System (UMLS) of the National Library of Medicine (NLM), initiated in 1986, is an attempt to integrate a number of medical terminologies into a unified knowledge representation framework [5, 30, 31, 34]. It also helps to improve the ability of computer programs to "understand" biomedical meaning in user inquiries and to use this understanding to retrieve and integrate relevant machine-based information. The UMLS provides users with accurate and up-to-date information which helps to improve decision making and ultimately the quality of patient care as well as research in the health-care field.

The UMLS [59] contains three Knowledge Sources: the Metathesaurus (META) [56, 58], the Semantic Network [37, 40], and the Specialist Lexicon. The META is the central vocabulary component of the UMLS, which represents medical knowledge in the form of names of concepts and links between those concepts [45]. Different names for a biomedical meaning are linked to a single Metathesaurus concept. Extensive additional information describing semantic characteristics, occurrence in machine-readable information sources, and how concepts co-occur in these sources is also provided, enabling a greater

comprehension of the concept in its various contexts. Thus the META serves as the central repository of concepts used in the biomedical field. The META also preserves the meaning, hierarchical connections, and other relationships between concepts presented in its source vocabularies. The latest version of the UMLS contains the Metathesaurus (META) with about 900,000 concepts (790,000 concepts in [59] and 871,000 concepts in [60]).

The purpose of the Semantic Network (SN) is to provide a consistent categorization of all concepts found in the META and to provide useful links between these concepts at the level of the semantic types [38, 40]. The SN contains 135 semantic types (134 in [60]), and hierarchical and non-hierarchical relationships between semantic types [40, 61]. Each concept in the META is assigned to one or more semantic types, so that the SN can provide some semantics for META's concepts. The assignment of concepts to semantic types involves algorithmic procedures as well as extensive review by domain experts based on two assumptions: 1) each concept is assigned to the most specific semantic type available; 2) semantic types are assigned according to the meaning or meanings that the concept has in its source vocabulary [38]. In this way, the SN serves as a high-level abstract view of the META [38, 40], which helps organize the large number of concepts in the biomedical field. This is expressed in [41] as follows: "The Semantic Network encompasses and provides a unifying structure for the Metathesaurus constituent vocabularies."

The SN contains a hierarchy consisting of two trees, rooted at the semantic types **Event**[1] and **Entity**, respectively [37]. This hierarchy is based on the IS-A relationship, which connects a more specialized semantic type (a child) to a more generalized semantic type (its parent). Each semantic type, except for **Event** and **Entity**, is a specialization

[1]Semantic types will be written in bold font in this dissertation except in tables and figures.

of exactly one semantic type (its parent) and inherits semantic relationships only from this unique parent. The child semantic type will inherit all semantic relationships from its parent unless a relationship is explicitly blocked from being inherited using the "DNI" (Defined but Not Inherited by the children of the Arguments) or the "blocked" mechanism.

While this tree structure is easy to implement and process, it is restrictive in that it does not allow multiple parents and the accompanying multiple inheritance of relationships from several semantic types. There are, in fact, some semantic types that could naturally be specializations of more than one semantic type. For example, **Gene or Genome** could conceptually be a child of two semantic types: one is its current parent **Fully Formed Anatomical Structure**; the other is **Molecular Sequence**. Hence, **Gene or Genome** should inherit from **Molecular Sequence** the semantic relationship *result_of* to **Mental Process**. In a case such as this, the modeling of the SN omits an aspect of current medical knowledge. Therefore, one important research goal is to enrich the SN from a two-tree structure into a Directed Acyclic Graph (DAG) to accommodate these omitted subsumption relationships. As a result of the enrichment, a new version of the SN, called the Enriched Semantic Network (ESN) is derived. The ESN will serve as an extended high-level abstract view of the META.

The SN, besides its IS-A links, contains about 7,000 semantic (non-hierarchical) relationships instances of 53 kinds which connect semantic types. This number of relationships makes it a large and complex framework, difficult for orientation and comprehension purposes. It is necessary to develop efficient visualization tools to help user orientation to the complex SN. Typically, a convenient way for a user to get oriented to such a large

knowledge structure is by studying a diagrammatic representation. People often prefer a graphical representation to an equivalent textual form, which may be quite extensive and unruly. However, a complete picture of the SN is by far too large for easy comprehension. As such, it is important to construct a compact higher-level abstraction network (metaschema) [6, 23] of the SN such that the metaschema can serve as the first view of the UMLS. A metaschema is based on a partition of the SN (or the ESN), which groups similar semantic types into a semantic-type group according to some criteria. Hence, it is interesting to study different partitioning techniques to yield different metaschemas. It is also important to develop an efficient method to evaluate these metaschemas.

1.2 Literature Review

In [38], McCray presented the structure of the UMLS's Semantic Network (SN) and described how to represent biomedical knowledge in the SN. The purpose of the SN is to provide a consistent categorization of all concepts in the META and to provide useful links between these concepts at the level of semantic types. The SN is organized in a hierarchy of two trees by the IS-A links. The appropriate place of a semantic type in the SN is determined by its definition, regardless of whether that definition is based on inherent or attributed features. Figure 1.1 shows part of the **Event** portion of the SN's IS-A hierarchy. The IS-A link in the SN allows a child semantic type in the IS-A hierarchy to inherit semantic relationships from its parent to ensure efficient information storage. When a semantic relationship is not allowed to be inherited, then the "DNI" or the "blocked" mechanism is used to prevent inheritance of this relationship.

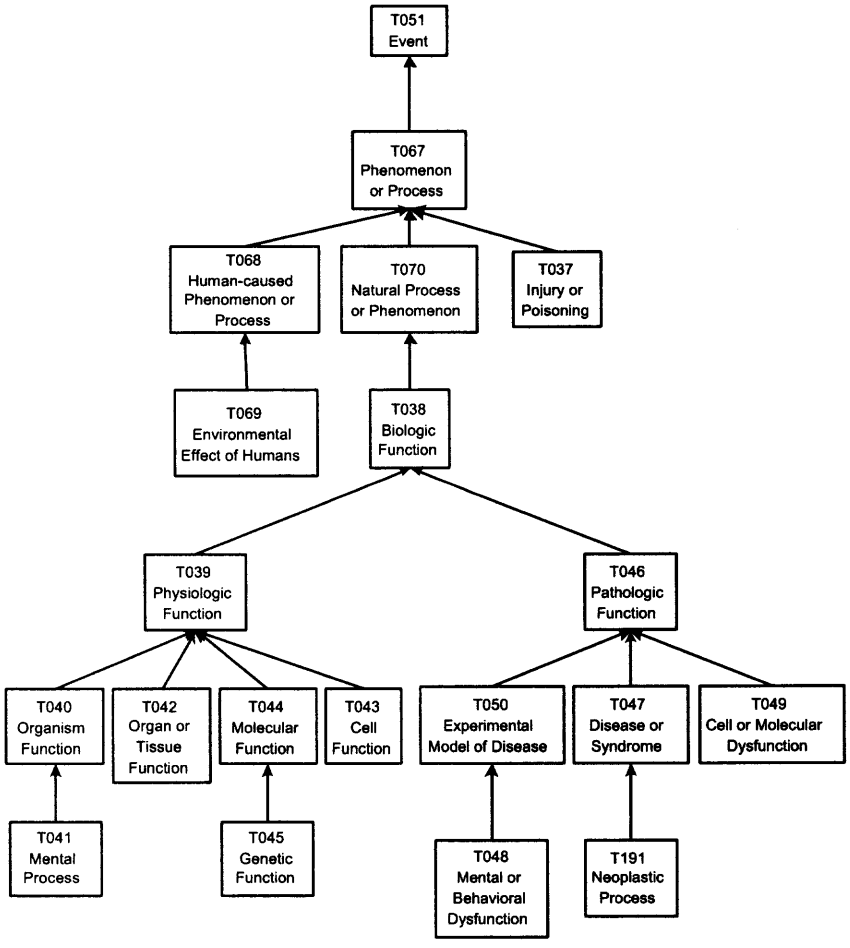


Figure 1.1 Part of **Event** portion of the SN's IS-A hierarchy.

Besides the IS-A links, semantic types in the SN are connected by semantic relationship of 53 kinds. Semantic relationships in the SN fall into five major categories which are themselves relationships: *Physically_related_to*, *temporally_related_to*, *functionally_related_to*, *spatially_related_to*, and *conceptually_related_to*. Semantic relationships, like semantic types, are given precise definitions and organized in a tree hierarchy rooted at *associated_with* which have the above five children. For example, *performs* is a child of *functionally_related_to* which is in turn a child of *associated_with*. Figure 1.2 shows *functionally_related_to* part of the relationship hierarchy.

In order to make the SN a high-level abstraction of the META, each concept in the

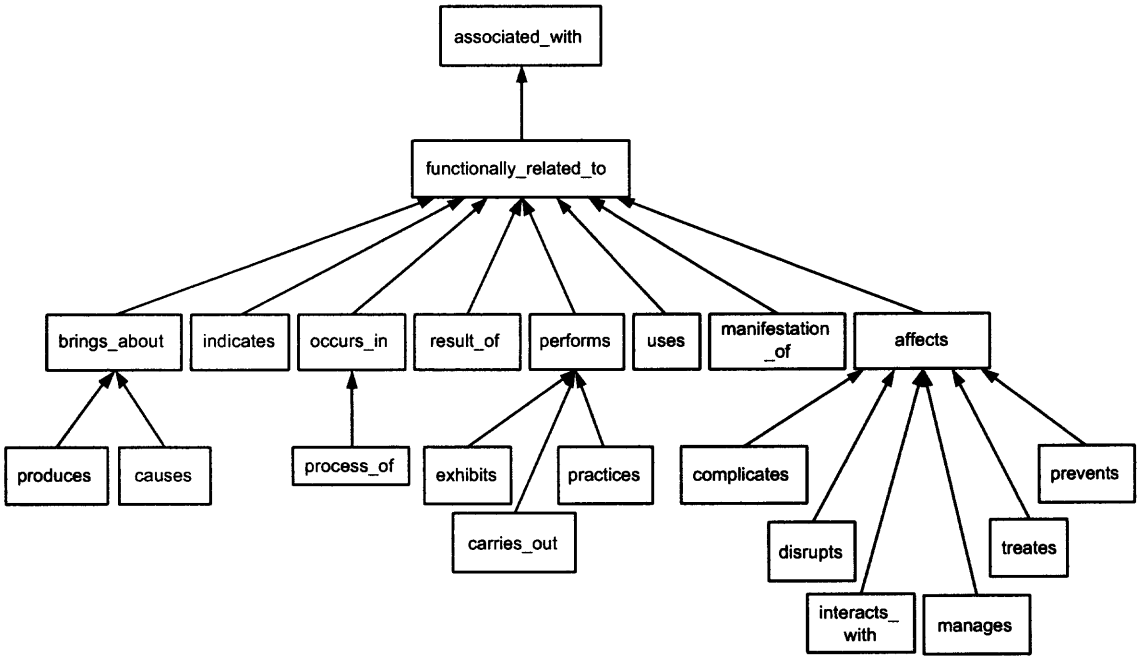


Figure 1.2 *functionally_related_to* part of the relationship hierarchy.

META is assigned to one or more semantic types in the SN according to the two principles in Section 1.1. A concept may be assigned to multiple semantic types since the concept might appear in different contexts in the source vocabulary and those contexts signal different semantic types in the SN. In this way, the SN provides a consistent categorization of the META's concept and thus serves as an abstraction of the META.

As stated in Section 1.1, an important goal of the research is to enrich the SN's hierarchy to accommodate multiple subsumption knowledge. In fact, the same idea was recommended to the NLM by other researchers in [11]. This paper addressed the suitability of the UMLS content for representing patient information in the large hospital-based Clinical Information System (CIS) at Columbia-Presbyterian Medical Center (CPMC). The SN was compared with the Medical Entities Dictionary (MED) developed at Columbia University [10, 32] in several specific aspects such as MED database entities vs. SN semantic types,

MED classes vs. SN semantic types, and MED semantic links vs. SN semantic relationships, etc. For example, **Event** in the SN is different from "event" in the MED in that in the SN events are broadly defined as any actions while in the MED events are the occurrences of actions with specific time or person. Although the SN provides a good coverage for classes of terms in the MED, some classes are still missing from the SN (i.e., Specimen and Drug Form). Most classes of data in the CPMC CIS and the important classes in the MED are well represented in the SN. But because the SN is based on concept classes and relationships from medical information sources (i.e., MEDLINE) and not from a clinical database, there are still some dissimilarities. Based on the comparison results, several recommendations were proposed to the NLM. Enriching the SN into a DAG structure was one of them. Multiple hierarchies in the SN would be helpful in overcoming the above dissimilarities.

During the enrichment of the SN, anatomical knowledge from the Foundational Model of Anatomy (FMA) developed at the University of Washington was used in this dissertation. A detailed introduction to the FMA is presented in [53, 54]. The FMA is an evolving ontology, containing entities and relationships necessary for coherently and consistently modeling knowledge about the human anatomy. The FMA is implemented in a frame-based system and stored in a relational database as a reusable and generalizable resource of anatomical knowledge. It is independent of any specific application and can be filtered according to different needs of different applications. The FMA currently contains 70,000 distinct anatomical concepts and 1.5 million relationship occurrences of over 170 kinds [43]. These anatomical concepts represent various anatomical structures

ranging in size from some macromolecular complexes and cell components to major body parts. Based on ten fundamental principles, the high-level scheme of the FMA now has four components: Anatomy Taxonomy, Anatomical Structural Abstraction, Anatomical Transformation Abstraction, and Metaknowledge. This organization captures the necessary information for describing the anatomy of the whole body as well as any structure or space that constitutes the body. The Protégé-2000 ontology editing and knowledge acquisition environment [47] was chosen for encoding the FMA, because of its frame-based architecture. An anatomical concept and a set of attribute (property)/value pairs of this concept is represented by a frame in Protégé-2000. Attributes and relationships of an anatomical concept are expressed by slots in the frame. A facet of a frame imposes constraints on the value that a slot can have. These frames are assigned as instances to different metaclasses which represent higher-level abstractions of the frames. Relationships and attributes can be inherited along hierarchical (IS-A) relationships among metaclasses. The concepts in the FMA will be added to the UMLS as an extension of the anatomy component of the UMLS. The FMA can be used as a reference ontology for bioinformatics, since its concept representation is independent of any specific applications. Furthermore, it is processable by computers and therefore, provides for machine-based inference.

Besides enriching the SN to accommodate more subsumption knowledge, there are other effort to extend the SN to accommodate concepts of other biomedical field. Since the UMLS is a large-scale knowledge source designed to facilitate retrieval and integration of information from multiple-readable biomedical information resources, it is quite important to insure that integrating a terminology of a new biomedical area into the UMLS is pos-

sible. In [62], the SN was extended to integrate concepts important for genomic research. Previously an ontology concerning genomic concepts was developed to specify important concepts and their relationships in the genomic domain. Based on this ontology, the SN was analyzed and extended to integrate these genomic concepts and relationships. Therefore, some concepts in the ontology were manually mapped to existing semantic types in the SN. This mapping was done by examining the network and looking at the definitions of semantic types provided by the UMLS. For those concepts that do not have corresponding semantic types in the SN, new semantic types were created and attached at appropriate places in the SN's hierarchy. As a result, it was observed that over 30 existing semantic types and most of the existing semantic relationships in the SN are relevant to the genome project. Six new semantic types were added and 16 new semantic relationships were introduced into the SN. The successful mapping between the genomic ontology and the SN shows the suitability and the adaptability of the SN for the representation of the growing domain of biomedical knowledge.

One of the methodologies to enrich the SN is based on the partition presented in [39]. In that paper, McCray, Burgun, and Bodenreider presented a partition of the SN into 15 groups, with each group representing a subject area. This partition was derived externally since the authors first picked different subject areas in medicine and then assigned each semantic type to a proper subject area. The groupings of semantic types were subject to a set of general principles including, *semantic validity* (the groups must be semantically coherent); *parsimony* (the number of groups should be as small as possible); *exclusivity* (each semantic type must belong to only one group); *completeness* (the groups

must cover the full domain); *naturalness* (the groups characterize the domain in a way that is acceptable to a domain expert); and *utility* (the groups must be useful for some purpose). Table 1.1 shows the 15 groups resulting from applying these rules. Two possible methods were presented to measure the degree of semantic coherence for each group in the resulting partition. One way is too see if all semantic types in a group are hierarchically related to each other. The other way is to analyze the semantic relationships exhibited by semantic types in a given group. The resulting partition can be used for display purposes to reduce conceptual complexity and to provide a broad overview of the SN. It might be also helpful to discover inconsistencies in the representation of the SN.

Other important goals of the research in this dissertation are to study different partitioning techniques to construct metaschemas as upper-level abstract views of the SN, to help user comprehension and to study the applications of these metaschemas. One possible application of a metaschema is to provide a partial graphic view of a specific subject area that is of interest to a user [49]. Another pertinent application is to audit the UMLS concept categorizations using a metaschema. Why can a metaschema be used to detect concept categorization errors in the UMLS? Every meta-semantic type in a metaschema represents a specific subject area of the SN, and thus of the underlying META. Therefore, concepts assigned to different semantic types in the SN may also be assigned to several meta-semantic types. It is more likely that a concept will be erroneously assigned to semantic types of different meta-semantic types than to semantic types of the same meta-semantic type because of larger semantic distance between different meta-semantic types. Based on this hypothesis, [18] concentrated on auditing concepts that were assigned to different meta-semantic

Table 1.1 Partition of the SN Presented in [39]

Group	Semantic Types in Group
Activities and Behaviors	Occupational Activity; Behavior; Activity; Event; Individual Behavior; Daily or Recreational Activity; Governmental or Regulatory Activity; Machine Activity; Social Behavior
Anatomy	Body Location or Region; Body System; Body Part, Organ, or Organ Component; Anatomical Structure; Embryonic Structure; Tissue; Cell; Body Space or Junction; Fully Formed Anatomical Structure; Body Substance; Cell Component
Chemicals and Drugs	Biomedical or Dental Material; Biologically Active Substance; Organic Chemical; Pharmacologic Substance; Chemical; Enzyme; Neuroreactive Substance or Biogenic Amine; Hormone; Chemical Viewed Structurally; Vitamin; Immunologic Factor; Indicator, Reagent, or Diagnostic Aid; Clinical Drug; Inorganic Chemical; Element, Ion, or Isotope; Antibiotic; Hazardous or Poisonous Substance; Receptor; Steroid; Eicosanoid; Chemical Viewed Functionally; Nucleic Acid, Nucleoside, or Nucleotide; Organophosphorus Compound; Amino Acid, Peptide, or Protein; Carbohydrate; Lipid
Concepts and Ideas	Classification; Quantitative Concept; Qualitative Concept; Temporal Concept; Idea or Concept; Conceptual Entity; Group Attribute; Language; Intellectual Product; Spatial Concept; Functional Concept; Regulation or Law
Devices	Research Device; Medical Device
Disorders	Finding; Injury or Poisoning; Pathologic Function; Experimental Model of Disease; Disease or Syndrome; Sign or Symptom; Anatomical Abnormality; Neoplastic Process; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Acquired Abnormality; Congenital Abnormality
Genes and Molecular Sequences	Molecular Sequence; Amino Acid Sequence; Carbohydrate Sequence; Nucleotide Sequence; Gene or Genome
Geographic Areas	Geographic Area
Living Beings	Organism; Population Group; Fungus; Alga; Virus; Human; Plant; Archaeon; Group; Professional or Occupational Group; Reptile; Family Group; Age Group; Patient or Disabled Group; Rickettsia or Chlamydia; Amphibian; Mammal; Fish; Bird; Animal; Vertebrate; Invertebrate; Bacterium
Objects	Entity; Manufactured Object; Physical Object; Substance; Food
Occupations	Occupation or Discipline; Biomedical Occupation or Discipline
Organizations	Health Care Related Organization; Self-help or Relief Organization; Professional Society; Organization
Phenomena	Phenomenon or Process; Human-caused Phenomenon or Process; Laboratory or Test Result; Natural Phenomenon or Process; Biologic Function; Environmental Effect of Humans
Physiology	Organism Attribute; Cell Function; Organ or Tissue Function; Organism Function; Genetic Function; Molecular Function; Physiologic Function; Mental Process; Clinical Attribute
Procedures	Laboratory Procedure; Health Care Activity; Molecular Biology Research Technique; Diagnostic Procedure; Educational Activity; Research Activity; Therapeutic or Preventive Procedure

types. In the auditing process, all concepts belonging to two or more semantic types where these semantic types are of different meta-semantic types were identified and reviewed by a domain expert. Different kinds of categorization errors were exposed in the process of review. The results showed that the metaschema is efficient in identifying groups of highly erroneous concepts in auditing the UMLS for concept categorization errors.

This dissertation also investigated partitioning techniques that can be used to partition the SN into semantic-type groups representing subject areas in the SN. Grouping semantic types into different subject areas is helpful in improving visualization and displaying the knowledge in a particular domain, and in other applications where high-level categories are sometimes needed and useful. Bodenreider and McCray presented in [2] a new technique to explore semantic-type groups through visual approach while assessing semantic coherence of these groups using their semantic relationships as important indicators. They first exhaustively examined semantic relationships existing between each pair of semantic-type groups, determining the nature and number of relationships for each pair. Based on this investigation, a radial diagram was developed to display the number of semantic relationships between a given semantic-type group and any other group in the partition of [39]. The radial diagrams proved useful for comparing the profiles and natures of various groups and exposing some inconsistence and errors in the SN's design. Semantic relationships exhibited among semantic types in a given group were studied. It was shown that a given relationship usually applies only to a limited number of semantic-type groups, which indicates that the constitution of the groups takes into account the semantics of the types as well as that of the relationships. If each sematic-type group is represented

by a node, then the number of semantic relationships exhibited by each group can be represented as the in-degree and out-degree of the corresponding node in a graph. Thus, for a specific kind of semantic relationship, a directed acyclic graph can be used to demonstrate the distribution of this relationship among semantic-type group pairs of the partition. For a given relationship, the semantic coherence should translate into a small number of nodes (pivot nodes) with high in-degree or out-degree, while most nodes are of degree 1 or 0 since such a relationship is only exhibited by a limited number of groups. This means that semantically coherent groups should result in a small number of subsets for a given kind of relationship in the whole SN. It was shown that semantic-type groups resulting from random partitions of the SN do not exhibit such a property. Finally correspondence analysis [15] was used for studying the association between semantic types and relationships of the semantic-type groups in the partition. The result was displayed in a two-dimensional graphical representation which was proven quite helpful in data visualization and knowledge navigation [26, 44]. The result showed that most of the semantic-type groups in the partition of [39] are semantically or semi-semantically coherent. The correspondence analysis also presented some improvement suggestions for the partition of [39].

1.3 Dissertation Overview

This dissertation is an amalgamation of four papers. They are organized as follows. The accepted journal paper [65] is the basis for Chapter 2 which presents the methodologies used to enrich the two-tree structured SN into a DAG structured ESN with a multiple subsumption hierarchy. Chapter 3 is based on a submitted journal paper [63] which demonstrates

the impact of the ESN's multiple subsumption hierarchy on the ESN's relationship distribution and concept configuration and presents the whole ESN, including its IS-A hierarchy, relationship distribution and concept configuration. Chapter 4 is based on the published journal paper [67] which extends a metaschema to be applicable to a DAG-structure network such as the ESN and presents the *qualified metaschema* and the *cohesive metaschema* of the ESN. The submitted journal paper [66], which is the basis for Chapter 5, presents a *lexical metaschema* for the SN, based on the lexical partition. The lexical partition groups lexically-related semantic types in the same semantic-type group. The evaluation of the lexical metaschema, using experts' opinions, is also described in this chapter. Finally, conclusions of this dissertation are in Chapter 6. Chapter 7 describes some future work based on this dissertation. A brief overview of these four research issues and how the fit together will be presented in the next section.

1.3.1 Enrich the Semantic Network's Hierarchy

In the first phase of research, the restriction of the SN's current two-tree structure is presented and analyzed. The current SN's tree structure does not allow semantic types to have multiple parents or ancestors, which should be a natural situation in the medical terminology. To enrich the SN tree-structure into a DAG structure allowing multiple parents, extra IS-A links have to be identified and added.

Two methodologies are presented in Chapter 2 to enrich the SN's IS-A hierarchy. The first methodology is based on a previous partition of the SN [39] which partitioned the SN into semantic-type groups representing different subject areas. The second methodology is based on the string matching between various semantic types' definitions and names.

Both methodologies identified and added extra viable IS-A relationships to the SN, after reviewed by a domain expert.

The addition of these IS-A links changes the SN from a two-tree network into a multiple subsumption hierarchy network. This new network is referred to as the Enriched Semantic Network (ESN).

1.3.2 ESN's Relationship Distribution and Concept Configuration

The ESN's multiple subsumption hierarchy allows semantic types to have multiple parents and thus to inherit new relationships from their new parents or ancestors. This kind of inheritance, called multiple inheritance, makes the ESN's relationship distribution different from that of the SN. Chapter 3 presents a technique to derive the ESN's relationship distribution based on that of the SN.

The ESN serves as a high-level abstraction of the underlying META, with each concept being assigned to one or more semantic types. It is impossible to assign the 900,000 concepts to the ESN's semantic types by hand. A mapping function is defined to derive the ESN's concept configuration based on that of the SN. This mapping function ensures that the ESN's concept configuration comply with the principle that each concept be explicitly assigned to the lowest (or most specialized) semantic type in the IS-A hierarchy, that is, this configuration is free from any redundant classifications. The whole ESN, including its IS-A hierarchy, relationship distribution and concept configuration, is also presented.

1.3.3 ESN's Metaschemas

The ESN, which is more complex than the SN, is a large network, which is difficult for user orientation and comprehension purposes. Therefore, it is helpful to build higher-level abstraction for the ESN to help user orientation. The previously mentioned metaschema can function as such an abstraction. In Chapter 4, the metaschema notion, which was developed for tree-structured networks, is extended to be applicable to DAG-structured networks such as the ESN. The requirements and derivation of such a metaschema are provided.

Two metaschemas are derived for the ESN in Chapter 4. One is the *Qualified metaschema* (in short, Q-metaschema) which is derived from the partition of the ESN in [65]. Another is the *Cohesive metaschema* (in short, C-metaschema) which is derived from a partition of the ESN based on the relationship structure of its semantic types. The two metaschemas are compared and evaluated. Applications of a metaschema, for example, the Q-metaschema, are demonstrated. The author shows how a user can take advantage of a metaschema to help him with comprehending the ESN and with navigation in the UMLS.

1.3.4 SN's Lexical Metaschema

A metaschema is always derived from a partition of the network. For the SN, different partitions yielded different metaschemas. Chapter 5 introduces a new metaschema, named the *lexical metaschema*, which is based on a lexical partition of the SN. The lexical partition groups semantic types that are lexically related into the same semantic-type group. Each such semantic-type group is represented by a meta-semantic type in the lexical metaschema. Chapter 5 presents the detailed derivation of the lexical partition and the lexical metaschema.

To evaluate the lexical metaschema, experts' responses in a study are used. *Cumulative metaschemas*, which represent different levels of experts' aggregation are built and compared to the lexical metaschema. Qualitative evaluation techniques are also used to measure the similarity of the lexical metaschema and the cumulative metaschemas. Results show that the lexical metaschema is similar to the cumulative metaschema which represents a simple majority of the experts' responses.

CHAPTER 2

ENRICHING THE SEMANTIC NETWORK'S STRUCTURE

2.1 Introduction

As analyzed in Section 1.1, the UMLS's Semantic Network's tree structure is restrictive since it does not allow multiple parents when warranted. Some semantic types could naturally be specializations of more than one semantic type. For example, **Gene or Genome** could conceptually be a child of two semantic types: one is its current parent **Fully Formed Anatomical Structure**; the other is **Molecular Sequence**. Hence, **Gene or Genome** should inherit from **Molecular Sequence** the semantic relationship *result_of* to **Mental Process**. In a case such as this, the modeling of the SN omits an aspect of current medical knowledge. It is quite natural to enrich the SN to accommodate this omitted subsumption knowledge in order to provide a more accurate modeling of the current medical knowledge. In [11], a study was conducted to evaluate how well the UMLS could support clinical information systems at Columbia-Presbyterian Medical Center as compared to the local Medical Entities Dictionary (MED) [10, 32]. A recommendation resulting from this study was that multiple parents be permitted in the SN.

Many researchers have suggested that concept-oriented [7, 9] and logic-based [3, 4, 51, 50] approaches are beneficial for creating categorical terminological structures [55], especially when their purpose is to support, as the UMLS does, cross-thesaurus mappings [55, 12]. However, the SN does not provide sufficient logic-based structures to apply such methods; alternative methods to improve the consistency and utility of the SN must be developed.

In this chapter, two methodologies are presented to structurally enrich the SN by transforming its hierarchy into a directed acyclic graph (DAG) structure that allows multiple parents. The methodologies are based on the identification of viable new IS-A relationships currently not included in the SN. These omissions may have been due to the tree-structure restriction on the SN, noted above. New semantic types are added to the SN as necessary to accommodate the new multiple subsumption framework. In the first methodology, the identification of new IS-As is guided by imposing connectivity on an existing partition of the SN [39]. In the second methodology, the identification is based on partial string matching between names of semantic types and the definitions of other semantic types. These identified potential IS-A relationships are then reviewed by a domain expert to decide whether they are semantically valid. With the addition of these new IS-A relationships, a new DAG version of the SN, referred to as the *Enriched Semantic Network* (ESN), is derived. Furthermore, a partition of the ESN consisting of groups that each have a tree structure is obtained. The ESN serves as an enhanced abstraction of the UMLS. The accompanying partition enables the creation of a metaschema [67], an additional abstract layer of the ESN that can help users in their orientation to the UMLS.

Section 2.2 presents the two methodologies to enrich the Semantic Network. In Section 2.2.1, the first methodology of identifying new IS-As is guided by the process of imposing connectivity on an existing partition of the SN [39]. In Section 2.2.2 the second methodology of identifying extra viable IS-A links is based on partial string matching between names of semantic types and the definitions of other semantic types. These identified potential IS-A relationships are then reviewed by a domain expert to decide whether

they are semantically valid. With the addition of these new IS-A relationships, a new DAG version of the SN, referred to as the Enriched Semantic Network (*ESN*), is obtained. Section 2.3 presents the results for the two methodologies and the final hierarchy of the ESN. Furthermore, a partition of the ESN consisting of groups, each of which has an internal tree structure is also presented in Section 2.3. The ESN serves as an enhanced abstraction of the UMLS. The accompanying partition enables the creation of a metaschema [67], an additional abstract layer of the ESN, that can help users in their orientation to the UMLS. Section 2.4 contains a discussion of the advantages of the ESN and an evaluation of the two methodologies. Section 2.5 contains conclusions.

2.2 Methods to Enhance the SN's IS-A Hierarchy

In this section, two methodologies are presented to enhance the SN's IS-A hierarchy. In Section 2.2.1 the author presents the first method, based on imposing connectivity on a previous partition in [39]. Four transformations are developed to explore this situation. Section 2.2.2 contains the second method based on the string matching between definitions and names of various semantic types in the SN.

2.2.1 Imposing Connectivity on an SN Partition

2.2.1.1 Basis

In [39], McCray, Burgun, and Bodenreider presented a partition of the SN into 15 groups, with each group representing a subject area. Six principles that such a partition should satisfy were proposed. Among all of the principles, semantic validity is perhaps the most important one [39]. Without semantic coherence, it is hard to see how useful such a parti-

tion would be for any given purpose. Therefore, it is quite important to assess the degree of semantic coherence for each group in the resulting partition. As stated in [39] "One way to measure semantic validity is to assess the degree to which the types in a group are hierarchically related to each other. This is so, since parents and children in a hierarchy share essential properties." In other words, one way for a group to satisfy semantic validity requires that all semantic types in the group together with the IS-A links connecting them be a connected subgraph [39] of the SN. This is referred to as the *connectivity property*. Since the SN's IS-A hierarchy consists of two trees, such a connected subgraph must form a tree with a unique root.

In the analysis of [39], it was noted that: "In some cases, it was not possible to resolve anomalies in our attempt to create a coherent and semantically valid set of groupings." In fact, some of the partition's groups do not satisfy the connectivity property. Such groups contain a forest, comprising two or more trees, or perhaps isolated semantic types. (Some groups have both.) For example, the *Physiology*[1] group contains a forest of two trees (Figure 2.1). There are no hierarchical relationships between a semantic type of one tree and a semantic type of the other tree. Therefore, the *Physiology* group is not connected.

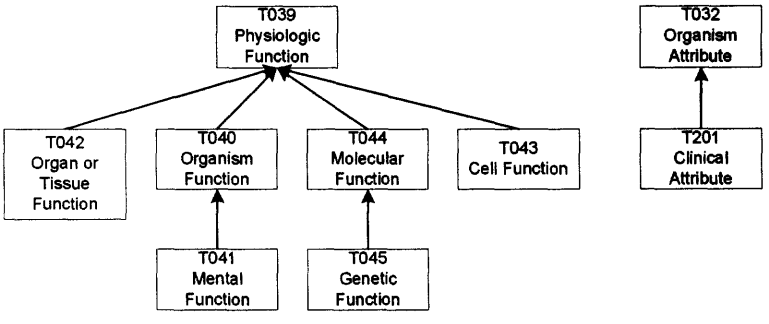


Figure 2.1 *Physiology* group.

[1]Group names will be written in Italics in this chapter.

In previous work [6], an alternative partition of the SN was presented based on the sets of relationships exhibited by its semantic types. In that technique, the hierarchy of each group of the partition was required to be a tree, exhibiting the connectivity property. In this way, a partition that is strictly semantically uniform was obtained. A difference between the partition of [39] and that of [6] is that the connectivity is only a preferred, not required, property in [39], while it is required and enforced in [6].

In the semantic technique, the partition of [39] will be used as a basis for augmenting the SN's hierarchy, and, in particular, for identifying new viable IS-A relationships. The basic idea is to bridge the gap between the two partitioning techniques by imposing the connectivity property on the partition of [39]. In order to convert the disconnected groups of [39] into connected groups, additional IS-As will be identified and added to the SN. This will yield a first version of the desired multiple subsumption hierarchy and an accompanying partition. Analysis of the definitions of semantic types within each disconnected group will guide the introduction of the new IS-A links. In this context, four kinds of transformations will be developed with respect to the groups of [39]. Another methodology employing exact string matching will then be utilized in a following subsection.

2.2.1.2 Four Transformations to Identify New IS-A Links

The possible transformations that can be applied to disconnected groups to make them connected are listed in the following. The choice of which transformation to utilize is based on reviews of the definitions of all semantic types within a group.

IS-A Addition Transformation: Identify a viable IS-A and add it to transform the group into a connected subtree. □

Split Transformation: Split a group into several groups, each of which is either a rooted tree structure or can be transformed into a rooted tree structure by adding IS-A relationships. □

Root-Addition Transformation: Create a new semantic type that will be an ancestor of all roots in the group. Make the new semantic type the group's root by adding additional IS-A relationships from all the roots of the group's connected components (either directly or via more new semantic types, if necessary). □

Root-Moving Transformation: Locate a semantic type (from another group) that is a lowest common ancestor of the roots of all the disconnected group's subtrees and/or isolated semantic types. Move that lowest common ancestor into the disconnected group, making it the root of the group and thereby connecting the group. Also, move all the new root's existing descendants into the group. □

The new network obtained by applying these transformations is called the *Enriched Semantic Network (ESN)*. It has a DAG structure rather than a two-tree structure. In the following, the various transformations will be demonstrated and their impact on different disconnected groups will also be analyzed.

As an example, the group *Disorders* demonstrates the IS-A addition transformation and split transformation (Figure 2.2). This group contains twelve semantic types, eleven of which belong to three trees rooted at **Pathologic Function**, **Anatomical Abnormality**, and **Finding**, respectively. **Injury or Poisoning** is an isolated member of the group. Clearly, *Disorders* does not satisfy connectivity.

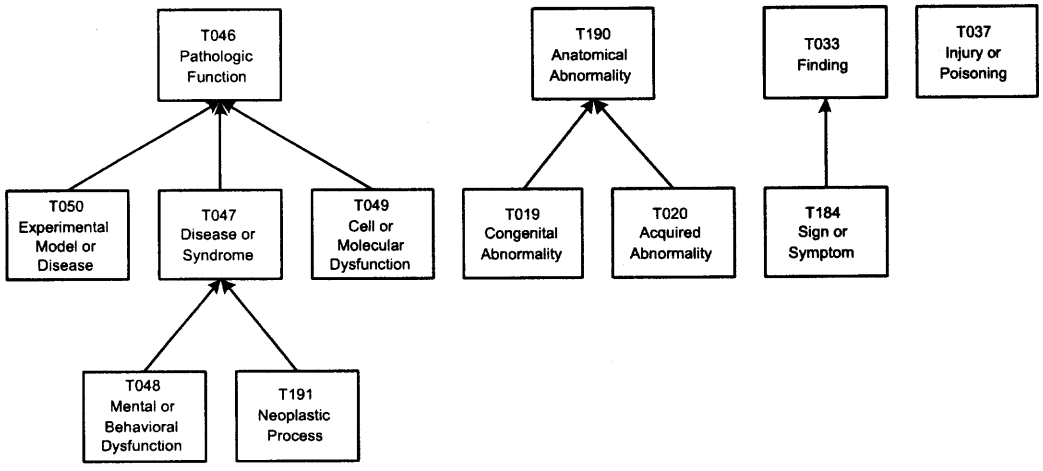


Figure 2.2 *Disorders* group.

The IS-A addition transformation is first applied to this group to connect **Injury or Poisoning** to the tree rooted at **Pathologic Function**. It is observed that **Injury or Poisoning** should have a subsumption relationship to **Disease or Syndrome** and inherit its semantic relationships. Thus an IS-A link is added to capture this. Since in the original SN, **Disease or Syndrome** is a descendant of **Phenomenon or Process**, the original IS-A from **Injury or Poisoning** to **Phenomenon or Process** can be removed because it can be inferred transitively via the new IS-A link from **Injury or Poisoning** to **Disease or Syndrome**.

At this point, the group is still a collection of disconnected trees. To rectify this, the split transformation is applied to form three new groups. According to the definitions of the twelve semantic types, it is clear that **Pathologic Function** and its six descendant semantic types, including the new descendant **Injury or Poisoning**, emphasize phenomenon or process and are in the **Event** tree, while the remaining semantic types emphasize an entity or object and are in the **Entity** tree. Furthermore, **Anatomical Abnormality** and its children are descendants of **Physical Object**, while **Finding** and its child are conceptual entities. Therefore, it is natural to partition this group into three smaller connected groups,

each comprising a tree. These groups, *Pathologic Function*, *Anatomical Abnormality*, and *Finding*, are shown in Figure 2.3. Note that using here a Root-Addition transformation for all or any two trees is not an option since this new root could not be placed anywhere in the SN due to the differences in the contents of the trees. The new groups are named after their roots.

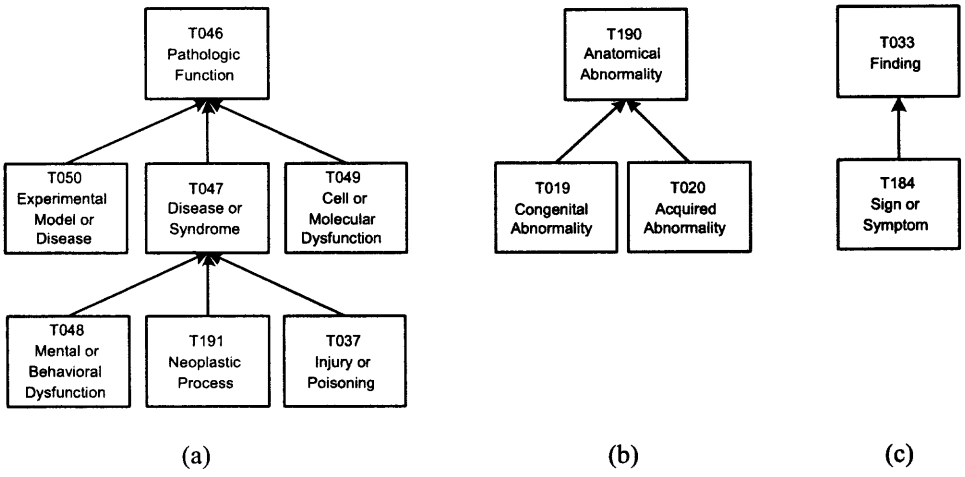


Figure 2.3 Three new groups: (a) *Pathologic Function*, (b) *Anatomical Abnormality*, and (c) *Finding* (derived via IS-A addition transformation and split transformation).

In the next example, the *Anatomy* group undergoes a root-addition transformation; that is, new semantic types are added to make the group connected. The group contains a tree of seven semantic types rooted at **Anatomical Structure** and four isolated semantic types, **Body Substance**, **Body System**, **Body Location or Region**, and **Body Space or Junction** (Figure 2.4). In carrying out this transformation, the analysis of [43] for definitions of anatomical concepts is used as reference. For example, the new semantic type **Material Physical Anatomical Entity** is defined as "IS-A Physical Anatomical Entity which has a mass" [43]. **Body Substance** is not an **Anatomical Structure** since it does not have a 3D shape, but it is a **Material Physical Anatomical Entity** since it has mass. Thus, both

Body Substance and **Anatomical Structure** are made children of the new semantic type **Material Physical Anatomical Entity**.

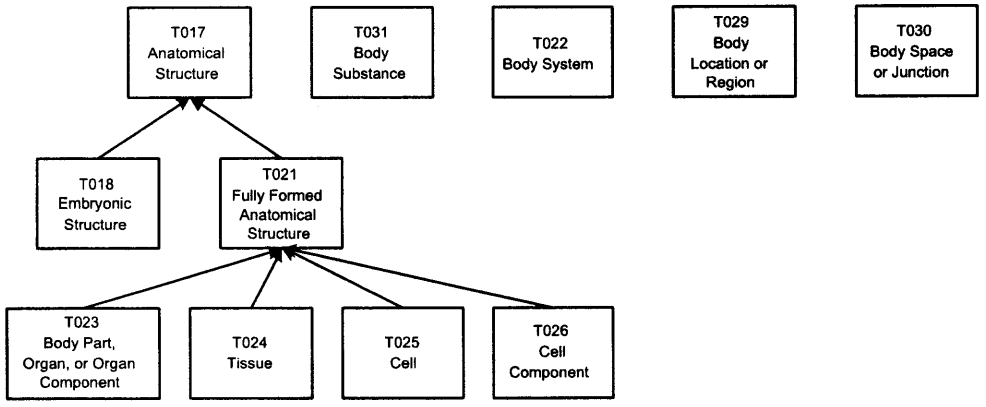


Figure 2.4 *Anatomy* group.

Furthermore, **Body Space or Junction** is not a **Material Physical Anatomical Entity**, but it is a **Physical Anatomical Entity** (defined in [43] to have spatial dimensions). Hence, both **Body Space or Junction** and **Material Physical Anatomical Entity** are made children of the newly introduced **Physical Anatomical Entity**, which in turn IS-A **Physical Object**. The original IS-A from **Anatomical Structure** to **Physical Object** is cut because it can be inferred from the new IS-A from **Anatomical Structure** to **Physical Anatomical Entity**.

On the other hand, **Body Location or Region** and **Body System** have neither mass nor spatial dimension and thus cannot be descendants of **Physical Anatomical Entity**. Nevertheless, both obviously should belong to the *Anatomy* group. Following [43], the new semantic type **Conceptual Anatomical Entity** is introduced, which in turn IS-A **Conceptual Entity**, to complement **Physical Anatomical Entity** and serve as the parent of **Body Location or Region** and **Body System**.

Finally, the new semantic type **Anatomical Entity** is added as the parent of both **Physical Anatomical Entity** and **Conceptual Anatomical Entity**. In turn, **Anatomical Entity** IS-A **Entity**. In this way, the whole *Anatomy* group is transformed into the new group *Anatomical Entity* (Figure 2.5). The dashed rectangles in the figure represent the newly added semantic types, and the dashed arrows represent the newly added IS-A links. It is important to note that each of the four new semantic types should have at least the corresponding concepts suggested in [43] assigned to it. (These concepts have been submitted to the NLM for inclusion in the next UMLS release.[2])

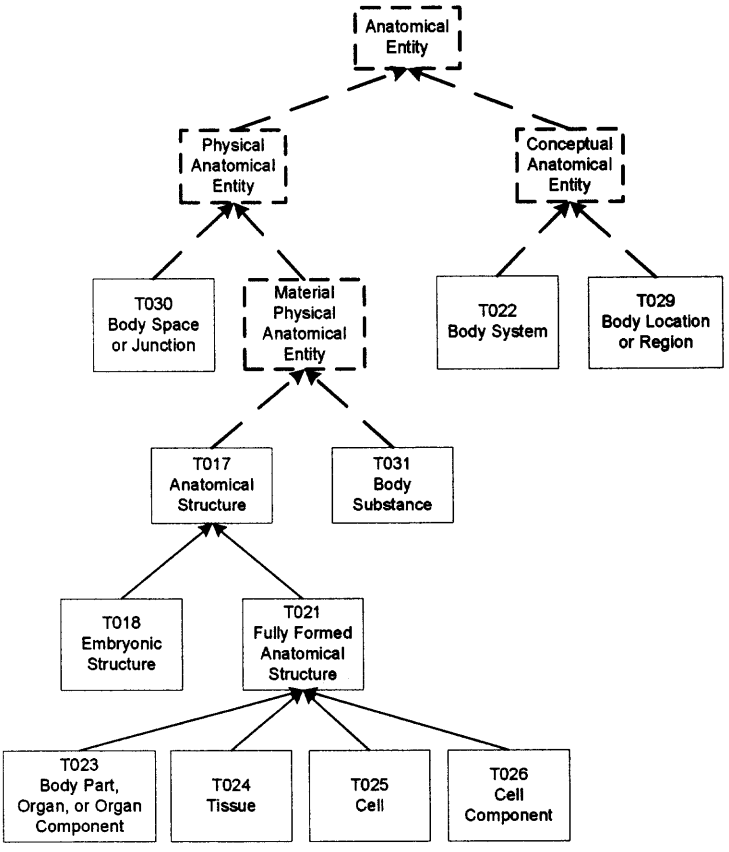


Figure 2.5 *Anatomical Entity* group.

[2]C. Rosse, personal communication, 2002.

In the next example, the Root-moving transformation is applied to the disconnected *Procedures* group to make it connected. The group contains seven semantic types, with two trees rooted at **Health Care Activity** and **Research Activity**, respectively, and the isolated **Educational Activity** (Figure 2.6). These three are children of **Occupational Activity**, which has another child **Governmental or Regulatory Activity**. Both of these semantic types, in turn, belong to the *Activities and Behaviors* group. In the context of the UMLS, these five semantic types refer to health-care related issues. They describe activities of health-care professionals. Thus, **Occupational Activity**, the lowest ancestor of the seven semantic types in the group, and its child **Governmental or Regulatory Activity** are moved to this group. By doing this, the group is transformed into the new *Occupational Activity* connected group (Figure 2.7).

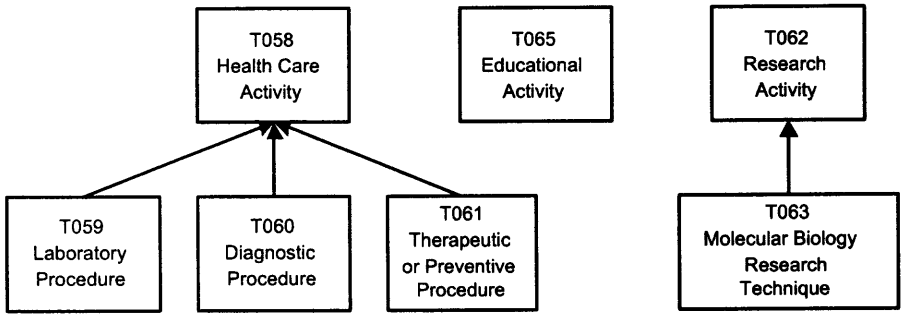


Figure 2.6 *Procedures* group.

2.2.2 String Matching

Additional IS-A links can be found by using string matching involving names and definitions of various semantic types in the SN. To be more formal, a string match is defined as follows:

Definition (String Match): A string match from a semantic type $\mathbf{T}_1$ to another semantic

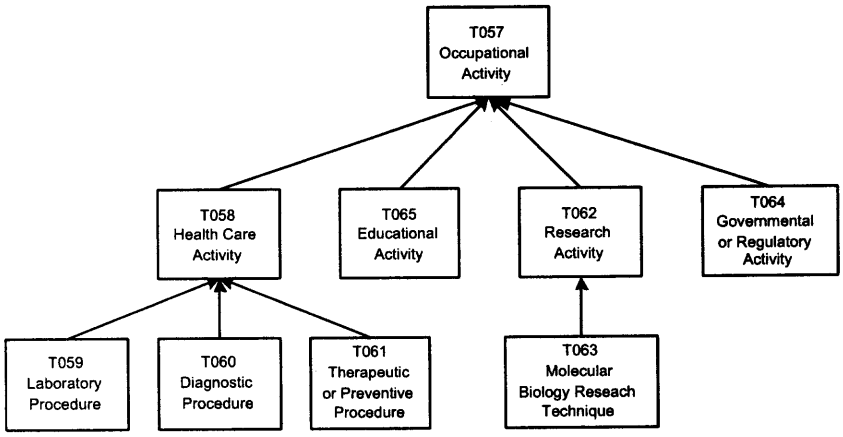


Figure 2.7 *Occupational Activity* group.

type $\mathbf{T}_2$ is a triple ($\mathbf{T}_1$; $\mathbf{T}_2$; S) such that S is a string appearing both in the definition of $\mathbf{T}_1$ and in the name of $\mathbf{T}_2$. S is called the common string and must contain one or more (not necessarily consecutive) complete words (ignoring case). □

For example, the definition of **Plant** contains the word "organism" which happens to be the name of a semantic type. Hence, a string match (**Plant**; **Organism**; "organism") exists.

The motivation for using this kind of string matching to find viable new IS-A links is based on the evaluation of string matches among the 132 pairs of semantic types that currently have IS-A relationships between them in the SN. By analyzing the definitions of the children in the pairs, there are string matches from 88 children to their respective parents. The string match (**Plant**; **Organism**; "organism") is one of them. Thus the sensitivity of this approach with *known* IS-A links is 67%. This finding leads to the following observation.

Observation: If $\mathbf{T}_1$ IS-A $\mathbf{T}_2$, then there is a high likelihood of a string match from $\mathbf{T}_1$ to $\mathbf{T}_2$. □

This leads to formulate the inverse hypothesis.

Hypothesis: If there is a string match from one semantic type to another, then it is likely to imply a viable subsumption relationship between them. □

Based on this hypothesis, the string matching method is developed to identify additional viable IS-A relationships not already appearing in the SN. This methodology is a human-computer interactive methodology and contains three steps:

Step 1: Preprocess names and definitions of semantic types to obtain the input file;

Step 2: Apply the "AllMatches" algorithm to the input file to get all string matches;

Step 3: Manually review all resulting string matches and determine which constitute additional viable IS-A links between semantic types.

In Step 1, some common techniques from the data mining and information retrieval fields are utilized for the preprocess [24].

Stop-words: All stop-words such as "a," "the," "of," "for," "with," and so on are removed from names and definitions.

Verb variant processing: All verbs and verb variants are removed from definitions of semantic types. In the string matching, consider verbs and verb variants will not be considered in string matching. The reason is that most semantic types' names consists only of nouns, adjectives, and adverbs.

Lexical normalization: The Specialist Lexicon (coupled with highly efficient "lexical variant generator" code) [42] is applied to stem-word variations. All adjectives and

adverbs are converted to nouns, and all plurals are converted to singular forms. Also, uppercase letters are changed to corresponding lowercase.

In Step 2, the following AllMatches algorithm is used to find string matches between any two semantic types not currently connected by a single IS-A link or a path of such links. The input file to the algorithm contains the names and definitions of semantic types after the preprocessing step.

In the description of the AllMatches algorithm, let $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_m$ be all semantic types in the SN. (In the 2002 version, $m = 134$). The notation $DEF(\mathbf{T}_i)$ is used to represent the definition of the semantic type $\mathbf{T}_i$, $1 \leq i \leq 134$, after preprocessing. $NAME(\mathbf{T}_i)$ is used to represent the name of $\mathbf{T}_i$, in the form of a string, after preprocessing. For example, suppose $\mathbf{T}_i$ = **Anatomical Structure**, which is defined as: "a normal or pathological part of the anatomy or structural organization of an organism." After preprocessing, $NAME(\mathbf{T}_i)$ ="anatomy structure" and $DEF(\mathbf{T}_i)$ = "normal pathology part anatomy structure organization organism."

In the following AllMatches algorithm, a list L is used to hold all common strings. The following functions defined for lists are also used in the algorithm:

Length(): Return the number of elements in the list

Retrieve(k): Retrieve the k^{th} element of the list

AllMatches algorithm: Find all string matches in the SN.

```
For (i = 1 to m)
    For all T_j  1 < j < m  &  j ≠ i)
        If (T_j is not the parent or an ancestor of T_i)
        {   L = FindCommonStrings(DEF(T_i), NAME(T_j));
            //write string matches to the output file
            For (k = 1 to L.Length())
```

```
        {   S = L.Retrieve(k);    // get the k^th element
            write (T_i; T_j; S) to output file;
        }
}
```

The function FindCommonStrings(R_1, R_2) is used to find all common strings involving a given pair of strings R_1 and R_2. During a call, R_1 is the definition of a semantic type $\mathbf{T}_i$ in a string format, and R_2 is the name of a semantic type $\mathbf{T}_j$ as a string. For each pair ($\mathbf{T}_i$, $\mathbf{T}_j$) that has no direct IS-A relationship or directed path of IS-A relationships between its components, the function FindCommonStrings(*DEF*($\mathbf{T}_i$) is called, *NAME*($\mathbf{T}_j$)) to get all possible common strings between *DEF*($\mathbf{T}_i$) and *NAME*($\mathbf{T}_j$). Each such common string is inserted into L. A match M is called redundant if its constituent common string S is a substring of another match's common string (again ignoring case). Hence, function FindCommonStrings(*DEF*($\mathbf{T}_i$), *NAME*($\mathbf{T}_j$)) identifies the redundant matches and does not return them. Consequently, L contains no redundant common strings. Finally, all string matches ($\mathbf{T}_i$; $\mathbf{T}_j$; S) are written to the output file. After AllMatches has been executed, the output file will contain all non-redundant string matches between pairs of semantic types not connected by IS-A relationships in the SN.

As an example, consider **Enzyme** whose definition is "a complex chemical, usually a protein, that is produced by living cells and which catalyzes specific biochemical reactions." The AllMatches algorithm finds three string matches:

(**Enzyme**; **Cell**; "cell")

(**Enzyme**; **Cell Component**; "cell")

(**Enzyme**; **Amino Acid, Peptide, or Protein**; "protein")

In Step 3, an expert is called upon to review all resulting string matches to find new

IS-A links not currently appearing in the SN. These newly discovered IS-A links can then be added to the ESN.

As it happens, in the case of the three string matches involving **Enzyme**, the third match implies the existence of a new IS-A link, since any enzyme must be a kind of protein. Hence, **Enzyme** IS-A **Amino Acid, Peptide, or Protein**.

As noted above, the sensitivity of the string matching approach, when applied to *known* IS-A links is 67%. In order to determine the sensitivity of this method for detecting *unknown* IS-A links, a gold standard is established by performing a manual review of randomly generated relationship pairs.

2.3 Results

By applying the two methods in Section 2.2, extra valid IS-A links are identified and added to the SN. These additions enrich the SN's IS-A hierarchy from a two-tree structure to a DAG structure. In this section, results of the application of the two methods are presented in Section 2.3.1 and Section 2.3.2, respectively. Section 2.3.3 contains the summary of the two results and shows the structure of the enriched semantic network.

2.3.1 Results of Imposing Connectivity on the Partition

Besides the three disconnected groups described in Section 2.2.1, the original partition of [39] contains six other disconnected groups. Table 2.1 presents the six groups and the six transformations applied to them. For each such group, the table shows the isolated semantic types or trees that existed in the group and the transformations used. In the second column of the table, each tree in the group is denoted by placing its constituent semantic types in

Table 2.1 Transformations Applied to Six Disconnected Groups of the Partition of [39]

Old Group Name	Isolated Semantic Types and/or Trees in a Disconnected Group	Transformation Type	Transformations Applied	New Group Name
Chemicals and Drugs	**Clinical Drug**	Split Transformation	Split into two connected groups. The group *Clinical Drug* contains just one semantic type.	Two groups: *Chemical*; *Clinical Drug*
Devices	**Research Device**; **Medical Device**	Root-moving Transformation	Move **Manufactured Object** from the *Objects* group and make it the new root of the *Devices* group	*Manufactured Object*
Genes and Molecular Sequences	**Gene or Genome**	IS-A addition Transformation	Add (**Gene or Genome** IS-A **Molecular Sequence**) link	*Molecular Sequence*
Living Beings	{Organism; Fungus; Alga; Virus; Human; Plant; Archaeon; Reptile; Rickettsia or Chlamydia; Amphibian; Mammal; Fish}; {Group; Family Group; Age Group; Population Group; Professional or Occupational Group; Patient or Disabled Group}	Split Transformation	Split into two smaller connected groups.	Two groups: *Organism*; *Group*
Phenomena	**Laboratory or Test Result**	IS-A addition Transformation	Add (**Laboratory or Test Result** IS-A **Phenomenon or Process**) link	*Phenomenon or Process*
Physiology	{Organism Attribute; Clinical Attribute}	IS-A addition Transformation	Add (**Organism Attribute** IS-A **Physiologic Function**) link	*Physiologic Function*

braces "{}". In the fourth column, the notation (**A** IS-A **B**) is used to denote a single IS-A link that was added to the group, where **A** and **B** are semantic types. The new groups are named after their respective roots.

Overall, using the four kinds of transformations, all disconnected groups are converted into new connected groups, each with an internal tree structure. During this process, a total of ten transformations were applied: the IS-A addition transformation was used four times; the split transformation was used three times; the root-addition transformation was used once (on the *Anatomy* group); and the root-moving transformation was used twice. Note that multiple transformations might have been applied to a single group (see the *Disorders* group). The application of the four transformations yielded the preliminary ESN with 15 new IS-A links. Its DAG structure allows semantic types to have multiple parents.

A total of 19 disjoint groups, which together constitute a partition of the ESN, was also obtained. See Table 2.2, where "*" is used to denote a group different from that originally appearing in [39]. Each group is a connected subgraph of the ESN. Hence, the partition satisfies the connectivity property preferred for semantic validity. The groups of the resulting partition and semantic types for each group is listed in Table 2.3.

2.3.2 Results of String Matching

For the manual review, 550 (3%) of the 17,396 possible pairs of semantic types for which no ancestor/descendant relationship currently exists were randomly selected. Neither of the two reviewers judged any of the 550 pairs to represent a true parent-child relationship. This corresponds to a prevalence of unknown pairs of 0%, with a 95% confidence interval of 0-0.54%.

Table 2.2 Partition of the ESN into 19 Connected Groups

Group	# of STs	Group	# of STs
Anatomical Abnormality *	3	*Anatomical Entity**	15
*Chemical**	25	*Clinical Drug**	1
Conceptual Entity	12	*Entity**	4
*Event**	7	*Finding**	2
Geographic Area	1	*Group**	6
*Manufactured Object**	4	*Molecular Sequence*	5
Occupation or Discipline	2	*Occupational Activity**	9
*Organism**	17	*Organization*	4
*Pathologic Function**	7	*Phenomenon or Process*	6
Physiologic Function	9		

A total of 665 string matches were found by the algorithm. Only 5 of these were judged to represent true parent-child relationships, for a precision of 0.75%. However, these 5 positive results suggest a prevalence of 0.029% $(5/17,396)$, which is within the 95% confidence interval of the gold standard analysis.

The semantic method resulted in the addition of 15 new IS-A links. However, 11 of these links involved the addition of new semantic types, leaving 4 previously undiscovered IS-A links. One of these, **Gene or Genome** IS-A **Molecular Sequence** was also detected by the string matching method. Thus, a total of 8 new parent-child relationships were discovered (prevalence $8/17,396 \simeq 0.046\%$, still within the range found by the gold standard). The string matching method detected 5 of the 8 true parent-child relationships discovered by both methods, yielding a sensitivity (or recall) of 62.5%. At the maximum prevalence suggested by the 95% confidence interval (0.54%), the sensitivity could be as low as 5.3%.

The four additional IS-A links are presented as follows. One is the new IS-A

Table 2.3 Partition of the ESN

Group	Semantic Types in Group
*Event**	Event; Activity; Behavior; Individual Behavior; Social Behavior; Daily or Recreational Activity; Machine Activity
*Anatomical Entity**	Anatomical Entity; Physical Anatomical Entity; Conceptual Anatomical Entity; Material Physical Anatomical Entity; Anatomical Structure; Embryonic Structure; Fully Formed Anatomical Structure; Body Part, Organ, or Organ Component; Tissue; Cell; Cell Component; Body Substance; Body System; Body Location or Region; Body Space or Junction;
*Chemical**	Chemical; Chemical Viewed Functionally; Biologically Active Substance; Receptor; Vitamin; Enzyme; Hormone; Neuroreactive Substance or Biogenic Amine; Immunologic Factor; Hazardous or Poisonous Substance; Pharmacologic Substance; Antibiotic; Biomedical or Dental Material; Indicator, Reagent, or Diagnostic Aid; Chemical Viewed Structurally; Organic Chemical; Amino Acid, Peptide, or Protein; Organophosphorus Compound; Nucleic Acid, Nucleoside, or Nucleotide; Carbohydrate; Lipid; Steroid; Eicosanoid; Inorganic Chemical; Element, Ion, or Isotope
Clinical Drug*	Clinical Drug
Conceptual Entity	Conceptual Entity; Group Attribute; Language; Idea or Concept; Functional Concept; Temporal Concept; Qualitative Concept; Quantitative Concept; Intellectual Product; Classification; Regulation or Law; Spatial Concept
*Manufactured Object**	Manufactured Object, Research Device; Medical Device; Medical Delivery Device
*Pathologic Function**	Pathologic Function; Experimental Model of Disease; Disease or Syndrome; Injury or Poisoning; Neoplastic Process; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction
*Anatomical Abnormality **	Anatomical Abnormality; Acquired Abnormality; Congenital Abnormality
*Finding**	Finding; Sign or Symptom;
Molecular Sequence	Molecular Sequence; Amino Acid Sequence; Carbohydrate Sequence; Nucleotide Sequence; Gene or Genome
Geographic Area	Geographic Area
*Organism**	Organism; Fungus; Alga; Virus; Human; Plant; Archaeon; Reptile; Rickettsia or Chlamydia; Amphibian; Mammal; Fish
*Group**	Group; Family Group; Age Group; Population Group; Professional or Occupational Group; Patient or Disabled Group
*Entity**	Entity; Physical Object; Substance; Food
Occupation or Discipline	Occupation or Discipline; Biomedical Occupation or Discipline
Organization	Organization; Professional Society; Health Care Related Organization; Self-help or Relief Organization
Phenomenon or Process	Phenomenon or Process; Human-caused Phenomenon or Process; Environmental Effect of Humans; Laboratory or Test Result; Natural Phenomenon or Process; Biologic Function
Physiologic Function	Physiologic Function; Organism Attribute; Clinical Attribute; Organ or Tissue Function; Organism Function; Mental Process; Molecular Function; Genetic Function; Cell Function
*Occupational Activity**	Occupational Activity; Health Care Activity; Laboratory Procedure; Diagnostic Procedure; Therapeutic or Preventive Procedure; Governmental or Regulatory Activity; Educational Activity; Research Activity; Molecular Biology Research Technique

link from **Enzyme** to **Amino Acid, Peptide, or Protein**, which was demonstrated in Section 2.2.2.

Another example relates to **Receptor**, for which there were five string matches:

(**Receptor**; **Cell Component**; "cell")

(**Receptor**; **Cell**; "cell")

(**Receptor**; **Anatomical Structure**; "structure")

(**Receptor**; **Amino Acid, Peptide, or Protein**; "protein")

(**Receptor**; **Hormone**, "hormone")

In accordance with the review of the domain expert, an IS-A link from **Receptor** to **Cell Component** was added. The other string matches did not imply IS-A links.

The third valid IS-A link involves **Vitamin**, which had four matches:

(**Vitamin**; **Pharmacologic Substance**; "substance")

(**Vitamin**; **Organic Chemical**; "organic chemical")

(**Vitamin**; **Body Substance**; "substance")

(**Vitamin**; **Animal**; "animal")

Based on the domain expert's review, two IS-A links were added: one IS-A from **Vitamin** to **Pharmacologic Substance**, and another IS-A from **Vitamin** to **Organic Chemical**.

2.3.3 Summary of Results of Two Methodologies

By adding the new IS-A links derived by the above two methodologies, a new network, referred to as the Enriched Semantic Network (ESN), was obtained. Compared to the original SN, the ESN has four new semantic types and 19 new IS-A links. Two IS-A links

appearing in the SN were not included in the ESN. Hence, the ESN has 150 IS-A links and 139 semantic types, among which twelve semantic types (about 8%) have multiple parents, giving the ESN a DAG-structured IS-A (subsumption) hierarchy. See Table 2.4 for these twelve semantic types and their parents.

Table 2.4 Semantic Types (STs) with Multiple Parents in the ESN

Child ST	Old Parent ST	New Parent ST	New Parent ST
Body Location or Region	**Spatial Concept**	**Conceptual Anatomical Entity**	—
Body Space or Junction	**Conceptual Entity**	**Physical Anatomical Entity**	—
Body Substance	**Substance**	**Material Physical Anatomical Entity**	—
Body System	**Functional Concept**	**Conceptual Anatomical Entity**	—
Conceptual Anatomical Entity	—	**Conceptual Entity**	**Anatomical Entity**
Enzyme	**Biologically Active Substance**	**Amino Acid, Peptide, or Protein**	—
Gene or Genome	**Fully Formed Anatomical Structure**	**Molecular Sequence**	—
Laboratory or Test Result	**Finding**	**Phenomenon or Process**	—
Organism Attribute	**Conceptual Entity**	**Physiologic Function**	—
Physical Anatomical Entity	—	**Physical Object**	**Anatomical Entity**
Receptor	**Biologically Active Substance**	**Cell Component**	—
Vitamin	**Biologically Active Substance**	**Organic Chemical**	**Pharmacologic Substance**

Figure 2.8 shows the portion of the ESN's hierarchy rooted at **Event**, and Figure 2.9 shows part of the portion rooted at **Entity**. To emphasize the changes from the original SN, dashed arrows are used to denote the new IS-A links and thick dashed rectangles to denote

new semantic types. Thin dashed rectangles denote semantic types that originally resided in the other tree of the SN. Ellipses in a rectangle indicate that the names of one or several semantic types are not shown due to lack of space.

2.4 Discussion

The ESN obtained in Section 2.3 has a DAG-structured IS-A hierarchy with more semantic types and more IS-A links. In this section, The advantages of the ESN is presented for semantic relationship modeling of the SN and the META's concept classification. The limitations of the two methodologies and a brief evaluation are also included in this section.

2.4.1 Advantages of the ESN

The ESN has twelve semantic types with multiple parents. As it happens, most such semantic types are leaves or parents of leaves. As such the changes are local, not influencing other semantic types. An exception is the modeling of the four new semantic types, **Anatomical Entity**, its two children **Conceptual Anatomical Entity** and **Physical Anatomical Entity**, and the child of the latter, **Material Physical Anatomical Entity**. This is the most visible difference from the original SN's two-tree structure, since it happens close to the root **Entity** rather than at the bottom levels of the SN. As such, it is not a local change.

The ESN has a number of advantages over the original SN. The multiple subsumption hierarchy enables better modeling of IS-A relationships for those semantic types having multiple parents. In the ESN, some semantic types will have more semantic relationships than they had in the SN. Specifically, semantic types with multiple parents will inherit relationships independently from each of those parents. Thus, such a semantic type

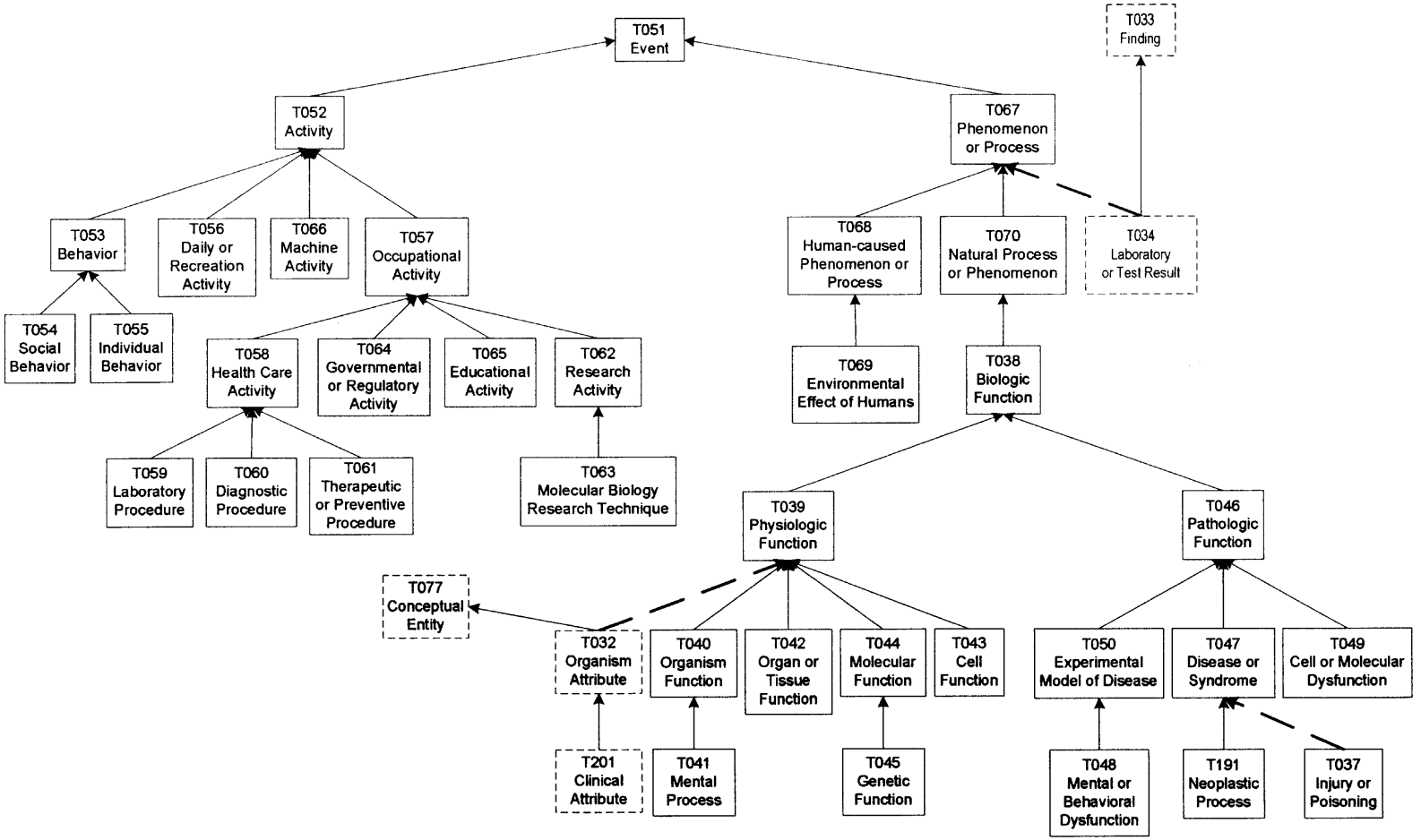


Figure 2.8 Event portion of the ESN.

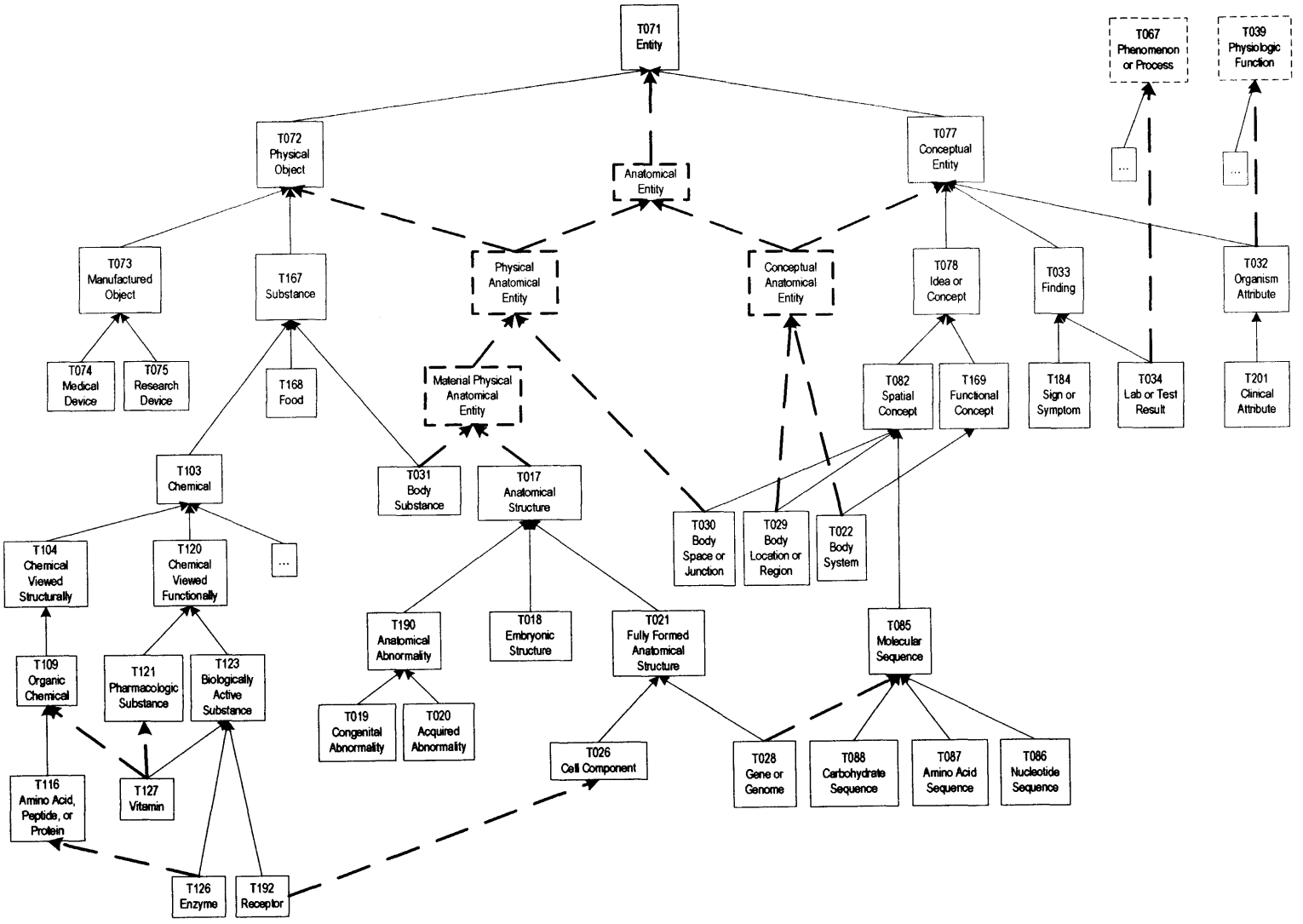


Figure 2.9 Part of the **Entity** portion of the ESN.

will have a larger relationship set than before. For example, **Organism Attribute** and its child **Clinical Attribute** will now have a relationship named *result_of* to **Anatomical Abnormality**. This relationship is inherited by **Organism Attribute** from its new parent **Physiologic Function**, and further inherited by its child **Clinical Attribute**.

One might consider the introduction of multiple inheritance as a potential problem in that inconsistent information from different parents might be inherited. However, when the placement of a concept into two classes is semantically correct, then the inheritance of definitional attributes from multiple parents is, by definition, also correct. Multiple inheritance will allow the identification of inconsistencies that were already there implicitly; it will not introduce new ones.

The addition of the IS-A links helps to expose missing classifications of concepts of the META to semantic types. The following example demonstrates this with regards to the concepts assigned to the semantic type **Vitamin**.

All 1,204 concepts from the META assigned to **Vitamin** are checked and found that 957 are also assigned to **Pharmacologic Substance**. One of them is also assigned to **Antibiotic**, which is a child of **Pharmacologic Substance**. The other 246 concepts are not assigned to **Pharmacologic Substance**. For example, the concept FOLATE[3] is not assigned to **Pharmacologic Substance**. However, some drugs, for example, vitamins given to pregnant women, contain folate to prevent possible congenital deficiencies of the baby. Hence, FOLATE should indeed be assigned to **Pharmacologic Substance**. As a matter of fact, all the remaining 246 concepts should also have been assigned to **Pharmacologic Substance** because all vitamins can be ingredients of drugs.

[3]Concept names will be written in a "small caps" style in this chapter.

Similarly, all concepts in **Vitamin** should also be assigned to **Organic Chemical** or one of its descendants. Among the 1,204 concepts assigned to **Vitamin**, 735 are also assigned to **Organic Chemical** or children of **Organic Chemical**. But there are 469 concepts assigned to **Vitamin** that are neither assigned to **Organic Chemical** nor to any of its children. An example is 24,25-DIHYDROXYVITAMIN D, which is a kind of vitamin D that is helpful for the absorption of Calcium and is certainly an organic chemical. In fact, all these 469 concepts should have been assigned to **Organic Chemical**.

Another advantage of the multiple-parent hierarchy is that it can simplify the assignment of META's concepts to semantic types. An important rule promoted by the SN's designers states that a concept should be explicitly assigned to the most specialized possible semantic type in the SN's IS-A hierarchy [41]. Suppose a concept was assigned to two semantic types $\mathbf{T}_1$ and $\mathbf{T}_2$ that originally had no IS-A path between them in the SN. If in the ESN, there is a direct IS-A link or path from $\mathbf{T}_1$ to $\mathbf{T}_2$ (i.e., $\mathbf{T}_2$ is now a parent or ancestor of $\mathbf{T}_1$), then the assignment of the concept to $\mathbf{T}_2$ is considered redundant [19, 48] and should be removed because it can be inferred from the assignment to $\mathbf{T}_1$.

As an example, consider the new IS-A link from **Vitamin** to **Pharmacologic Substance**. After adding this IS-A link, the 957 assignments of **Vitamin**'s concepts to **Pharmacologic Substance** should be removed since **Vitamin** is now more specialized than **Pharmacologic Substance** in the network. The 957 concepts should only be assigned to **Vitamin**, because implicit assignments to **Pharmacologic Substance** can be inferred via the new IS-A. After the addition of **Vitamin** IS-A **Organic Chemical**, the 735 assignments to **Organic Chemical** should also be removed for the same reason.

In the preceding discussion, the 469 additional concepts are proposed to assigned to **Organic Chemical** and then are removed subsequently. However, the proposed additions were strictly in the context of the current SN hierarchy, where simultaneous assignment to **Vitamin** and **Organic Chemical** is not redundant and is in fact warranted. In the ESN, with **Vitamin** now being a child of **Organic Chemical**, such assignments become redundant and therefore, unnecessary. This further supports the validity of the new IS-A link and demonstrates that the ESN hierarchy requires fewer explicit assignments of META's concepts to the semantic types.

The partition of the ESN can enable the design of a metaschema [49], a higher-level abstraction network that can aid in user orientation. Among other things, a metaschema will allow a user to focus on a subject area of interest, without losing sight of the overall ESN layout.

Regarding limitations, the first methodology was applied only to the partition presented in [39], and decisions were made with respect to the current definitions of semantic types. Of course, there are many possible partitions of the SN. If other partitions were used as references, different IS-A links might be identified.

The string matching methodology is dependent on the definitions of the semantic types. It is important to realize that the current definitions are not necessarily the only ones possible for the given semantic types. Another team of designers might come up with slightly different definitions. Because the exact wording of the definitions is utilized, the results may very well be altered by alternative definitions. Furthermore, the average time complexity of the algorithm is about $O(n^2)$, and this limits its scalability. It is thus

applicable only to a compact upper-level abstraction ontology (like the SN), not a full-scale ontology. For example, the algorithm will be very time-consuming if it is applied to finding string matches for the META's concepts. Word-level synonymy (or phrase synonymy) was not considered in the algorithm. If utilized, it may increase the string match cases and maybe the number of new viable IS-A links found. However, this would likely erode the algorithm's efficiency, which is already low, and might increase the number of false positives, which is already high.

2.4.2 Evaluation

Without an exhaustive examination of all 17,396 pairs of unrelated semantic types, it is impossible to know the exact prevalence of undiscovered parent-child relationships. However, all of the methods used (semantic modeling, manual review, and string matching) suggest that the number of such relationships is very low. In the absence of a precise figure for prevalence, estimating the sensitivity of the automated methods is impossible. However, the semantic modeling revealed 15 links and the string matching revealed 5 links; either of these counts represent a significant contribution to the number of links in the SN; taken together as 19 (since one was repeated), they increase the number of links by 14.3%.

While at first glance the precision of the string matching method (0.75%) appears poor, applying it to the SN reduces the number of semantic-type pairs that must be manually reviewed by 94%. It is important to note that there is no inherent reason why a string match from the definition of a semantic type to the name of another semantic type would necessarily indicate an IS-A link. The match may indicate another kind of connection, such as a semantic relationship. The hypothesis was that if there is a shared string, then

the likelihood of a parent-child relationship is substantially higher. The results support the hypothesis.

2.5 Summary

In this chapter, the UMLS's Semantic Network (SN) hierarchy was enhanced by adding new IS-A links and new semantic types to accommodate multiple parents. A new semantic network that has a DAG structure instead of a two-tree structure was obtained. This new semantic network, containing 139 semantic types and 150 IS-A relationships, is referred to as the Enriched Semantic Network (ESN). The ESN expresses cases of multiple subsumption for several semantic types. Furthermore, a partition of the ESN comprising 19 groups was derived; each group in the partition exhibits connectivity and semantic uniformity. This new partition enables the design of a metaschema [67] which helps to further improve user orientation to the ESN.

CHAPTER 3

ESN'S RELATIONSHIPS DISTRIBUTION AND CONCEPT CONFIGURATION

3.1 Introduction

In Chapter 2, the creation of the Enriched Semantic Network (ESN) as an extension of the SN was described. Its key characteristic is an IS-A hierarchy permitting multiple parents for a single semantic type. The ESN thus exhibits a directed acyclic graph (DAG) hierarchy, in contrast to the SN's tree-structured hierarchy. The ESN also contains some additional semantic types that were included to support the new multiple subsumption framework. Overall, the ESN contains 139 semantic types and 150 IS-A links.

As in the SN, semantic types of the ESN are also connected by semantic (non-IS-A) relationships of 53 different kinds. Such relationships can be directly introduced at a semantic type or inherited by it. When a relationship is defined at a semantic type but not at its parent, that semantic type is called an *introduction point* of the relationship. All the descendants of an introduction point inherit this introduced relationship, unless the inheritance is explicitly blocked. There are two mechanisms for blocking inheritance in the SN. The first mechanism, called "blocking," nullifies the definition of an inherited relationship. The second mechanism allows a newly introduced relationship to be designated as "defined but not inherited (DNI)." This means that the relationship is not inherited by any of the children (and thus descendants) of the semantic type that is introducing it.

The entire set of relationships exhibited by a semantic type—including those inherited and those introduced—is called the "relationship structure" of the semantic type. Collectively, the collection of relationship structures of all semantic types is referred to as

the relationship distribution of the SN. The relationship distribution plays a major role in the analysis of a partition of the SN [2]. The relationship structure of a given semantic type in the ESN may in general differ from that of the same type in the SN. This is a result of the fact that in the ESN a semantic type can have more than one parent and inherit relationships independently from each—a situation referred to as "multiple inheritance." The ESN was designed so that all semantic types should at least preserve their relationship structures. That is, the relationship structure of a semantic type in the ESN will be a (not necessarily proper) superset of that in the SN. It will be noted though that introduction points for relationships may have changed.

In Chapter 2, only the IS-A hierarchy of the ESN was presented without presenting the details of the relationship structures. In this chapter, a technique to derive the ESN's entire semantic relationship distribution is presented and its application is analyzed, with particular emphasis placed on those semantic types having more than one parent. The introduction points for all relationships and the relationship structures of all types are examined in the context of the new multiple subsumption network. All newly inherited relationships are audited for semantic validity, and those deemed invalid are excluded from the ESN.

As with the SN, the ESN is designed to serve as a high-level abstraction of the underlying META, with each concept being assigned to one or more semantic types. Collectively, such assignments of concepts to semantic types is referred to as the "concept configuration." Rather than redoing all the work of the UMLS's maintenance personnel, the ESN's concept configuration is derived automatically from that of the SN. In this chapter, a mapping function is defined through which this derivation takes place. An important

issue in the development of this mapping is compliance with the principle that each concept be explicitly assigned to the lowest (or most specialized) semantic type in the IS-A hierarchy [41]. In previous work [20, 48], many situations were found where a concept was assigned both to a descendant semantic type and its ancestor type simultaneously. Such a situation, which is refer to as a "redundant categorization," must be avoided in the ESN's concept configuration. The mapping function ensures that the ESN is free of any redundant categorizations.

3.2 Derivation of Relationship Distribution

In the SN, there are 53 different kinds of semantic relationships. However, there are typically many different occurrences for each kind of relationship. For example, there is an *affects* relationship from **Anatomical Abnormality** to **Alga**; meanwhile, there is also an *affects* relationship from **Amino Acid, Peptide, or Protein** to **Biologic Function**. Each of them is an occurrence of *affects*, with different source and target semantic types.

The notation $r(\mathbf{X}, \mathbf{Y})$ is used to denote an occurrence of the relationship of kind r from semantic type **X** to semantic type **Y**. Here, r is the kind of relationship; **X** and **Y** are the source semantic type and the target semantic type of the relationship, respectively.

In the original SN, there are 6,977 semantic relationship occurrences of the 53 different kinds. Hence, the average number of occurrences per semantic type is about 50. A semantic type may be the source of several occurrences of the same kind of relationship, with different targets. For brevity, "occurrence" and "relationship" will be used interchangeably whenever there is no possibility of confusion.

Relationships fit into two categories:

- Introduced relationship
- Inherited relationship

Let **X** and **Y** be two semantic types, and let $\mathbf{P}_X$ be the parent of **X**. A relationship r(**X**, **Y**) is an *introduced relationship* of **X** if there does not exist a relationship $r(\mathbf{P}_X, \mathbf{Y})$ in the SN; otherwise, it is an *inherited relationship* of **X** unless $r(\mathbf{P}_X, \mathbf{Y})$ is a DNI relationship at $\mathbf{P}_X$ or a blocked relationship at **X**.

There are 422 introduced relationships in the SN and in total $6,977 - 422 = 6,555$ inherited relationships. There are only 27 DNI relationships (about 6% of the introduced relationships) and ten "blocking" relationships.

The ESN's relationship distribution is derived from that of the SN according to the following three rules and review step.

Rule 1: An introduced relationship r(**X**, **Y**) in the SN implies an introduced relationship r(**X**, **Y**) in the ESN;

Rule 2: An inherited relationship r(**X**, **Y**) in the SN implies an inherited relationship r(**X**, **Y**) in the ESN;

Rule 3: If a semantic type **T** has multiple parents (or ancestors) in the ESN, then initially **T** inherits all the relationships of its new parents (or ancestors) except for those that have been explicitly blocked or are DNI relationships.

Review Step: A domain expert manually checks the semantic validity of all newly inherited relationships in the ESN. Only those relationships that are deemed semantically

valid are retained; otherwise, blocking or DNI is used to avoid inheritance of an invalid relationship.

All existing introduced relationships in the SN are preserved in the ESN according to Rule 1; and all existing inherited relationships in the SN are also preserved in the ESN in Rule 2. For each semantic type having multiple parents, Rule 3 will find all newly inherited relationships that can be inherited from the new parent(s).

As an example of Rule 3, **Gene or Genome** has a new parent **Molecular Sequence** in the ESN. According to Rule 3, it will inherit all non-blocked and non-DNI relationships from the new parent. There is a relationship *result_of*(**Molecular Sequence, Mental Process**) that is not defined at either **Gene or Genome** or its unique parent **Fully Formed Anatomical Structure** in the SN. Therefore, according to Rule 3, **Gene or Genome** will initially inherit the *result_of* relationship in the ESN. That means there is a relationship *result_of*(**Gene or Genome, Mental Process**) in the ESN waiting to be reviewed by a domain expert in the Review Step. For this relationship, it is deemed valid according to the expert's review. Therefore, **Gene or Genome** will truly have a relationship *result_of*(**Gene or Genome, Mental Process**) in the ESN.

Rule 3 implies that a semantic type with multiple parents might have more relationships in the ESN than in the SN, because it could inherit new relationships from its new parents. The same is true for its descendants.

3.3 Concept Configuration Mapping

To complete the abstraction provided by the ESN, all of META's concepts must be assigned to one or more of the ESN's semantic types. As noted, the UMLS documentation actually views these assignments in the opposite direction, with semantic types being assigned to concepts. Moreover, the entire set of assignments, which is referred to as the *concept configuration*, is considered a part of META and distributed in the file MRSTY. Because the ESN was created as a new abstraction mechanism on top of META, it is better to keep the ESN's concept configuration separate from META itself. In this regard, the author talks of the assignment of concepts to types. This arrangement also avoids any upheaval in META's current representation. Of course, a new file, say, MRSTYE could be constructed to serve in the same role for the ESN that MRSTY does for the SN.

The simplest way to construct the ESN's concept configuration is to use that of the SN without any change. That is, if a concept C in META was assigned to a set of semantic types $\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_m\}$ in the SN, then in the ESN, the concept C will also be assigned to these same types. This mapping, although direct and simple, will possibly yield two kinds of redundant categorizations in the ESN. In the first case, an already existing redundant categorization is copied over to the ESN. In the second case, a new redundant categorization arises as a result of a semantic type having more parents than it did before. The mapping function deals with these situations in order to prevent introducing any redundant categorizations in the ESN. For the latter case, it is necessary to check each pair of semantic types having a new IS-A path between them in the ESN that did not exist in the SN. If such a new IS-A path has the potential for introducing new redundant

categorizations, then this must be accounted for in the mapping.

For example, besides the current parent **Conceptual Entity**, **Organism Attribute** has a new parent **Physiologic Function** in the ESN (see Figure 3.1). Among the 2,381 concepts assigned to **Organism Attribute**, 14 concepts, e.g., INTRACHAMBER DIASTOLIC,[1] are also assigned simultaneously to **Physiologic Function** (see Table 3.1 for the whole list). Since **Physiologic Function** is now a parent of **Organism Attribute** in the ESN, the 14 assignments to **Physiologic Function** would be redundant categorizations if they were to appear in the ESN's concept configuration. Therefore, The mapping must prevent introducing these 14 assignments to **Physiologic Function**.

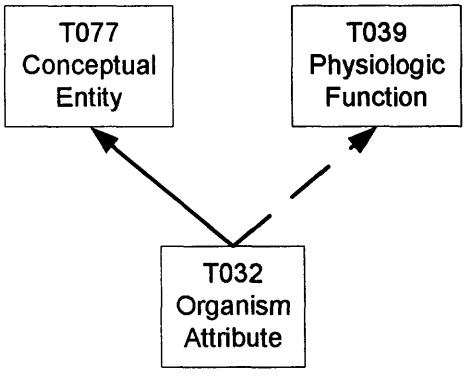


Figure 3.1 Organism Attribute with its parents in the ESN.

With the above considerations in mind, the mapping function can be defined as follows. Suppose concept C was assigned to the semantic types $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_m$ in the SN. The assignments for concept C in the ESN's concept configuration can be obtained according to the following three rules.

Rule 1: If in the ESN no pair of types $(\mathbf{T}_i, \mathbf{T}_j)$ $(i \neq j)$ has an IS-A path between them, then C is assigned in the ESN to each of the types $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_m$. □

[1]Concept names will be written in a "small caps" style in this chapter.

Table 3.1 Concepts Assigned in the SN to **Organism Attribute** and **Physiologic Function**

Concept ID	Concept Name
C0489560	INTRACHAMBER DIASTOLIC
C0489561	INTRACHAMBER MEAN
C0489562	INTRACHAMBER SYSTOLIC
C0489564	INTRAVASCULAR END DIASTOLIC
C0489565	INTRAVASCULAR MEAN
C0489566	INTRAVASCULAR SYSTOLIC
C0489568	INTRAVASCULAR SYSTOLIC.INSPIRATION - EXPIRATION
C0489703	FOREARM BLOOD PRESSURE SYSTOLIC
C0489708	HEPATIC CAPILLARY WEDGE PRESSURE
C0489726	LEFT UPPER ARM BLOOD PRESSURE MEAN
C0489728	MAXIMUM SYSTOLIC BLOOD PRESSURE
C0489731	MEAN SYSTOLIC BLOOD PRESSURE
C0489733	MINIMUM SYSTOLIC BLOOD PRESSURE
C0489763	RIGHT THIGH BLOOD PRESSURE SYSTOLIC

Rule 2: If among the m semantic types, $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_m$, there exists a pair $(\mathbf{T}_i, \mathbf{T}_j)$ $(i \neq j)$ such that $\mathbf{T}_i$ is an ancestor of $\mathbf{T}_j$ in the SN, then the assignment of C to $\mathbf{T}_i$ will be excluded from the ESN. □

Rule 3: If among the m semantic types, there exists a pair $(\mathbf{T}_i, \mathbf{T}_j)$ $(i \neq j)$ such that $\mathbf{T}_i$ is a new ancestor of $\mathbf{T}_j$ in the ESN, then the assignment of C to $\mathbf{T}_i$ will be excluded from the ESN. □

Rule 1 is used to preserve all non-redundant categorizations in the SN. Rule 2 excludes all redundant categorizations currently existing in the SN's concept configuration from the ESN's concept configuration. Rule 3 averts the introduction of new redundant categorizations arising from multiple-parent cases in the ESN. Overall, the application of the three rules yields an ESN concept configuration that preserves the SN's concept

configuration while purging existing redundant categorizations and avoiding new ones.

Note that in the SN no concept is assigned to the four new semantic types of the ESN. However, each of the four should be assigned at least the corresponding concept, of the same name, suggested in [43, 54]. (These concepts have been submitted to the NLM for inclusion in the next UMLS release.[2]) Furthermore, a domain expert should review the concepts assigned to the parents and children of each of the four new semantic types to check whether any assignments of their concepts should be switched to one of these four.

The application of the three mapping rules involves the use of algorithms that detect all existing or potential redundant categorizations. For Rule 2, the algorithm in [48] (here referred to as "DetectRedundantCatgs") is used to scan through the SN's concept configuration (as supplied in MRSTY) and mark all assignments it determines to be redundant categorizations. Subsequently, these marked assignments are not introduced into the ESN's concept configuration.

For Rule 3, the following DetectNewRedundantCatgs algorithm is applied to detect and mark all potentially new redundant categorizations arising from new IS-A paths in the ESN. In the algorithm, E_T denotes the set of concepts assigned to a semantic type **T** in the SN. NewAncestors(**T**) is the set of all new ancestor semantic type(s) (including the new parent(s)) of **T** in the ESN. $E_{T_1} \cap E_{T_2}$ is the intersection of the concept sets of $\mathbf{T_1}$ and $\mathbf{T_2}$. Following the UMLS convention, the notation (C | **T**) is used to denote the assignment of concept C to the semantic type **T**.

DetectNewRedundantCatgs algorithm: mark potentially new redundant categorizations.

[2]C. Rosse, personal communication, 2002.

```
for (each semantic type T with a new parent(s)
     or a new ancestor(s))
{
    for (each semantic type Y ∈ NewAncestors(T))
    {
        if (E_T ∩ E_Y ≠ ∅ in the SN)
            //potential redundant categorizations
            for (each concept C ∈ E_T ∩ E_Y)
                Mark the assignment (C | Y)
    }
}
```

With the algorithms to detect and mark the existing and potentially new redundant categorizations, the GenerateESNConceptConfig algorithm can be defined to implement the mapping function. It creates the ESN concept configuration free of any redundant categorizations. In the following, assume that the SN's concept configuration is available as a set:

$$F = \{(\mathrm{C} \mid \mathbf{T}) \mid \text{ C is a concept, } \mathbf{T} \text{ is a semantic type}\}$$

GenerateESNConceptConfig algorithm: assign META's concepts to ESN's semantic types.

```
Call DetectRedundantCatgs();

Call DetectNewRedundantCatgs();

for (each assignment (C | T) ∈ F)
    if ((C | T) is not marked)
        Assign C to T in the ESN;
```

Note that the mapping function handles a redundant categorization in a way such that the assignment of the concept to the parent (or ancestor) will always be the one excluded from the ESN. But it is possible that in the original SN, the assignment of the concept to the parent (or ancestor) is actually correct while the assignment to the child is

wrong. In such a case, the assignment to the parent should be preserved in the ESN while the assignment to the child should be excluded. If such a case is found by a human expert and corrected in the original SN, this algorithms can be re-run after the correction to guarantee that the concept is assigned to the correct type in the ESN.

3.4 RESULTS

This section first presents the ESN's relationship distribution derived from that of the SN, taking into account the newly inherited relationships. Then the ESN's concept configuration is presented as the result of the mapping function defined in Section 3.3. The ESN's concept configuration is free from any redundant categorizations.

3.4.1 ESN Relationship Distribution

By applying the three rules and the Review Step in Section 3.2, the ESN's relationship distribution is obtained. Rule 1 obtained 422 introduced relationships, and Rule 2 yielded 6,555 inherited relationships. In Rule 3, 426 newly inherited relationships are obtained through multiple inheritance. In the Review Step, all 426 new relationships were audited by professor James J. Cimino, our medical expert.

Among the 426 new relationships, twelve involve the four semantic types appearing exclusively in the ESN and not in the SN. These were deemed valid upon review. For example, for the new semantic type **Anatomical Entity**, there is a relationship *issue_in*(**Anatomical Entity**, **Biomedical Occupation or Discipline**) that is inherited from its parent **Entity** in the ESN. Table 3.2 lists all twelve relationships.

Table 3.2 New Relationships for the Four New Semantic Types in the ESN

issue_in(Anatomical Entity, Biomedical Occupation or Discipline)
issue_in(Anatomical Entity, Occupation or Discipline)
part_of(Anatomical Entity, Organism)
issue_in(Physical Anatomical Entity, Biomedical Occupation or Discipline)
issue_in(Physical Anatomical Entity, Occupation or Discipline)
part_of(Physical Anatomical Entity, Organism)
issue_in(Conceptual Anatomical Entity, Biomedical Occupation or Discipline)
issue_in(Conceptual Anatomical Entity, Occupation or Discipline)
part_of(Conceptual Anatomical Entity, Organism)
issue_in(Material Physical Anatomical Entity, Biomedical Occupation or Discipline)
issue_in(Material Physical Anatomical Entity, Occupation or Discipline)
part_of(Material Physical Anatomical Entity, Organism)
Total: 12

The remaining 414 ($426 - 12 = 414$) newly inherited relationships involve currently existing semantic types having multiple parents (or ancestors) in the ESN because only these semantic types might inherit new relationships from their parents (or ancestors). Among all 135 semantic types in the SN, 21 semantic types have multiple parents (or ancestors) in the ESN. They are **Anatomical Structure** with its ten descendants, **Organism Attribute** with its child **Clinical Attribute**, and eight other leaf semantic types. Hence, at most 21 semantic types can exhibit different relationships structures in the ESN from those in the SN. The 414 newly inherited relationships involve these 21 semantic types.

A review of the 414 new relationships found that 314 out of the them (about 75.8%) are valid and are thus retained in the ESN. These are inherited by a total of 12 semantic types out of the 21 types having multiple parents. The other 100 relationships are semantically invalid and are blocked from being inherited by the children from their new parents. Therefore, those twelve semantic types have different relationship structures in

the ESN from those in the SN. For example, **Body substance** in the ESN has a different relationship structure from that in the SN since it inherits a valid *part_of* relationship to **Organism** from its new parent **Conceptual Anatomical Entity**. As an example of an invalid new relationship, **Organism Attribute**'s new parent **Physiologic Function** has a *process_of* relationship to **Organism** that might be inherited by **Organism Attribute** in the ESN. After being reviewed by a domain expert, *process_of*(**Organism Attribute, Organism**) is deemed invalid and is excluded ("blocked") in the ESN. Table 3.3 presents these twelve semantic types, number of newly inherited relationships reviewed, number of valid relationships in the ESN, and number of invalid relationships that are blocked in the ESN.

Now consider the semantic types for which blocking occurs. In the ESN, **Injury or Poisoning** has a new parent **Disease or Syndrome**. This new IS-A relationship causes 112 newly inherited relationships for **Injury or Poisoning**. After being reviewed, 92 are deemed valid and are retained, while 20 are invalid and excluded. For example, there is a new relationship *affects*(**Injury or Poisoning, Organism**) inherited from **Disease or Syndrome**. The review concluded that this relationship is valid and retained. Meanwhile, another new relationship *degree_of*(**Injury or Poisoning, Pathologic Function**) was found invalid and is excluded. Table 3.4 shows the 20 invalid relationships. All the 92 valid relationships are not listed because of space limitations.

Another example involves **Laboratory or Test Result** which has the new parent **Phenomenon or Process**. This new IS-A relationship causes 22 new relationships for **Laboratory or Test Result** that might be inherited from **Phenomenon or Process**. In the

Table 3.3 Relationships Inherited from New Parent Semantic Types in ESN

Child Semantic Type	New Parent Semantic Type	# new relationships reviewed	valid	invalid
Body Location or Region	Conceptual Anatomical Entity	1	1	0
Body Space or Junction	Physical Anatomical Entity	1	1	0
Body Substance	Material Physical Anatomical Entity	1	1	0
Body System	Conceptual Anatomical Entity	1	1	0
Clinical Attribute	Physiologic Function	92	52	40
Enzyme	Amino Acid, Peptide, or Protein	1	1	0
Gene or Genome	Molecular Sequence	1	1	0
Injury or Poisoning	Disease or Syndrome	112	92	20
Laboratory or Test Result	Phenomenon or Process	22	22	0
Organism Attribute	Physiologic Function	92	52	40
Receptor	Cell Component	67	67	0
Vitamin	Pharmacologic Substance	23	23	0
Total: 12		414	314	100

Review Step, all of them were deemed valid and were retained in the ESN. For example, the new relationship *result_of*(**Laboratory or Test Result, Acquired Abnormality**) is deemed valid on review. Table 3.5 shows all the 22 new valid relationships.

There are in total 7,297 relationships (including both introduced and inherited relationships) in the ESN vs. 6,977 in the original SN. Among the 139 semantic types in the ESN, 122 have the same relationship structures as in the SN, and 16 have different relationship structures. Among them, four are new semantic types, twelve are semantic types having newly inherited relationships. Table 3.6 shows these 16 semantic types and their numbers of relationships in the SN and ESN.

Table 3.4 Invalid Semantic Relationships of **Injury or Poisoning** Blocked in the ESN

degree_of(Injury or Poisoning, Pathologic Function)
degree_of(Injury or Poisoning, Cell or Molecular Dysfunction)
degree_of(Injury or Poisoning, Disease or Syndrome)
degree_of(Injury or Poisoning, Experimental Model of Disease)
degree_of(Injury or Poisoning, Mental or Behavioral Dysfunction)
degree_of(Injury or Poisoning, Neoplastic Process)
manifestation_of(Injury or Poisoning, Pathologic Function)
manifestation_of(Injury or Poisoning, Physiologic Function)
manifestation_of(Injury or Poisoning, Cell Function)
manifestation_of(Injury or Poisoning, Cell or Molecular Dysfunction)
manifestation_of(Injury or Poisoning, Disease or Syndrome)
manifestation_of(Injury or Poisoning, Experimental Model of Disease)
manifestation_of(Injury or Poisoning, Genetic Function)
manifestation_of(Injury or Poisoning, Injury or Poisoning)
manifestation_of(Injury or Poisoning, Mental Process)
manifestation_of(Injury or Poisoning, Mental or Behavioral Dysfunction)
manifestation_of(Injury or Poisoning, Molecular Function)
manifestation_of(Injury or Poisoning, Neoplastic Process)
manifestation_of(Injury or Poisoning, Organ or Tissue Function)
manifestation_of(Injury or Poisoning, Organism Function)

As an example of a new semantic type, **Anatomical Entity** has three relationships in the ESN: one is the introduced *part_of* relationship; the other two are occurrences of *issue_in* (with different targets) inherited from the parent **Entity**. The four new semantic types obviously did not have any relationships in the SN. As an example for semantic types having newly inherited relationships, **Vitamin** has 86 semantic relationships in the SN as opposed to 109 semantic relationships in the ESN.

Anatomical Structure is a special case for relationship structure. Although it has the same relationship structure in the ESN and in the SN, its relationship introduction pattern is different. In the SN, it is the introduction point of the relationship *part_of*(**Anatomical**

Table 3.5 Laboratory or Test Result's New Relationships Inherited from **Phenomenon or Process**

result_of(Laboratory or Test Result, Acquired Abnormality)
result_of(Laboratory or Test Result, Anatomical Abnormality)
result_of(Laboratory or Test Result, Biologic Function)
result_of(Laboratory or Test Result, Cell Function)
result_of(Laboratory or Test Result, Cell or Molecular Dysfunction)
result_of(Laboratory or Test Result, Congenital Abnormality)
result_of(Laboratory or Test Result, Disease or Syndrome)
result_of(Laboratory or Test Result, Environmental Effect of Humans)
result_of(Laboratory or Test Result, Experimental Model of Disease)
result_of(Laboratory or Test Result, Genetic Function)
result_of(Laboratory or Test Result, Human-caused Phenomenon or Process)
result_of(Laboratory or Test Result, Injury or Poisoning)
result_of(Laboratory or Test Result, Mental Process)
result_of(Laboratory or Test Result, Mental or Behavioral Dysfunction)
result_of(Laboratory or Test Result, Molecular Function)
result_of(Laboratory or Test Result, Natural Phenomenon or Process)
result_of(Laboratory or Test Result, Neoplastic Process)
result_of(Laboratory or Test Result, Organ or Tissue Function)
result_of(Laboratory or Test Result, Organism Function)
result_of(Laboratory or Test Result, Pathologic Function)
result_of(Laboratory or Test Result, Phenomenon or Process)
result_of(Laboratory or Test Result, Physiologic Function)

Structure, **Organism**), but in the ESN it inherits this relationship from its new parent **Physical Anatomical Entity** instead of itself introducing this relationship. Therefore, *part_of*(**Anatomical Structure**, **Organism**) is an introduced relationship in the SN, while it is an inherited relationship in the ESN. Since the relationship structure of **Anatomical Structure** did not change, the relationship structures for nine of its descendants did not change either in the ESN (the tenth descendant **Gene or Genome** inherits a new relationship from its new parent **Molecular Sequence**.).

Table 3.6 Semantic Types with Different Relationship Structures in the SN and the ESN

Semantic Type	# relshps in SN	# relshps in ESN	Diff.
Anatomical Entity	0	3	3
Physical Anatomical Entity	0	3	3
Conceptual Anatomical Entity	0	3	3
Material-Physical Anatomical Entity	0	3	3
Body Space or Junction	42	43	1
Body Location or Region	34	35	1
Body System	5	6	1
Body Substance	28	29	1
Gene or Genome	72	73	1
Enzyme	86	87	1
Injury or Poisoning	86	178	92
Laboratory or Test Result	105	127	22
Organism Attribute	69	121	52
Clinical Attribute	69	121	52
Receptor	86	153	67
Vitamin	86	109	23
Total: 16			326

3.4.2 ESN Concept Configuration

The mapping function did not include all 5,653 existing redundant categorizations in the ESN's concept configuration. For example, **Enzyme** has the old parent **Biologically Active Substance** in the ESN. Among the 19,226 concepts assigned to **Enzyme**, 54 were also assigned to **Biologically Active Substance**. Therefore, the assignments of the 54 concepts to **Biologically Active Substance** would be redundant categorizations in the ESN because they can be inferred by the assignments to **Enzyme**. All 54 of those redundant categorizations are not included in the mapping process.

Altogether, the mapping function prevented 21,297 potential new redundant cate-

gorizations in the process of constructing the ESN's concept configuration. For example, the semantic type **Enzyme** has a new parent **Amino Acid, Peptide, or Protein**. There are 19,226 concepts assigned to **Enzyme**. Among them, 18,941 concepts are also assigned to **Amino Acid, Peptide, or Protein**, and 88 are assigned to **Organic Chemical**, which is the parent of **Amino Acid, Peptide, or Protein**. The new IS-A relationship would have made these assignments redundant categorizations if the mapping function did not prevent the assignments to the new parent and ancestor. Table 3.7 shows all the potential new redundant categorizations prevented by the mapping function. Column 2 shows the number of concepts in the child semantic type. Column 3 shows the new parent(s) (or ancestors) of the child semantic type, the assignments to which would become redundant categorizations. Column 4 contains the number of prevented redundant categorizations with respect to the different new parents or ancestors.

Table 3.7 Redundant Categorizations Involving Two Semantic Types having New IS-A Links

Child ST	# Concepts	New Parent (Ancestor) ST	# Joint Concepts
Organism Attribute	2,381	Physiologic Function	14
Injury or Poisoning	30,778	Disease or Syndrome	556
		Pathologic Function	105
Enzyme	19,266	Amino Acid, Peptide, or Protein	18,941
		Organic Chemical	88
Vitamin	1,208	Pharmacologic Substance	948
		Organic Chemical	644
		Chemical Viewed Structurally	1
Total: 4	53,633	8	21,297

Another example is **Vitamin**, which has two new parents in the ESN (see Fig-

ure 3.2): **Organic Chemical** and **Pharmacologic Substance**. Among the 1,208 concepts assigned to **Vitamin**, 644 were also assigned to **Organic Chemical**, and 948 were also assigned **Pharmacologic Substance**. The mapping function also prevented these potential redundant categorizations.

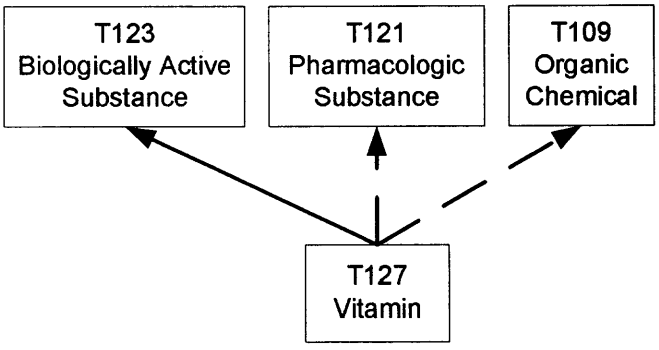


Figure 3.2 Vitamin with its parents in the ESN.

In total, the mapping function avoids the generations of 26,950 (5,653+21,297) redundant categorizations in the construction of the ESN's concept configuration.

3.5 Discussion

The ESN has 12% more IS-As than the SN: 150 vs. 133. In contrast, the increase in the number of (semantic) relationships is only about 4.8% ($314 + 12 = 326$ new relationships). The main reason for the relatively low impact of the extra IS-As on the increase in the number of relationships in the ESN is the position of these IS-As. Most of the semantic types with multiple parents are leaf semantic types or parents of leaves. Thus, most of the increase in relationship numbers happens at leaf semantic types where no further inheritance occurs—and thus the expansion is limited.

An interesting issue regarding the design of the SN is whether its lack of a multiple-

parent configuration is due to (A) an *a priori* imposition of the tree structure, or (B) the fact that those who defined the IS-As did not see the need for multiple parents in proper modeling. Insight into this issue can be gained by examining whether the designers tried to compensate for the lack of multiple parents by explicitly introducing relationships at a semantic type that would have otherwise inherited them if multiple parents existed. To be more specific, if a semantic type **A** lacks an IS-A to another type **B**, the designers of the SN could have duplicated at **A** those semantic relationships defined at **B**, since they would not be inherited. If the designers of the SN had taken such steps, then the new IS-As of the ESN would not imply much of a difference between the relationship distributions of the ESN and the SN.

The observation in Section 3.4.1 is that such actions were not undertaken in the design of the SN. Only three such duplicate relationship introductions appear in the SN; they involve the relationship *result_of* at **Organism Attribute** and **Clinical Attribute** and *part_of* at **Anatomical Structure**. In the ESN, these three relationships were obtained by the respective types via inheritance rather than explicit introduction. On the other hand, the 314 new relationships that appear in the ESN were not defined previously at the proper semantic types in the SN. Hence, there is no evidence of an effort to compensate for the inability to model multiple parents with duplicate relationship introductions in the SN.

A similar issue can be raised regarding the assignment of concepts to semantic types. If, as before, an IS-A from **A** to **B** is lacking, the domain experts doing the concept assignment could have assigned to **B** all the concepts that were assigned to **A**. In this way, each such concept would be both in **A** and **B**, even though **A** IS-A **B** could not be

modeled. Such an effort was actually seen in the assignment of 18,941 concepts to **Amino Acid, Peptide, or Protein**, which are among the 19,226 concepts assigned to **Enzyme**. The only other such meaningful effort appears in the assignment of most of the vitamins to **Pharmacologic Substance** and **Organic Chemical**. See Table 3.7 for more details. Thus, the redundant categorizations that are (potentially) caused by the addition of an IS-A to the SN exactly expose the efforts utilized to accurately model the knowledge in the SN, where the IS-A did not originally appear. As in Table 3.7, this approach can be found in a few of the cases, but it does not seem to be a widespread policy applied systematically by the domain experts doing the concept assignments.

In summary, judging from studies of the impact of adding IS-As to the SN on the relationship distribution and concept configuration, a general systematic effort cannot be identified in the design of the UMLS to compensate for the lack of multiple parents. Nevertheless, there is more of a tendency to compensate in the assignments of concepts to types than in duplicate introductions of semantic relationships. The latter activity was found to be practically nonexistent.

3.6 Summary

The semantic relationship distribution in the ESN is more complex than that of the SN due to the new multiple-parent IS-A hierarchy. In this arrangement, relationships can be inherited from more than one source. In this chapter, a technique is presented for deriving the relationship distribution of the ESN from that of the SN. The technique sought to preserve relationship introductions and existing relationship inheritance. All the newly

inherited relationships were audited for semantic validity. Based on the audit step in the technique, the ESN's relationship distribution is obtained, consisting of a total of 7,303 relationships.

The entire set of assignments of concepts to types in the ESN was derived automatically according to three rules. The process ensured that a concept is only assigned to the most specialized semantic types that are appropriate. In this way, redundant categorizations were avoided completely, unlike in the SN.

The resulting complete ESN contains 139 semantic types, 150 IS-A links, and 7,303 semantic relationships. There are in total 1,013,876 concept assignments with an average of 7,294 per semantic type. Compared to the SN, the ESN serves as an extended and more refined abstraction of the UMLS's META.

CHAPTER 4
DESIGNING METASCHEMAS FOR THE ESN

4.1 Introduction

While the SN of the UMLS is an important abstraction of the META, it is still a difficult mechanism to employ for comprehension due to its large number of semantic types and semantic (i.e., non-IS-A) relationships. Some previous work has been done to help with the visualization and navigation of the UMLS knowledge. In [46], a Hypercard browser of Meta-1 (MetaCard) was adapted to enable users to continue the browsing process, extended from the Metathesaurus to a variety of different knowledge sources. In [57], a review about visualization and navigation of knowledge in the medical domain was presented. The notion of a metaschema was introduced in some previous work [23, 49], based on a partition of the SN [6]. A metaschema is a higher-level network that serves as a compact abstraction of the SN. As shown in [23, 49], the notion of metaschema offers various compact (partial) views which can help users in their orientation to the SN.

In the current version of the SN with its two-tree hierarchy, each semantic type has at most one parent semantic type and can inherit relationships only from this unique parent. Some semantic types are naturally specializations of more than one semantic type. The tree structure does not allow for this kind of multiple parents arrangement. To improve the SN's structure, two methodologies were presented in Chapter 2 to add IS-A links and obtain the Enriched Semantic Network (ESN), a network similar to the SN but permitting multiple parents.

Because the ESN has a more complex hierarchy than the current SN, it is even

more critical to develop an ESN metaschema to help in its orientation. This chapter will concentrate on extending the notion of metaschema to make it applicable to a DAG hierarchy network and thus to the ESN. A methodology to derive such a metaschema is also provided.

The rest of this chapter is organized as follows. Section 4.2 provides a brief review of the ESN. Section 4.3 introduces the notion of metaschema for a network having a DAG hierarchy. The requirements that a higher-level network must satisfy in order to be a metaschema are discussed. A method by which a metaschema can be derived from a partition of a network like the SN or the ESN is described. The separate description is intended to emphasize that for the same network, there may exist several useful metaschemas, corresponding to various partitions of the network. Section 4.4 presents two metaschemas of the ESN based on two different partitioning techniques that have previously appeared [6, 64]. One metaschema is the "qualified metaschema" ("Q-metaschema" for short) based on the partitioning technique in [64] which is a modification of the partition of the SN in [39]; another is the "cohesive metaschema" ("C-metaschema") based on the technique in [6]. Section 4.5 contains a comparison and evaluation of the two metaschemas. Section 4.5.1 introduces a general example to demonstrate how a user can employ a metaschema to help in orientation to the ESN. Other applications of the metaschema to auditing for classification errors and to the prevention of redundant classifications in the UMLS are also briefly discussed. A summary appears in Section 4.6.

4.2 Background

A partition of the SN into 15 groups was previously presented in [39]. Each group in this partition represents a subject area. Six qualities were proposed as desired for such a partition: semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. The semantic validity quality means that each group must be semantically coherent [39]. One way to assess a group's semantic validity is to see if its semantic types together with their IS-A links form a connected subgraph of the SN. This is called the *connectivity property* [64]. Since the SN's IS-A hierarchy consists of two trees, such a connected subgraph must form a tree with a unique root.

Some groups in the partition of [39] do not satisfy the connectivity property. Each such group comprises two or more trees. For example, the *Phenomena* group (Figure 4.1) contains two trees; one of them consists solely of **Laboratory or Test Result** having no IS-A links to any other members.

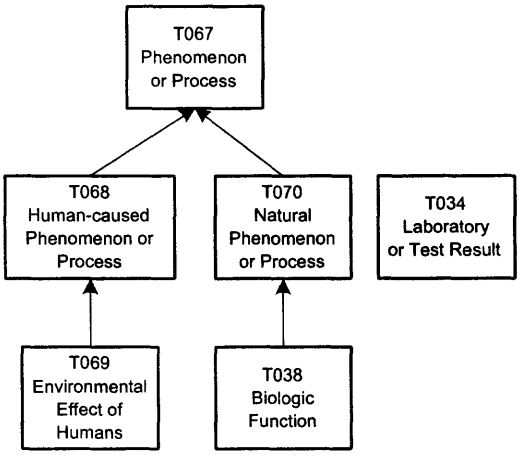


Figure 4.1 *Phenomena* group.

Another partitioning technique was developed to derive a cohesive partition of the SN in [23, 49, 6] which requires that all groups in the partition be connected. Following

the cohesive partition in [23, 49, 6], the *connectivity property* was enforced for all groups in the partition of [39] in the design of the ESN in Chapter 2. Four transformations were presented to convert each disconnected group into a new connected group, based on reviews of the definitions of all semantic types within a given disconnected group. During the transformations, new potential IS-A links were identified and then, where appropriate, were added. In Chapter 2, another methodology was also described to identify additional potential IS-A links for the SN. This methodology is based on string matching between names and definitions of various semantic types. Using this, four extra IS-A links were identified and added to the SN.

Based on above work, a new semantic network, referred to as the Enriched Semantic Network (ESN), was obtained, with an accompanied derived partition of the ESN. For an excerpt of the ESN hierarchy containing some of the descendants of **Entity**, see Figure 4.2. To emphasize the changes from the original SN, dashed thick arrows are used to denote the added IS-A links and thick dashed rectangles to denote new semantic types. Thin dashed rectangles denote semantic types that originally resided in the **Event** tree of the SN. An ellipsis in a rectangle indicates that some semantic types are not shown due to lack of space.

In the ESN, as in the SN, a pair of semantic types can be linked by 54 kinds of non-hierarchical (semantic) relationships. Each semantic type inherits all the semantic relationships of its parents via IS-A unless such an inheritance is explicitly blocked. Each concept of META is assigned to one or more of the semantic types. Thus the ESN is fairly complicated.

It is necessary to develop a metaschema to help in the orientation to the ESN. How-

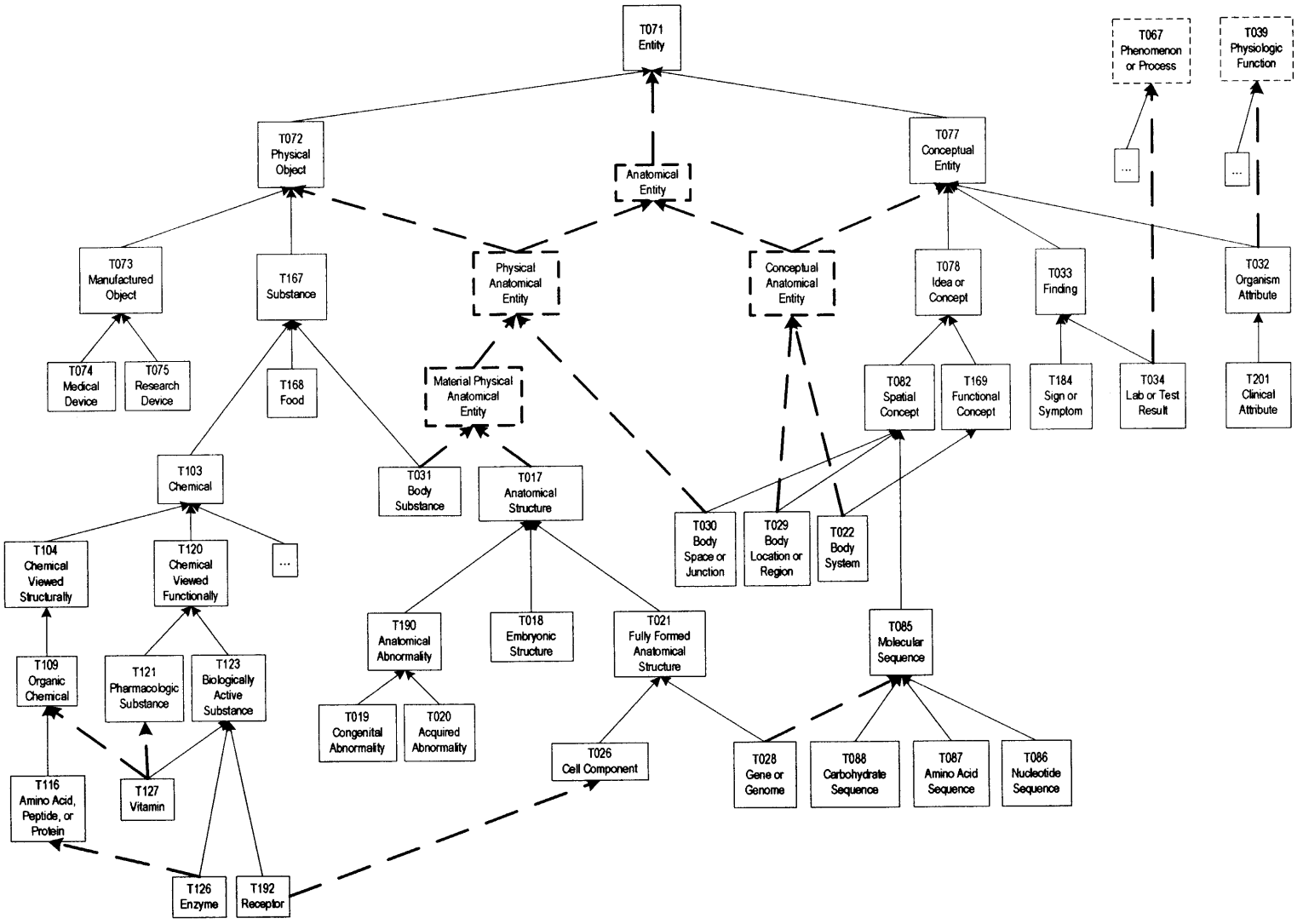


Figure 4.2 Part of the **Entity** component of the Enriched Semantic Network.

ever, since the ESN is a DAG rather than two trees, the definition of metaschema (as proposed in [23, 49]) is not applicable to the ESN. For example, in [23, 49] the hierarchical relationships of the metaschema were derived under the assumption that each semantic type had at most one parent. This is not true for the ESN. In the next section, the definition of a metaschema for a network with a DAG-structured hierarchy will be presented.

4.3 Methods

The requirements for and the actual derivation of a metaschema are presented in this chapter. In Section 4.3.1, the properties of a metaschema for a given semantic network are presented, independent of the way an actual metaschema is derived. For this, the requirements a network should satisfy to qualify as a metaschema of a DAG hierarchy network are described. The derivation of a metaschema is described in Section 4.3.2. The separate description is intended to emphasize that for the same semantic network, there may exist several useful metaschemas, corresponding to various partitions of the network.

4.3.1 Metaschema Requirements

For the requirements of a metaschema, some definitions are necessary.

Definition (Partition): A *partition* of a set V of elements is a family of subsets $\{V_1, V_2, \ldots, V_k\}$ such that $\bigcup_{i=1}^{k} V_i = V$, and $V_j \cap V_l = \emptyset$ when $j \neq l$. □

That is, a partition of V is a set of disjoint subsets such that each element of V belongs to exactly one subset.

A partition of the set of semantic types of the SN was presented in [39]. For ex-

ample, the *Phenomena* group of [39] is {**Phenomenon or Process, Human-caused Phenomenon or Process, Natural Phenomenon or Process, Laboratory or Test Result, Environmental Effect of Humans, Biologic Function**} (Figure 4.1). However, the SN is more than the set of its semantic types; it is a network where the semantic types are connected via hierarchical (IS-A) and non-hierarchical (semantic) relationships. Thus, it is important to consider a partition of a graph (network) rather than a set, particularly a partition of the hierarchy of the SN consisting of the semantic types and all the IS-A relationships connecting them. For this, the following definition is provided. In all the discussions a graph refers to a directed graph.

Definition (Induced Subgraph): An *induced subgraph* of a graph $G = (V, E)$ induced by a subset of nodes V' ($V' \subseteq V$) is a graph $G' = (V', E')$ where E' contains all the edges of E for which both endpoints are in V'. □

In other words, the V'-induced subgraph of G contains the nodes in V' and all the edges of G connecting them. For example, when G is the hierarchy of the SN, the graph induced by the *Phenomena* group of [39] appears in Figure 4.1.

Definition (Partition of a DAG): A *partition* of a DAG $G = (V, E)$ based on a partition $\{V_1, V_2, \ldots, V_k\}$ of V is a collection of subgraphs $\{G_1, G_2, \ldots, G_k\}$ where $G_j = (V_j, E_j)$, $1 \leq j \leq k$, is the subgraph of G induced by V_j. □

Definition (Connected Partition): A partition of a graph is a *connected partition* if each of its subgraphs is a connected graph having a unique root. □

Note that a connected subgraph of a tree must have a unique root, but this is not necessarily true for a DAG. Thus, when dealing with the ESN having a DAG hierarchy,

rather than the SN having a tree hierarchy, the requirement for a unique root must be added to the definition.

The partition of the SN hierarchy of [39] is not a connected partition since, for example, the subgraph of the *Phenomena* group is not connected. (See left subgraph of Figure 4.3). On the other hand, the partition of the ESN in [64] is a connected partition. For example, see the subgraph of the *Phenomenon or Process* group in the right subgraph of Figure 4.3.

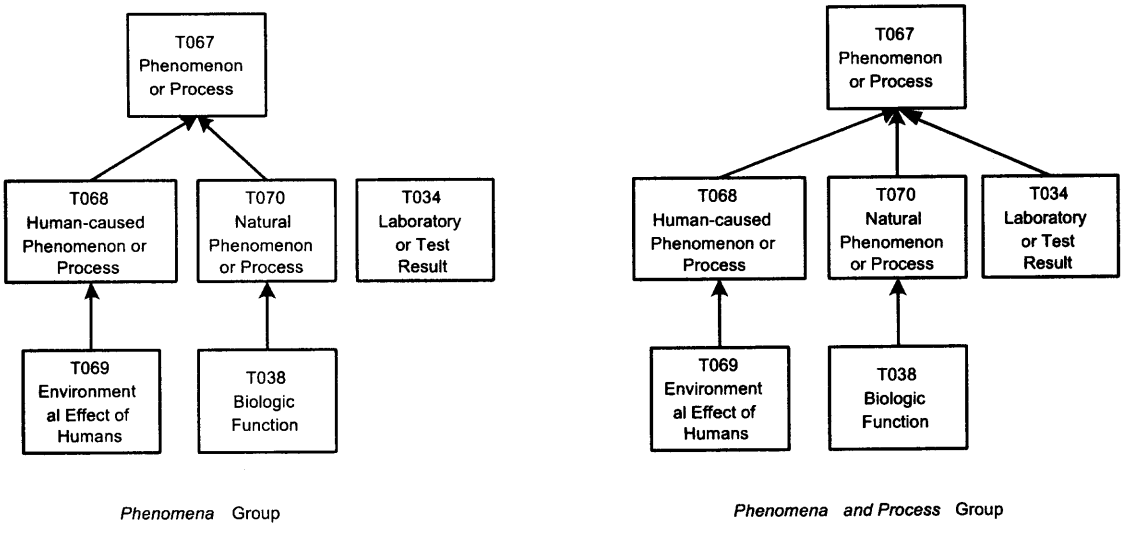


Figure 4.3 *Phenomena* Group vs. *Phenomena or Process* group.

Based on the above definitions, the notion of a metaschema for a DAG can be defined as follows.

Definition (Metaschema): A *metaschema* of a network G with a DAG hierarchy is a directed network which consists of a set of nodes called meta-semantic types (MSTs) connected via hierarchical *meta-child-of* relationships and non-hierarchical meta-relationships satisfying the following two conditions:

1. The set of MSTs represents a connected partition of the given DAG hierarchy.

2. The hierarchy of the metaschema which consists of MSTs and all the *meta-child-of* relationships connecting them is a DAG.

The reason for condition 1 is that an MST standing for a set of semantic types, say, S represents the subgraph of G induced by S. That is, a set of semantic types together with all their hierarchical relationships and semantic relationships. The set of subgraphs of G's hierarchy induced by the set of MSTs in a metaschema make up a connected partition of G. The reason for condition 2 is obvious: in order to qualify as a hierarchy a network must be a DAG; a cycle contradicts the notion of a hierarchy of its nodes.

4.3.2 Metaschema Derivation

A metaschema will be derived based on a connected partition. For each group of the partition, a meta-semantic type (MST) is defined to represent the group. The MST is named after the unique root of the corresponding group. The term "root of an MST" denotes the semantic type which is the root of the semantic-type group represented by this MST. After defining the MSTs, the *meta-child-of* relationships and the meta-relationships for the metaschema will be derived.

Let $\{G_1, G_2, \ldots, G_k\}$ be a connected partition of a network G with a DAG hierarchy. Then semantic type A is the unique root of the semantic-type group represented by MST A (called the root of A for short).[1] Since G has a DAG hierarchy, **A** may have several parents $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_j$. There are two cases.

Case 1: All j parents are associated with a single MST B.

[1] An italic font will be used for MSTs in this chapter.

Then a *meta-child-of* relationship in the metaschema is defined from A to B. All semantic types associated with A are descendants of the root semantic type **A**. Since all **A**'s parents are descendants of the root semantic type **B** of B, all semantic types in A are descendants of semantic type **B** of B. □

Case 2: The j parents $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_j$ are not associated with one MST.

Suppose these j parents are associated with l MSTs $M_1, M_2, \cdots, M_l$. Then there should be a *meta-child-of* relationship from A to each of the l MSTs. Therefore, all semantic types associated with A are descendants of each of the roots M_i $(1 \leq i \leq l)$ of the l MSTs. □

After the hierarchical *meta-child-of* relationship is derived for the metaschema, the meta-relationships between two MSTs is obtained as follows.

Let **A** be the root of the MST A, and let $\mathbf{B}_i$ be a semantic type in the MST B. If in the original network there exists a semantic relationship `rel` connecting **A** to $\mathbf{B}_i$, then in the metaschema there exists a link labeled "`rel`"[2] connecting A to B. Such a link is called a meta-relationship.

Note that semantic type $\mathbf{B}_i$ does not need to be the root of B, but the source semantic type of the `rel` relationship must be the root **A** of A. Sometimes in the original network, there is a semantic relationship $\texttt{rel}_1$ from semantic type **C** to semantic type **D** where **C** is not a root of an MST. This does not mean that there exists a meta-relationship $\texttt{rel}_1$ from the MST associated with **C** to the MST associated with **D**. The reason for this asymmetry in the requirements for the source and target semantic types of meta-relationships is as

[2]A courier font will be used for semantic relationships and meta-relationships in this chapter.

follows. For a meta-relationship `rel` to be defined from MST *A* to MST *B*, it is ideal to have a situation that for each semantic type $\mathbf{A}_j$ of MST *A*, there should be some semantic type $\mathbf{B}_i$ of MST *B* such that $\mathbf{A}_j$ `rel` $\mathbf{B}_i$. For this `rel` should be defined at the root semantic type **A** of MST *A*, so `rel` is inherited by all semantic types of MST *A*, which are all descendants of the root semantic type **A**. Such a requirement is not needed for the target semantic type $\mathbf{B}_i$ of the relationship, since not every semantic type in MST *B* has to be a target of such a relationship. It is enough that there exists some semantic type in MST *B* which is a target of `rel` for each source semantic type $\mathbf{A}_j$ of MST *A*.

To reflect the relationship inheritance of the original network, the inheritance of meta-relationships can be defined along the hierarchical *meta-child-of* relationships in the metaschema. Suppose there exist three MSTs *A*, *B*, and *C*, where a *meta-child-of* link connects *B* to *A*. If there is a meta-relationship `rel` from *A* to *C*, then *B* also has a meta-relationship `rel` to *C*, and to all MSTs that have *meta-child-of* links or a chain of *meta-child-of* links to *C*.

The relationships of the metaschema should reflect the relationships in the SN. For example, if *A* is *meta-child-of B*, then every semantic type in *A* should be a descendent of some semantic type in *B*. Similarly, if there is a meta-relationship `rel` from *A* to *B*, then there should be a relationship `rel` defined for every semantic type in *A* to some semantic type in *B*.

In Section 4.4, the metaschema derivation described will be applied to the ESN network with its DAG hierarchy.

4.4 Results: Two Metaschemas

For a given semantic network, any connected partition leads to a metaschema. Each such metaschema will be named after its partition. In this section, two possible metaschemas for the ESN are presented, both derived using the method given in the previous section.

4.4.1 Qualified Metaschema of the ESN

Definition (Qualified Partition): A partition of a set is called a *qualified partition* if it possesses the six qualities (principles) listed in [39]: semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. □

Note that "Q-partition" is used as an abbreviation for "qualified partition" throughout the remainder of the chapter.

The partition of the SN in [39] is a Q-partition but not a connected partition. Thus, it cannot be used to derive a metaschema for the SN. However, the partition of the ESN obtained in [64] is a connected Q-partition. Thus, a metaschema can be defined based on the connected Q-partition of the ESN. The resulting metaschema is referred to as the qualified metaschema (Q-metaschema for short).

The hierarchy found in each group in the Q-partition [64] is a tree with a unique root. An MST whose name is the root of the group is defined to represent each group. Therefore, a metaschema of 19 MSTs (see Table 4.1) is obtained.

Now, it is time to derive the hierarchical *meta-child-of* relationships for the Q-metaschema relating to the above Q-partition. For example, the root of MST *Phenomenon or Process* is the semantic type **Phenomenon or Process** which is a child of **Event**. **Event**

Table 4.1 MSTs, Semantic Types (STs), and Meta-relationships in the Q-metaschema

MST	# of STs contained in	# of outgoing meta-relationships
Anatomical Abnormality	3	11
Anatomical Entity	15	1
Chemical	25	2
Clinical Drug	1	0
Conceptual Entity	12	0
Entity	4	1
Event	7	1
Finding	2	8
Geographic Area	1	2
Group	6	7
Manufactured Object	4	2
Molecular Sequence	5	0
Occupation or Discipline	2	1
Occupational Activity	9	3
Organism	17	1
Organization	4	3
Pathologic Function	7	14
Phenomenon or Process	6	2
Physiologic Function	9	4
Total: 19 MSTs	139	63

is associated with *Event*; hence, there is a *meta-child-of* from *Phenomenon or Process* to *Event* in the Q-metaschema. The root of *Pathologic Function*, the semantic type **Pathologic Function**, is a child of **Biologic Function** which resides in *Phenomenon or Process*. Thus, there exists a *meta-child-of* from *Pathologic Function* to *Phenomenon or Process*.

By applying this *meta-child-of* derivation process to all 19 MSTs, the entire Q-metaschema hierarchy consists of 17 *meta-child-of* links. Figure 4.4 shows this hierarchy. Each node contains the name of the MST and the number of constituent semantic types

written in parenthesis. It is interesting to note that no root of a group in the Q-partition actually has more than one parent. Multiple parents occur only for non-root semantic types in the Q-partition. Hence, the hierarchy of the Q-metaschema has a two-tree structure, as did the original SN.

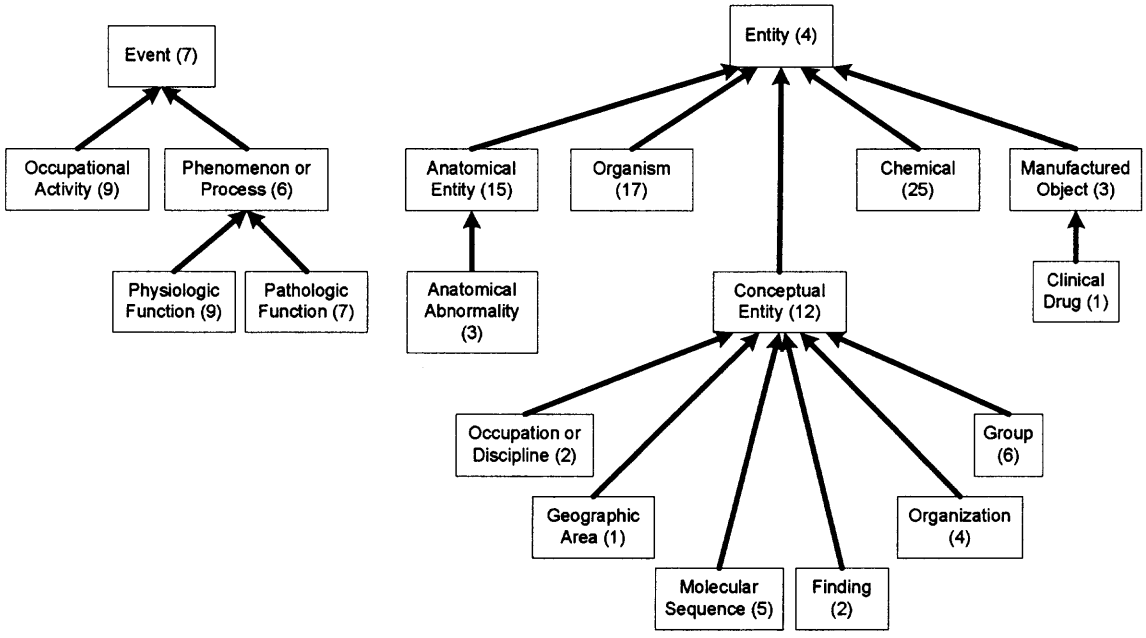


Figure 4.4 The Q-metaschema hierarchy of the ESN based on the Q-partition.

Besides the *meta-child-of* relationships, the metaschema also has meta-relationships. For example, **Pathologic Function** introduces the `manifestation_of` relationship to **Physiologic Function**. Since **Pathologic Function** is the root of *Pathologic Function*, and **Physiologic Function** is in *Physiologic Function*, there is a `manifestation_of` meta-relationship from *Pathologic Function* to *Physiologic Function* in the metaschema. There is a relationship `occurs_in` from **Pathologic Function** to **Group**. Thus, there is also an `occurs_in` meta-relationship from *Pathologic Function* to *Group*. Meanwhile, **Pathologic Function** also defines `co-occurs_with`, `complicates`, `manifestation_of`, and `occurs_in` relationships to **Injury or Poisoning**, which is in *Finding*. Thus, there

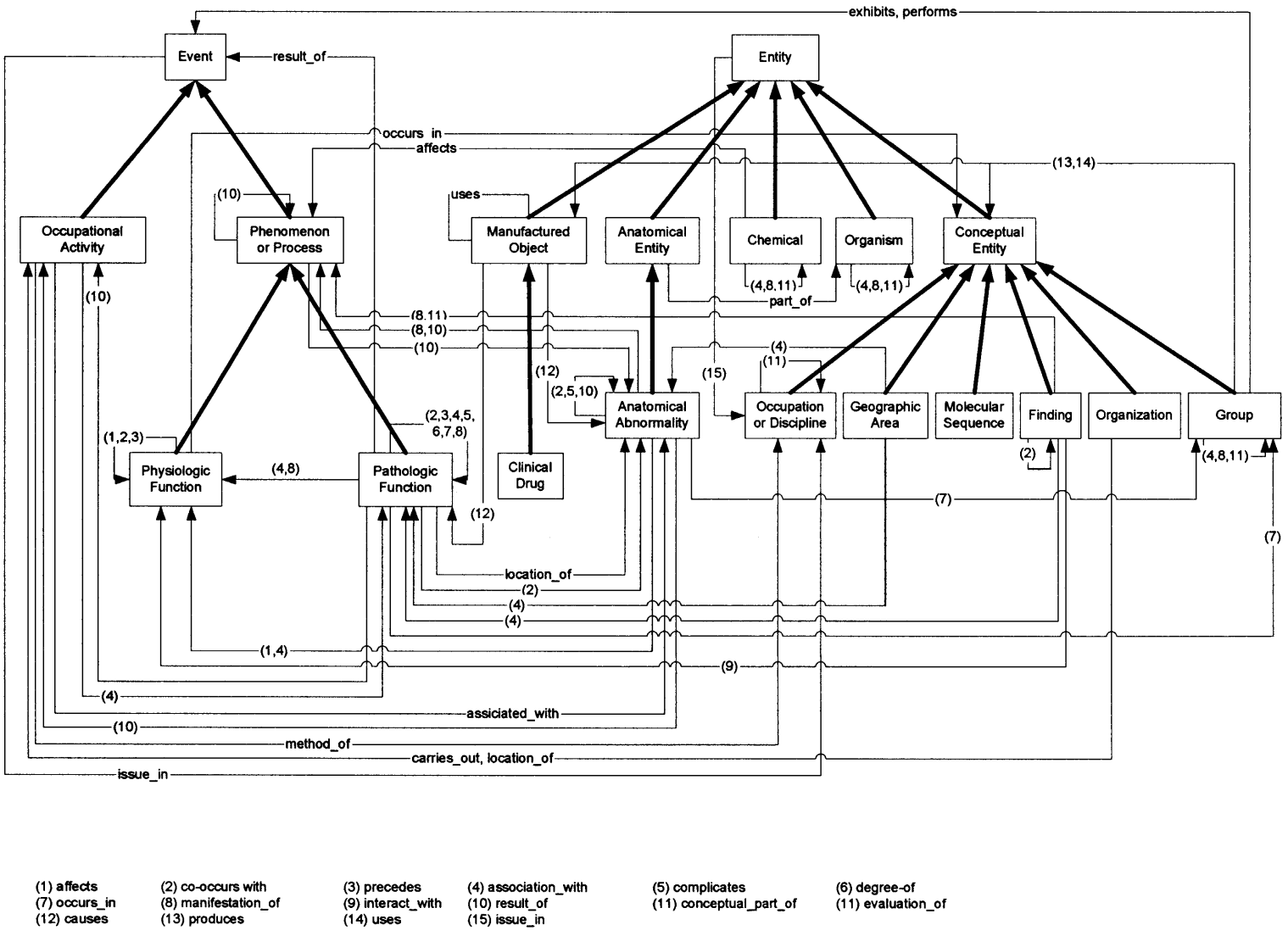


Figure 4.5 Q-metaschema of the ESN based on the Q-partition of [64].

are also meta-relationships `co-occurs_with`, `complicates`, `manifestation_of`, and `occurs_in` from *Pathologic Function* to *Finding*.

In total, there are 63 meta-relationships belonging to 22 kinds of relationships. Figure 4.5 shows the whole Q-metaschema, including its 19 MSTs, 17 *meta-child-of* relationships, and 63 meta-relationships. Note that only the meta-relationships introduced at an MST are shown in the figures; the inherited meta-relationships are not shown to avoid clutter. The existence of these additional relationships is easily derived from the figure. A thick arrow denotes a *meta-child-of* relationship, while a labeled thin arrow denotes a meta-relationship. This metaschema, which is displayed on only one page, serves as a compact abstraction of the ESN and can help with user orientation.

4.4.2 Cohesive Metaschema of the ESN

The technique for deriving a metaschema for the SN described in [23, 49] first defined the "structure" of a semantic type as the set of its defined relationships, either introduced directly or inherited. Semantic types with the same structure were grouped as one semantic-type group. Thus, a structural partition of the SN was obtained. However, that partition was not connected. By applying the three rules defined in [49], a cohesive partition was obtained, consisting of cohesive (singly rooted) semantic-type collections. An MST was then defined to represent each cohesive semantic-type collection. It should be noted that elements of the structural partition were called groups to distinguish from elements of the cohesive partition that were called collections. Based on the cohesive partition, the cohesive metaschema of the SN was derived in [23, 49].

Now the second metaschema of the ESN, referred to as the cohesive metaschema based on an application of the methodology of [23, 49], will be derived. First, A structural partition of the ESN will be obtained. Note that the structural partition of the ESN will differ from the structural partition of the SN due to the multiple parent configuration and the new distribution of inherited relationships. Then the three rules are applied to derive a cohesive partition from the structural partition. Finally, the method of Section 4.3.2 is used to obtain the cohesive metaschema of the ESN. The term "C-metaschema" and "C-partition" are used as abbreviations for the cohesive metaschema and the cohesive partition of the ESN, respectively.

4.4.2.1 Cohesive Partition of the ESN

Since the structural partition depends on the relationships defined at semantic types, it is important to note the relationships of the four new semantic types of the ESN. Following the precedent set by the Digital Anatomist Foundational Model [43], the new **Anatomical Entity** semantic type in the ESN is defined as "a biologic entity which forms the whole or part of or is an attribute of the structural organization of a biological organism." Thus, **Anatomical Entity** introduces the `part_of` relationship directed at **Organism**[3] instead of having its descendant **Anatomical Structure** introduce it, as in the current SN. Thus, in the ESN, **Anatomical Structure** inherits `part_of` from **Anatomical Entity**; it still introduce the `location_of` relationship. The introduction of these relationships is relevant to the structural partition of the ESN, as each of these two semantic types is a root of a semantic-type group.

[3]Cornelius Rosse, personal communication, 2002.

For the ESN, its structural partition consists of 74 semantic-type groups. Most of these contain only one semantic type. Such groups are called singletons. See Table 4.2 for the distribution of the numbers of groups according to their sizes.

Table 4.2 Size Distribution of Semantic-type Groups

# of STs in group	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of groups	50	10	5	4	3	0	0	1	0	0	0	0	0	1

To obtain the C-partition of the ESN, the following three rules [49] are applied to the 74 semantic-type groups.

Rule 1: Each semantic-type group with a non-leaf unique root becomes a semantic-type collection and is named after its root. □

Rule 2: If a leaf semantic type L is a singleton in the structural partition, then L is added to its parent's semantic-type collection. □

Rule 3: Let the semantic types $A_1, A_2, \ldots, A_n$ $(n \geq 2)$ be roots of the same semantic-type group G of the structural partition. If there exists a lowest common ancestor A of $A_1, A_2, \ldots, A_n$ in the IS-A hierarchy, then add all the semantic types of G to the semantic-type collection of A. □

However, in applying Rule 2, there are eight leaf singletons that have multiple parents. Note that some leaves with multiple parents are not singletons as they share the same structure (relationship set) and thus the same group with one of the parents. For example, **Vitamin** has three parents, but it has the same structure as its parent **Biologically Active Substance** and is thus in the same group as that semantic type.

Each of the eight leaf singletons has a different relationship set from all its parents. Besides this, its parents exhibit different structures and thus are not in the same semantic-

type group. Rule 2 stated that a leaf singleton should be merged into its parent's semantic type collection. In such a case of multiple parents, it is required to determine to which semantic-type collection each singleton should be added since each semantic type must belong to exactly one semantic-type collection in the C-partition. For this, it is necessary to differentiate between different kinds of parents of such a singleton. Among the parents, only one parent will be identified as the "primary parent" of the singleton; other parents will be considered "secondary parents." The singleton will then be merged into the group of its primary parent. Of course, if the singleton has only one parent, then this parent is considered the primary parent. The process of identifying the primary parent is discussed in the following subsection.

4.4.2.2 Identifying the Primary Parent among Multiple Parents

The process of differentiating multiple IS-A links from a singleton to all its parents is guided by the analysis of the names and definitions of the singleton semantic type and its parents. The following guidelines are provided, which are modifications of the guidelines in [21, 22].

A definition of a singleton semantic type is distinguished among three kinds: the descriptive kind, the functional kind, and the characterizing kind. The descriptive kind captures the essence or nature of the semantic type. The functional kind captures the functionality or usage of the semantic type. The characterizing kind does not describe the essence of the knowledge or its function, but characterizes what kind of knowledge is represented. A definition sometimes has both a descriptive part and a functional part.

For each singleton semantic type having multiple parents, it is important to find,

from all its parent semantic types, which parents are descriptive, which parents are functional, and which parents are characterizing. Typically, all parents contribute to the definition of the child; a descriptive parent highlights the essence or nature of the child semantic type; a functional parent highlights the function or usage of the child semantic type; a characterizing parent classifies the kind of knowledge rather than concentrating on the knowledge itself.

Case 1: *Some of the parents are descriptive and the others are functional.*

First check the descriptive part and the functional part of the singleton's name or definition, and determine which part is the primary part.

If the primary part is the descriptive part and there is only one descriptive parent, then choose this descriptive parent as the primary parent; otherwise, choose the primary parent from among the group of descriptive parents using Case 2.

If the primary part is the functional part and there is only one functional parent, choose this functional parent as the primary parent; otherwise, choose the primary parent from among the group of functional parents using Case 3. □

Case 2: *All parents are descriptive* (*or the primary part of the singleton's name or definition is descriptive*). Among these descriptive parents, distinguish the primary parent by linguistic analysis of the name or definition of the singleton. If the name of one parent is used as a noun and the names of the other parents are used as adjectives in the singleton's name or definition, then the noun defines the primary parent.

If the names of all parents are used as nouns in the name or definition of the singleton, then the last noun is considered the primary noun. The corresponding parent is

chosen as the primary parent. If the names of all parents are used as adjectives in the name or definition of the singleton, then the adjective closest to the noun in the name or definition is considered the primary adjective. The corresponding parent is chosen as the primary parent. □

Case 3: *All parents are functional (or the primary part of the singleton's name or definition is functional).*

Again, use the linguistic analysis described in Case 2 to identify the primary (functional) parent. □

Case 4: *Some parents are characterizing.*

Examples of such parents are: **Physical Object**, **Functional Concept**, **Spatial Concept**, and **Conceptual Entity**.

The only case where such a parent semantic type will be the primary parent of a child semantic type is when the child is also considered characterizing. In all other circumstances, another parent will be picked as primary using the other three cases after removing the characterizing parents from consideration. □

In each case, the singleton is merged into the collection of its primary parent in the partition. To capture the situation of a singleton with more than one parent, Rule 2 defined in [23, 49] must be restated as:

Rule 2′: If a leaf semantic type L is a singleton in the structural partition, then L is added to its primary parent's semantic-type collection.

The following will demonstrate how to identify the primary parent for the eight

singletons with multiple parents, following the method described above. For example, consider **Enzyme**, a singleton in the structural partition of the ESN having two parents. One is the old parent **Biologically Active Substance**; the other one is the new parent **Amino Acid, Peptide, or Protein**. **Enzyme** is defined as "a complex chemical, usually a protein, that is produced by living cells and which catalyzes specific biochemical reactions." The descriptive part in the definition is "a complex chemical, usually a protein," while the functional part is "that is produced by living cells and which catalyzes specific biochemical reactions." The two parents' definitions are reviewed. **Biologically Active Substance** is defined as "a generally endogenous substance produced or required by an organism, of primary interest because of its role in the biologic functioning of the organism that produces it." This definition emphasizes the role (or usage) of the substance, in this case **Enzyme**, in an organism. Hence, **Biologically Active Substance** is a functional parent. **Amino Acid, Peptide, or Protein** is defined as "amino acids and chains of amino acids connected by peptide linkages." This describes the chemical composition of **Enzyme**. (Enzyme is a kind of protein.) Therefore, **Amino Acid, Peptide, or Protein** is a descriptive parent. Since one parent is functional and the other one is descriptive, both the descriptive part and the functional part of **Enzyme**'s definition has to be checked. Finally the primary part of the definition must be determined. It is clear that what makes enzyme different from other proteins lies in its function (usage), which is catalyzing specific biochemical reactions of an organism. Thus, the functional part of **Enzyme** is the primary part of its definition. So, the functional parent **Biologically Active Substance** is chosen as the primary parent, and **Enzyme** will be merged into the *Biologically Active Substance* group.

As another singleton example, **Gene or Genome** has two parents in the ESN. One is the old parent **Fully Formed Anatomical Structure**; the other one is **Molecular Sequence**. First, the definition of **Gene or Genome** is reviewed, which is defined as "a specific sequence, or in the case of the genome the complete sequence, of nucleotides along a molecule of DNA or RNA (in the case of some viruses) which represent the functional units of heredity." In the definition, the descriptive part is "a specific sequence, or in the case of the genome the complete sequence, of nucleotides along a molecule of DNA or RNA (in the case of some viruses)." The functional part is "which represent the functional units of heredity." Next the definitions of its two parents are also checked. **Fully Formed Anatomical Structure** is defined as "an anatomical structure that exists only before the organism is fully formed; in mammals, for example, a structure that exists only prior to the birth of the organism. This structure may be normal or abnormal." This definition is descriptive as no function is discussed. **Molecular Sequence** is defined as "a broad type for grouping the collected sequences of amino acids, carbohydrates, and nucleotide sequences." This definition is also descriptive since it does not discuss the function or usage of **Gene or Genome**. Since both parents are descriptive, the linguistic analysis has to be used to distinguish the primary parent from the secondary one. In the definition of **Gene or Genome**, the primary noun is "sequence"; therefore, **Molecular Sequence** is the primary parent, and **Gene or Genome** will be merged into the *Molecular Sequence* group.

Some leaf singleton semantic types with two parents have one parent which is a characterizing parent, while the semantic type is not of the characterizing kind. Both **Body Location or Region** and **Body Space or Junction** have the characterizing **Spatial Con-**

cept as a parent. **Body System** has the characterizing **Functional Concept** as parent. All these parents are considered to be secondary while the primary parent semantic types are **Physical Anatomical Entity** and **Conceptual Anatomical Entity** respectively (where by linguistic analysis "anatomical" is the primary adjective being closer to the noun in the name of the semantic type). Note that although these two primary parents have a characterizing part in their names, namely Physical and Conceptual, these two parts are considered secondary in the names of the parents.

Table 4.3 Primary/Secondary Parents for Singletons Having Multiple Parents

Singleton	Primary parent ST	Secondary parent ST
Body Location or Region	Conceptual Anatomical Entity	Spatial Concept
Body Space or Junction	Physical Anatomical Entity	Spatial Concept
Body System	Conceptual Anatomical Entity	Functional Concept
Body Substance	Material Physical Anatomical Entity	Substance
Enzyme	Biologically Active Substance	Amino Acid, Peptide, or Protein
Gene or Genome	Molecular Sequence	Fully Formed Anatomical Structure
Laboratory or Test Result	Phenomenon or Process	Finding
Receptor	Biologically Active Substance	Cell Component

By using the above guidelines, the primary parent for each leaf singleton having multiple parents is determined (see Table 4.3). Those singletons will be merged into the groups of their primary parents according to the revised Rule 2′. When applying the three rules to the 74 semantic-type groups, 29 collections of semantic types, called cohesive semantic-type collections, are obtained (see Table 4.4). The "*# of STs*" column is the

number of semantic types in each semantic-type collection. The "*# of rel.*" column in Table 4.4 is the number of semantic relationships introduced by the root of each semantic-type collection in the ESN. These relationships will imply the meta-relationships in the derivation of the ESN's cohesive metaschema. The 29 collections together form a partition, called the cohesive partition ("C-partition" for short).

Table 4.4 Semantic-type Collections of the ESN C-metaschema

Semantic-Type Collection	# STs	# rel.	Semantic-Type Collection	# STs	# rel.
Anatomical Abnormality	3	11	Anatomical Entity	8	1
Anatomical Structure	2	1	Animal	9	1
Behavior	3	10	Biologic Function	1	4
Biologically Active Substance	7	7	Chemical	16	2
Entity	8	1	Event	4	1
Finding	2	8	Fully Formed Anatomical Structure	5	7
Group	6	7	Health Care Activity	4	1
Idea or Concept	12	2	Manufactured Object	5	2
Natural Phenomenon or Process	1	2	Occupation or Discipline	2	1
Occupational Activity	3	3	Organism	6	1
Organism Attribute	2	6	Organization	4	3
Pathologic Function	7	14	Pharmacologic Substance	2	11
Phenomenon or Process	4	2	Physiologic Function	7	4
Plant	2	1	Research Activity	2	7
Substance	2	3			
Total: 29 Groups				139	124

It is important to stress here that the IS-A link from the singleton to the secondary parent is still part of the ESN. It is just labelled so the user can determine uniquely the

groups of the partition on which the metaschema is based. Interestingly in most cases, the secondary parent was the original parent in the SN, while the connection to the primary parent is a newly added IS-A link.

4.4.2.3 Derivation of the Cohesive Metaschema

The derivation of the cohesive metaschema (C-metaschema) for the ESN is based on the above C-partition. For each cohesive semantic-type collection, an MST is defined to represent it. It is named after the root of the collection. The *meta-child-of* relationships and meta-relationships are derived as described in Section 4.3.2. The C-metaschema contains 29 MSTs, 28 *meta-child-of* relationships, and 124 meta-relationships belonging to 31 kinds of relationships. Figure 4.6 shows the cohesive metaschema hierarchy of the ESN with 29 MSTs. Note that this metaschema has a DAG hierarchy, which will be discovered further in Section 4.5. The number in each rectangle denotes the number of semantic types in the MST. Interestingly, the choice of the primary parents for the singleton leaves does not have influence on the metaschema itself, since no leaf is an MST in the metaschema. However, there is an impact on the underlying partition, reflected in the number of semantic types for some groups. Figure 4.7 shows the C-metaschema including all *meta-child-of* relationships and most meta-relationships. Unfortunately, there is insufficient space to draw all the meta-relationships.

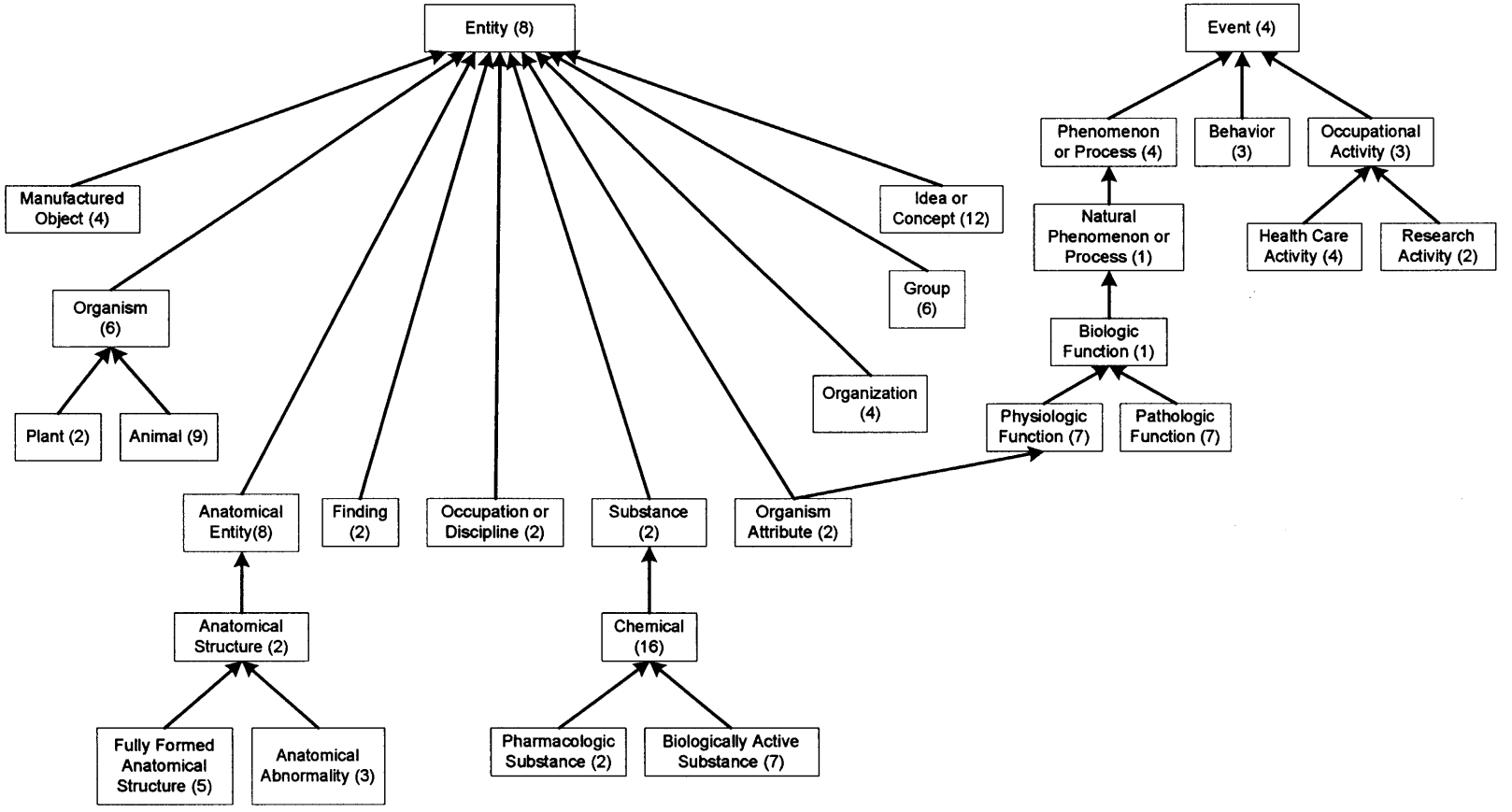


Figure 4.6 The C-metaschema hierarchy of the ESN.

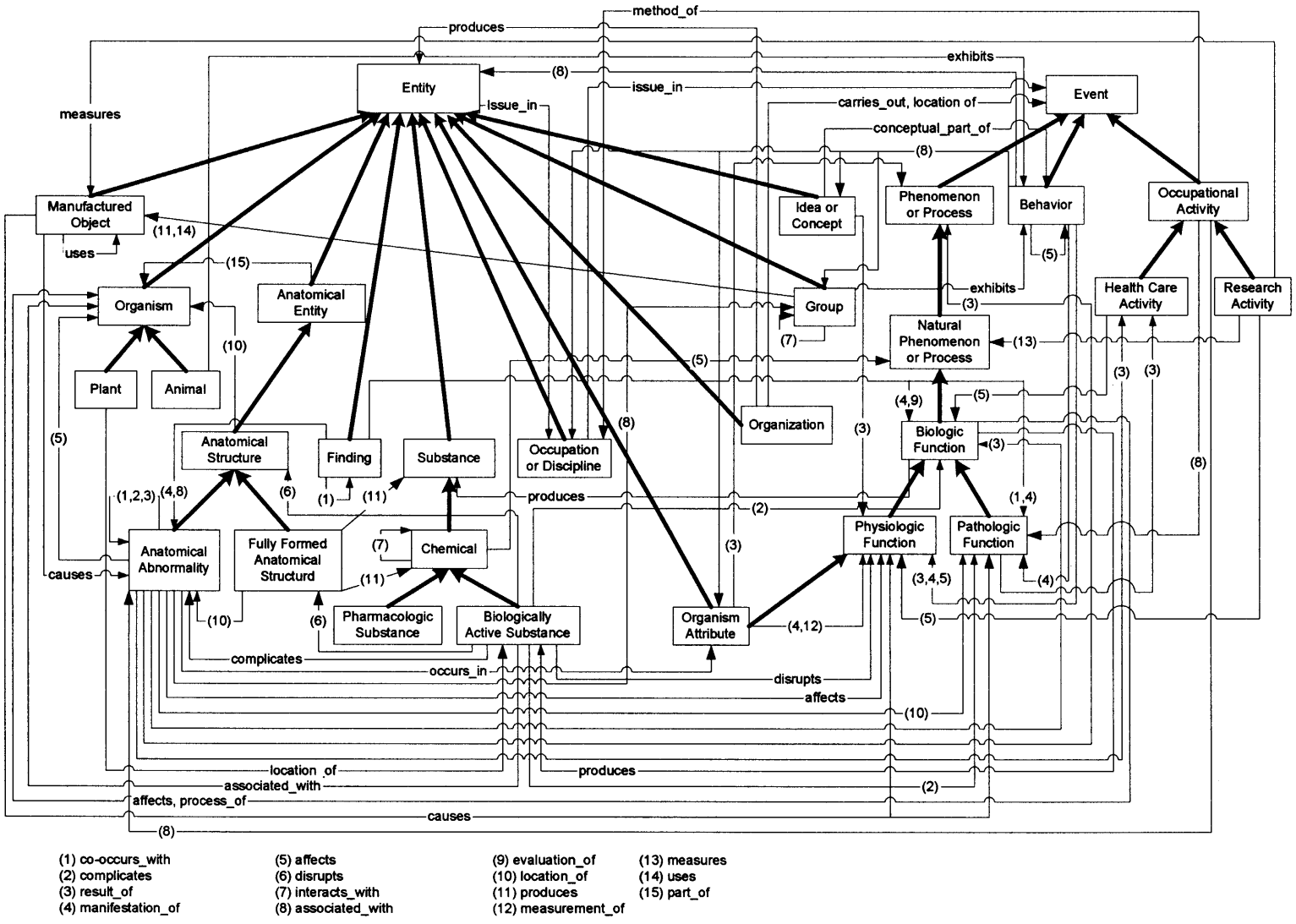


Figure 4.7 The C-metaschema of the ESN with most of its meta-relationships.

4.5 Comparison of Two Metaschemas

Based on the Q-partition of the ESN in [64], the Q-metaschema of the ESN was obtained (Figure 4.4). Modifying the method in [23, 49], the C-partition and the C-metaschema were derived (Figure 4.6) of the ESN. Each of the metaschemas provides an abstract view of the ESN.

The Q-metaschema contains 19 MSTs, while the C-metaschema contains 29 MSTs. There are some common MSTs between the two metaschemas. Among the 19 MSTs in the Q-metaschema, six also appear in the C-metaschema, representing the same semantic-type collections in both the Q-partition and the C-partition. That means both metaschemas agree that these six MSTs are quite important in the abstraction of the ESN, providing the metaschema with the representation of natural units of semantic types. These six MSTs are: *Anatomical Abnormality*, *Finding*, *Group*, *Occupation or Discipline*, *Organization*, and *Pathologic Function* (see Table 4.5). Together they cover 24 semantic types (i.e., 17.4% of the ESN).

There are some obvious differences between the two metaschemas and their underlying partitions. The Q-metaschema contains two trees, while the C-metaschema is a DAG. In the Q-partition, semantic type **Organism Attribute** and its child **Clinical Attribute** are part of the *Physiologic Function* group. However, in the C-partition, these two semantic types form a separate semantic-type collection due to structural differences; hence, there is an MST named *Organism Attribute* in the C-metaschema. This MST has two parents in the C-metaschema: one is *Entity*, the other is *Physiologic Function*. These two *meta-child-of* relationships make the C-metaschema a DAG.

Table 4.5 Identical MSTs in Q-metaschema and C-metaschema

MST	Semantic-type collection
Anatomical Abnormality	Anatomical Abnormality; Acquired Abnormality; Congenital Abnormality
Finding	Finding; Sign or Symptom
Group	Group; Family Group; Age Group; Population Group; Professional or Occupational Group; Patient or Disabled Group
Occupation or Discipline	Occupation or Discipline; Biomedical Occupation or Discipline
Organization	Organization; Professional Society; Health Care Related Organization; Self-help or Relief Organization
Pathologic Function	Pathologic Function; Experimental Model of Disease; Disease or Syndrome; Injury or Poisoning; Neoplastic Process; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction

In the Q-metaschema, the MSTs *Clinical Drug* and *Geographic Area* each represent a semantic-type collection that contains only a leaf singleton semantic type. In the C-metaschema, there is no such case because Rule 2′ explicitly merges each leaf singleton into its parent's group. On the other hand, the C-metaschema contains two MSTs, *Natural Phenomenon or Process* and *Biologic Function*, that each represent a semantic-type collection consisting of only one internal (non-leaf) semantic type. This is because a semantic type like **Natural Phenomenon or Process** has a different structure (relationship set) from its parent and its child, and it is not merged into its parent's group since it is an internal node of the DAG.

There are also some other differences between the two metaschemas and their underlying partitions. Some semantic-type collections in the Q-partition are split into several different semantic-type collections in the C-partition, which results in several separated

MSTs in the C-metaschema. These MSTs in the C-metaschema are more refined than the corresponding MSTs in the Q-metaschema (Table 4.6). For the number of semantic types in the respective MSTs, see the parentheses alongside the MSTs' names in Figure 4.4 and Figure 4.6.

Table 4.6 Refined MSTs in the C-metaschema

MST in Q-metaschema	MST in C-metaschema
Anatomical Entity	Split into three MSTs: Anatomical Entity; Anatomical Structure; and Fully Formed Anatomical Structure
Chemical	Split into three MSTs: Chemical; Pharmacologic Substance; and Biologically Active Substance
Event	Split into two MSTs: Event; Behavior
Occupational Activity	Split into three MSTs: Occupational Activity; Health Care Activity; and Research Activity
Organism	Split into three MSTs: Organism; Plant; Animal
Phenomenon or Process	Split into three MSTs: Phenomenon or Process; Natural Phenomenon or Process; and Biologic Function
Physiologic Function	Split into two MSTs: Physiologic Function; Organism Attribute

For example, the *Chemical* group in the Q-partition is split into three semantic-type collections in the C-partition. One is *Pharmacologic Substance* containing **Pharmacologic Substance** and its child. One is *Biologically Active Substance* containing **Biologically Active Substance** and its children. The third is *Chemical* containing **Chemical** and all its descendants, except those in the *Pharmacologic Substance* and *Biologically Active Substance* semantic-type collections. This is because **Pharmacologic Substance** introduces the five relationships `complicates`, `diagnoses`, `disrupts`, `prevents`, and `treats`, and **Biologically Active Substance** introduces `associated_with`, `complicates`, and `disrupts`. Since these relationships are not defined at **Chemical**, **Pharmacologic Sub-**

stance and **Biologically Active Substance** start new MSTs.

Another example is the MST *Anatomical Entity* in the Q-metaschema. This MST represents a group of 15 semantic types. This group is split into three semantic-type collections in the C-partition. One collection contains **Anatomical Entity** and its seven descendants which are not in the other two collections; the second collection contains **Anatomical Structure** and **Embryonic Structure**; and the third contains **Fully Formed Anatomical Structure** and its children. This is because **Anatomical Structure** introduces a new relationship `location_of` that is not defined for its ancestors, and **Fully Formed Anatomical Structure** defines two new relationships, `contains` and `produces`. Hence, **Anatomical Structure** and **Fully Formed Anatomical Structure** both begin new MSTs in the C-metaschema.

On the other hand, the semantic-type collection *Manufactured Object* in the C-partition, containing **Manufactured Object**, **Medical Device**, **Research Device**, and **Clinical Drug**, is split into two groups in the Q-partition. One group is *Clinical Drug*, containing only **Clinical Drug**; the other group is *Manufactured Object*, consisting of the remainder of the three semantic types. This is because in the C-partition, the leaf singleton **Clinical Drug** is merged into the group of its parent semantic type **Manufactured Object**, while in the Q-partition, there is no rule to avoid leaf singletons.

From the above comparison, it is clear that the C-metaschema generally provides a more refined abstract view of the ESN than the Q-metaschema. The collections that are similar in the two metaschemas, up to the refinement level, cover 92 semantic types (i.e., 66.7% of the ESN).

There are 22 semantic types of the ESN that are assigned to MSTs differently in the two metaschemas. The MSTs involved are *Entity*, *Conceptual Entity*, *Molecular Sequence*, and *Geographic Area* in the Q-metaschema and *Entity*, *Idea or Concept*, and *Substance* in the C-metaschema. There are also cases where MSTs with the same name in the two metaschemas represent different semantic-type collections in the underlying partitions. For example, *Entity* appears in both metaschemas, but it represents different semantic-type collections in each. Please note that the major differences in the two metaschemas involve only 15.9% of the ESN.

An interesting measure for the two metaschemas is how many semantic relationships of the ESN are not reflected by the meta-relationships of the metaschema. There are 422 defined semantic relationships in the ESN, but when the inherited semantic relationships are taken into account, the number is 7,303. For the Q-metaschema, there are 699 out of the 7,303 semantic relationships (about 9.6%) that are not reflected. For the C-metaschema, there are only 285 out of the 7,303 semantic relationships (about 4%) that are not reflected. Hence, the C-metaschema is better at capturing the relationship structure of the ESN. The reason for this is not just the larger number of MSTs; it is also due to the fact that the initial design of the collections is based on the grouping of all semantic types with the same set of relationships. This organization minimizes the cases of relationships introduced at a non-root semantic type of a collection. Furthermore, all 285 semantic relationships that are not reflected by the C-metaschema are defined at leaves and are not inherited. This is not the case for the Q-metaschema.

Although the Q-metaschema captures less ESN's semantic relationships than the

C-metaschema, it contains less MSTs. Therefore, its network is more compact and simpler than that of the C-metaschema. Hence, the whole Q-metaschema with all its meta-relationships can be displayed on one page. To summarize, both metaschemas have their advantages and disadvantages and each can serve as a compact abstraction of the ESN.

Note that conceptually there is loss of knowledge in a metaschema view versus the complete ESN diagram. The loss occurs both in the nodes and in the links. In the nodes, only the roots of the collections are appearing and represent the rest of the semantic types. In the links, only the meta-relationships were presented, standing for the semantic relationships defined at the roots of the semantic-type collections. Hence, semantic relationships whose sources are non-root semantic types were missed. Furthermore, for the meta-relationships the knowledge of the target semantic type for each relationship is not reflected is the metaschema. Such knowledge loss is unavoidable in the process of capturing a large network in a compact abstract view.

However, there is no permanent loss of any knowledge as the metaschema is just the first view a user will employ when orienting herself to the ESN. The user will still have access to all the ESN's elements. Section 4.5.1 demonstrates how various partial graphical views, based on the metaschema, provide complete knowledge of small, comprehensible portions of the ESN. In particular, the fact that Figure 4.7 of the C-metaschema cannot show all the 124 meta-relationships defined for it is not so critical, as the missing meta-relationships and the semantic relationships represented by them will be displayed in the various partial views.

4.5.1 Applications of a Metaschema

In this section, three applications of a metaschema are briefly described. (These applications were described in detail in [49].)

The first application uses the metaschema notion for auditing the classification of concepts in the UMLS, where concepts of the META are assigned to one or more semantic types of the ESN. Auditing the META concept classification is a persistent, and perhaps overwhelming, task for UMLS professionals. There is a need to design auditing techniques for the UMLS which will minimize the effort and maximize the probability of finding errors.

Previously published papers have exploited UMLS knowledge to help audit the META. For example, in [8], Cimino used semantic methods to uncover UMLS classification errors. Gu et al. [19] and Bodenreider [1], respectively, described techniques to support the maintenance of the META by constructing object-oriented models of the UMLS. Hole demonstrated a new method to find missed synonymy in the META [27].

Metaschemas, too, can be used to help uncover classification errors in the META. In a metaschema, closely related semantic types are grouped into semantic-type collections and abstracted these into meta-semantic types. Since a concept may be assigned to several semantic types, it may also be associated with several meta-semantic types. However, it is more likely that a concept will be erroneously assigned to several semantic types residing in different meta-semantic types than to several semantic types of the same meta-semantic type. The reason is that, in general, two semantic types of the same meta-semantic type belong to the same domain. On the other hand, if two semantic types are in two different

meta-semantic types, they belong to two different domains. This observation leads to the idea of an audit that concentrates on concepts which are associated with several meta-semantic types. The idea is that such concepts are more likely to be in error than other concepts, and the effort to review them is limited since their number is not very large. For more details and examples, see [17].

One example is the concept SERIAL ANALYSIS OF GENE EXPRESSION that was assigned to **Plant** and **Research Activity** simultaneously. In the C-metaschema, these two semantic types belong to MSTs *Plant* and *Research Activity*, respectively. The MST *Plant* consists of semantic types residing in the **Entity** part of the ESN, while the *Research Activity* contains semantic types residing in the **Event** part. They are quite different in nature. Hence, the classification of a concept assigned to these two MSTs is suspicious. As a matter of fact, from the name of the concept, it is clear that the assignment of the concept to **Plant** is erroneous and should be removed. A typical user for this application is an NLM employee who is an auditor of the UMLS concept classifications. Such a person can utilize the metaschema to help in detecting classification errors.

The second application is using various kinds of graphical views, based on the metaschema, to enhance user orientation to the ESN. These views include:

1. A collection subnetwork which is a subgraph of the ESN induced by a semantic-type collection (See Figure 4.8).

2. The focus MST submetaschema which contains an MST in which the user is interested (a focus MST) and all its neighboring MSTs (See Figure 4.9).

3. The bi-collection subnetwork which is the subgraph of the ESN induced by two

neighboring collections (i.e., the corresponding MSTs are neighbors). (See Figure 4.10.)

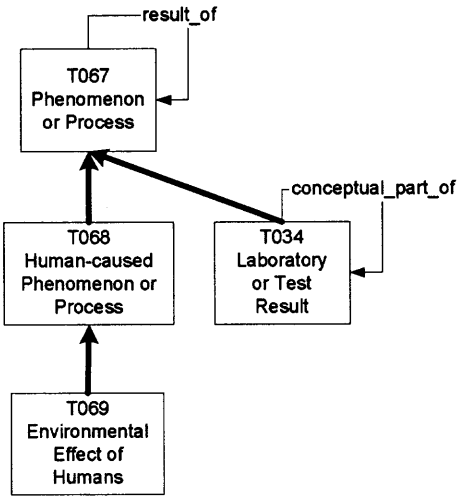


Figure 4.8 *Phenomenon or Process* collection subnetwork.

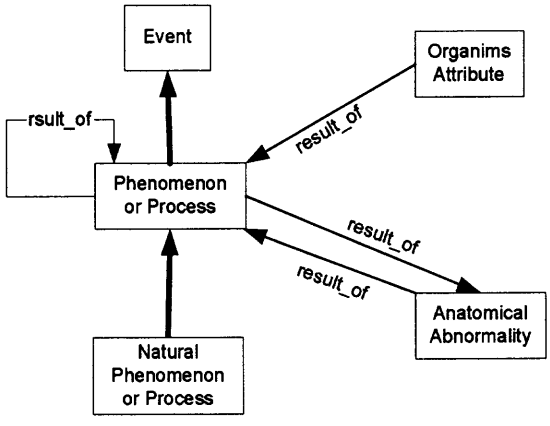


Figure 4.9 Focus *Phenomenon or Process.* submetaschema

The following example describes a scenario of a user employing these graphical views to gain an orientation. The user starts by viewing the metaschema hierarchy (Figure 4.6) to identify which MST is closest to her interest. Suppose it is *Phenomenon or Process*. Then the viewer looks at the *Phenomenon or Process* collection subnetwork (Figure 4.8), and she can see all the semantic types in the collection and all relation-

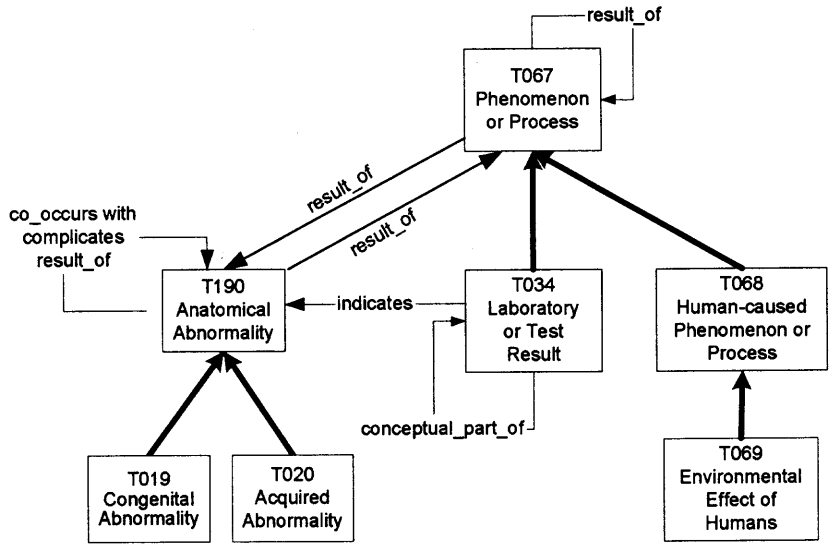


Figure 4.10 Bi-collection subnetwork of *Phenomenon or Process* and *Anatomical Abnormality*.

ships connecting them. Once the user gains this knowledge, she might want to see the interaction between semantic types of this collection and other external semantic types. But the number of relationships between semantic types of this collection and other semantic types may be overwhelming. Thus, the user can first view an abstraction of this interaction by viewing the *Phenomenon or Process* focus MST submetaschema where the relationships to and from the various neighboring MSTs of *Phenomenon or Process* are shown (Figure 4.9). If, for example, the user identifies an interest in the interaction between *Phenomenon or Process* and *Anatomical Abnormality*, she can choose to view the *Phenomenon or Process/Anatomical Abnormality* bi-collection subnetwork (Figure 4.10). The subnetwork contains all the interactions in the ESN between the semantic types of these two collections. Note that this view may show relationships from non-root semantic types of a collection which were not reflected in the metaschema, e.g., the `indicates` relationship from **Laboratory or Test Result** to **Anatomical Abnormality**. That is, the

loss of relationship knowledge in the metaschema is not a permanent loss, and the "lost relationships" appear in the bi-collection views. If the user wants to learn about all the external relationships of the *Phenomenon or Process* collection, then she can view a sequence of bi-collection subnetworks, one for each pair of neighboring MSTs in the focus MST submetaschema. In this way, the overwhelming task of reviewing all the relationship interactions of one collection is divided into a sequence of manageable tasks, supporting user comprehension efforts. A potential user for this application is a medical informatics student or professional who is not familiar with the SN of the UMLS and is trying to achieve an orientation to the SN.

For the third application, the user is an NLM employee classifying concepts of the UMLS who can use the graphical views, provided by a metaschema framework, to help detect and avoid redundant classifications within an MST. A classification of a concept to a semantic type while it has a simultaneous assignment to a descendant of the semantic type is called a redundant classification and is forbidden in the UMLS [41]. This situation can be demonstrated with regards to classifications involving chemicals and will use the *Chemical* collection subnetwork view (Fig 4.11).

As an example, consider the concept CONCENTRIN assigned to semantic types **Steroid**, **Lipid**, and **Organic Chemical**. From the *Chemical* collection subnetwork in Fig 4.11, **Organic Chemical** is the parent of **Lipid**, which in turn is the parent of **Steroid**. Therefore, the assignment of concept CONCENTRIN to **Organic Chemical** and **Lipid** is redundant since it can be inferred from the assignment to **Steroid**.

In another example, there are two concepts, FLUOR PROTECTOR and AELITEFIL,

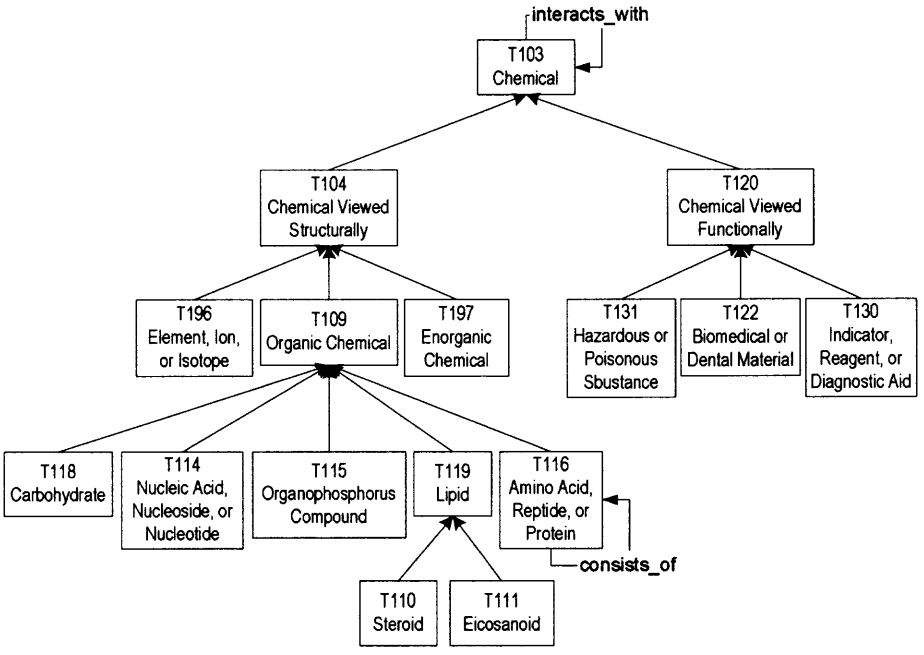


Figure 4.11 *Chemical* collection subnetwork.

assigned to three semantic types within the same subnetwork: **Chemical Viewed Structurally**, **Organic Chemical**, and **Inorganic Chemical**, where **Chemical Viewed Structurally** is the parent of the other two semantic types. Hence, the assignment of the two concepts to **Chemical Viewed Structurally** is redundant. Furthermore, a concept cannot be both an organic chemical and an inorganic chemical simultaneously. As a matter of fact, the two concepts are organic chemicals.

The following statistics demonstrate that such users might need the help of graphical views in determining concept classifications. In [19], while reviewing all intersections of semantic types in the SN of the 1998 version of the UMLS, it was discovered that 8,622 concepts had redundant classifications. This group of redundant classifications was reported to the NLM so they could be omitted in subsequent releases. Recently, a follow-up audit was performed on the 2001 UMLS to determine the status of these 8,622 concepts.

It was found that a portion (38.3%) of the redundant classifications was properly removed. However, a large number of them (57%) were still present. A third portion (4.7%) of the redundant classifications was partially treated. For instance, an existing redundant classification was removed, and a new assignment to another semantic type was added instead, only to create a new redundancy. The graphical views provided by a metaschema framework might help such users in concept classification, especially in avoiding the creation of new redundant classifications while removing existing redundant classifications.

4.6 Summary

The UMLS's Semantic Network (SN) provides an abstract view for its Metathesaurus and helps with its comprehension. However, the SN itself can be hard to comprehend since it is complex and large. At the same time, the SN does not allow for multiple parents and multiple inheritance. The ESN with its DAG structure in Chapter 2, enabling multiple parents, is more accurate but also more complex than the SN. In this chapter, the author presented the requirements for and derivation of metaschemas that support the comprehension of the ESN. The "qualified metaschema" (Q-metaschema) based on the qualified partition (Q-partition) and the "cohesive metaschema" (C-metaschema) based on the cohesive partition (C-partition) were obtained. The two metaschemas and their underlying partitions were compared. The Q-metaschema is a more compact metaschema, and the C-metaschema is more refined. Each metaschema can be used as a compact abstract layer of the ESN to help in its comprehension. Potential applications of metaschemas were described.

In [16, 35, 36] techniques to design an upper-level schema for the MED terminology

were developed. Similar techniques can be applied to other medical terminologies such as the SNOMED-CT to abstract its huge concept hierarchy into a schema of classes or groups of structural similar concepts. The role of this schema for the given terminology is similar to the role of the Semantic Network for the META of the UMLS. The technique presented in this chapter can then be applied to derive a simplified metaschema to serve as a higher-level compact view of the schema and indirectly of the concept hierarchy. The metaschema can be used as the first view presented to users to help in their orientation.

CHAPTER 5

METASCHEMAS FOR THE SN

5.1 Introduction

While the SN is an important abstraction of the META, it is still a difficult source to employ for orientation purposes due to its extensive content. To give an idea of the SN's complexity, its **Event** subnetwork is shown in Fig 5.1. Note that the figure displays neither the incoming relationships from semantic types out of the scope of the figure (i.e., from the **Entity** side) nor the inherited relationships of the semantic types. A circle with a question mark inside denotes a semantic type that is a relationship target in the **Entity** part. This figure clearly demonstrates the need to provide comprehensible access to the SN through simpler and more compact views to help user orientation. In previous work [49], the notion of metaschema was introduced, which is a higher-level network derived from a partition of the SN [6]. A metaschema serves as an abstraction of the SN. As shown in [49], a metaschema offers various compact (partial) views that can help users in their orientation to the SN. Additional applications were described in [49, 17]. In Chapter 4, the notion of metaschema was extended to encompass a directed acyclic graph (DAG) semantic network. Two metaschemas were obtained for the DAG-structured Enriched Semantic Network (ESN) in Chapter 4.

In this chapter, a new kind of lexical partitioning technique is presented based on string matching from definitions of semantic types to the names of their parents. In this technique, a child and parent that are "lexically related" will be grouped together in the same element of the lexical partition. A metaschema, called the *lexical metaschema*, based

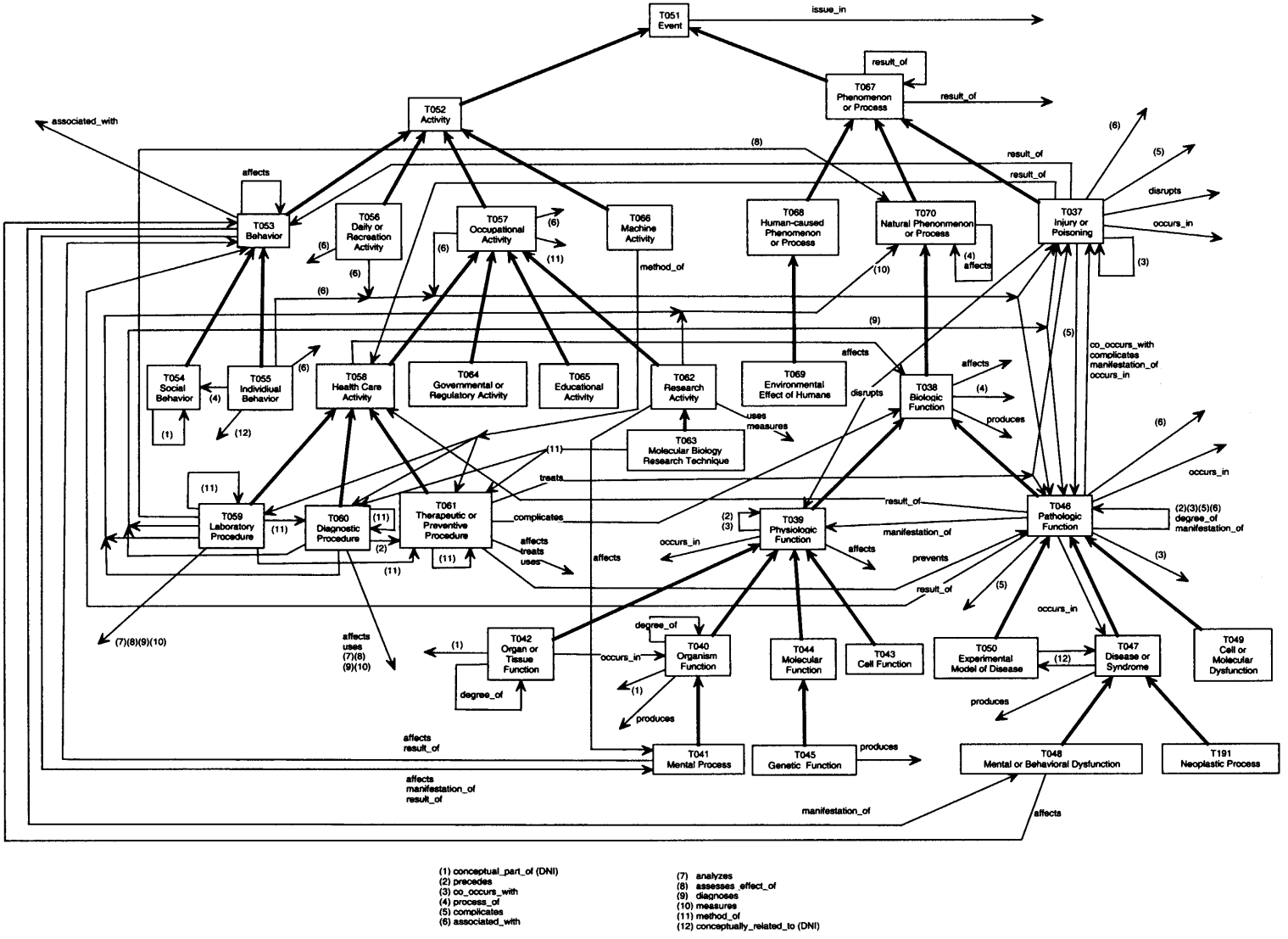


Figure 5.1 Event subnetwork of the SN.

on the lexical partition is then derived.

To evaluate the quality of the lexical metaschema, it is compared to a collection of "cumulative metaschemas" derived from a group of UMLS experts. These cumulative metaschemas are obtained from the application by experts of a manual partitioning procedure defined for them. Each cumulative metaschema represents a different level of aggregation of the experts' responses.

The notion of metaschema was introduced in [49] as an abstraction of the SN. A metaschema is based on a connected partition of the SN where the SN's IS-A hierarchy is partitioned into disjoint *semantic-type groups*. A partition is said to be *connected* if each of its semantic-type groups satisfies the condition that its semantic types together with their respective IS-A links constitute a connected subgraph of the SN with a unique root. Additionally, while a semantic-type group can be a singleton (i.e., can contain only one semantic type), that semantic type cannot be a leaf in the SN's hierarchy. This condition is imposed because the metaschema should manifest some size reduction, which singletons do not contribute to. However, a singleton containing a non-leaf semantic type is allowed, since it may express an important internal branching point in the metaschema.

In a metaschema, each semantic-type group of the partition is represented by a single node, called a *meta-semantic type*. Two kinds of relationships connect meta-semantic types. The hierarchical *meta-child-of* relationships between meta-semantic types are derived as abstractions of the SN's IS-A links. The non-hierarchical relationships, called *meta-relationships*, are derived from the SN's semantic relationships. Details of these derivations were presented in [49, 67], and a summary appears in Section 5.2.2.

For example, the hierarchy of the **Event** portion could be partitioned into five semantic-type groups as in Fig 5.2. Each semantic-type group is represented by a meta-semantic type in the corresponding metaschema. A meta-semantic type PHENOMENON OR PROCESS[1] is defined to represent the semantic-type group rooted at **Phenomenon or Process** in Fig 5.2. The metaschema hierarchy derived from the partition of the **Event** portion is shown in Fig 5.3.

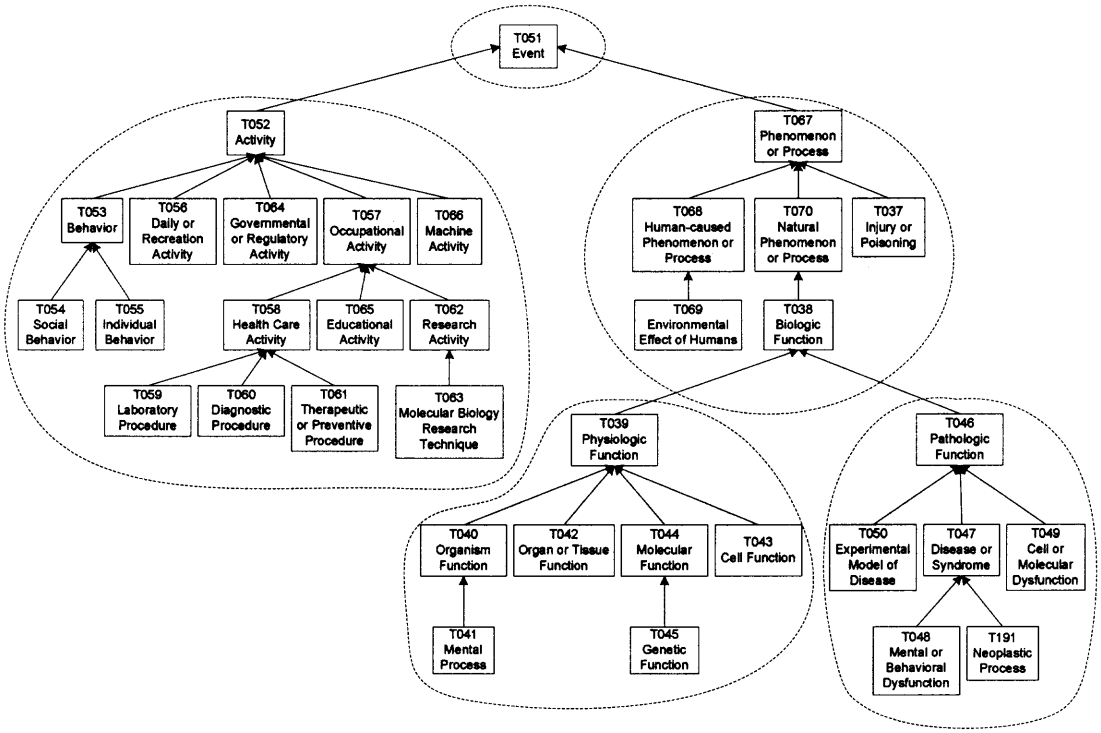


Figure 5.2 Partition example for **Event** portion of the SN.

Overall, a diagram of a metaschema serves as a good visualization mechanism of the SN and, in turn, the META, and helps in the navigation of the UMLS knowledge. In [49] various partial graphical views of groups of semantic types supported by the metaschema paradigm were introduced. These views can help in orientation of a user to the full scope of the SN's semantic relationships. In addition to the notion of metaschema, other previous

[1]Meta-semantic types will be written in "small caps" style in this chapter.

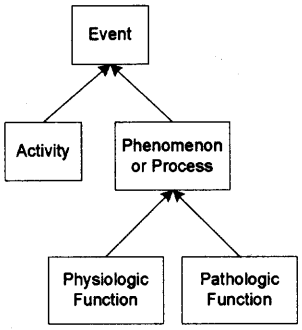


Figure 5.3 Metaschema hierarchy of the partition of **Event** portion.

work has focused on different methods to facilitate UMLS knowledge comprehension and visualization. Bodenreider and McCray described how to use visualization of semantic relationships as important indicators to explore coherence of semantic groups and help in auditing and validating the SN [2]. In [46], Nelson and Sherertz, *et al.*, presented the Hypercard browser MetaCard to enable users to extend the browsing process from META to a variety of different knowledge sources. In [33], knowledge exploration tools using levels of indentation to represent items standing in hierarchical relationships were used for displaying biomedical hierarchies in environments such as Protégé-2000. A review of knowledge visualization and navigation in the medical domain was presented by Tuttle *et al.* in [57].

5.2 Methods

This section first introduces the lexical partitioning technique for generating a lexical partition. Then it describes how to derive the lexical metaschema based on the lexical partition. After that, the instructions given to the various UMLS experts are described and the derivation of the cumulative metaschemas from their responses is presented. Finally, the evaluation techniques used to judge the quality of the lexical metaschema compared to the

experts' responses are introduced.

5.2.1 A Lexical Partitioning Technique Based on String Matching

The lexical partitioning technique is based on string matches among pairs of child and parent semantic types. The notion of "string match" is defined in the following, where the term "child/parent pair" ("CP-pair" for short) is used to denote a pair of semantic types ($\mathbf{T}_1$, $\mathbf{T}_2$) such that $\mathbf{T}_1$ is a child of $\mathbf{T}_2$ in the SN.

In Section 2.2.2, it was observed that some child semantic types referred to the names or part of the names of their respective parents in their definitions. Term "string match" was defined to describe such a case. For example, the definition of **Plant** contains the word "organism" which happens to be the name of its parent **Organism**. Hence, there is a string match (**Plant**; **Organism**; "organism"). On the other hand, there is no string match from **Biologically Active Substance** to its parent **Chemical Viewed Functionally**. The string match between a child semantic type and its parent semantic type reflects the lexical relationship in this CP-pair. From this overlapping word usage, the following definition is given:

Definition (Lexically related): A CP-pair ($\mathbf{T}_1$, $\mathbf{T}_2$) is said to be lexically related if there exists a string match between $\mathbf{T}_1$ and $\mathbf{T}_2$. □

For example, the CP-pair (**Plant**, **Organism**) is lexically related, while (**Biologically Active Substance**, **Chemical Viewed Functionally**) is not lexically related. The child in a CP-pair that is not lexically related is called *lexically independent*. The two roots of the SN, **Entity** and **Event**, are by definition lexically independent since they do not have parents.

In order for a metaschema to help users in their orientation to the SN, its associ-

ated partition must have semantic-type groups that capture various subject areas within the medical field. An underlying assumption of the lexical partitioning technique is that if a CP-pair is lexically related, then both semantic types belong to the same subject area and should therefore be in the same semantic-type group. If, on the other hand, a CP-pair is not lexically related, then the child can be seen as a transition to a new (although related) subject area. The child in this case will be made the root of a new semantic-type group in the lexical partition. Its own lexically related children and, in turn, all their lexically related children, etc., will be part of this semantic-type group, too.

For example, **Biologically Active Substance** will start a new subject area and thus will be the root of a new semantic-type group. In contrast, **Plant** is deemed to be in the same subject area as **Organism**, and thus resides in the same semantic-type group.

To construct the lexical partition, it is necessary to identify all lexically related CP-pairs. That is, it's required to check if string matches exist for the 133 CP-pairs in the SN. In the following, the partitioning process will be described as a series of four steps.

Step 1: Apply the string match method presented in Section 2.2.2 to identify all string matches in all CP-pairs of the SN;

Step 2: For each CP-pair, if there exists a string match, mark the CP-pair as "lexically related"; otherwise, mark it as "lexically unrelated."

Step 3: For each lexically unrelated CP-pair: if the child is not a leaf, then the child marks the root of a new semantic-type group in the partition; otherwise, the child is assigned to the same semantic-type group as its parent. For each lexically related CP-pair: the child is assigned to the same semantic-type group as its parent.

Please note that even though "lexically related" is not transitive, the following is a consequence of the rules. If (**B**, **A**) is a lexically related CP-pair, then **B** is assigned to the same semantic-type group as **A**. Meanwhile, if (**C**, **B**) is a lexically related CP-pair, then **C** is assigned to the same semantic-type group as **B**. Therefore, **A**, **B**, and **C** will be in the same semantic-type group in the lexical partition.

5.2.2 Metaschema Derivation

With the lexical partition in place, the lexical metaschema can be derived. A metaschema comprises three kinds of components: meta-semantic types, *meta-child-of* relationships, and meta-relationships. These are defined below along with their derivations.

A *meta-semantic type* is a node defined to represent a single semantic-type group. It is given the name of the semantic-type group's root. The root of the semantic-type group is also called the root of the meta-semantic type. The size of a meta-semantic type is the number of semantic types in the group it represents.

A *meta-child-of* relationship ("*meta-child-of*" for short) is a link between two meta-semantic types representing the IS-A relationships between the two corresponding semantic-type groups. More specifically, let **A** and $\mathbf{B}_r$ be semantic types in the semantic-type groups of meta-semantic types A and B, respectively. Furthermore, let $\mathbf{B}_r$ be the root of B and $\mathbf{B}_r$ IS-A **A**. Then in the metaschema, there exists a *meta-child-of* directed from B to A. Note that the semantic type **A** does not need to be the root of its meta-semantic type. Only the source $\mathbf{B}_r$ has to be a root in order for a new *meta-child-of* to be induced in the metaschema. The derivation of the *meta-child-of* links is motivated in detail in [49].

A *meta-relationship* is a link between two meta-semantic types representing a spe-

cific semantic relationship (non-IS-A relationship) between the two corresponding semantic-type groups. Specifically, let $\mathbf{A}_r$ and **B** be semantic types in the semantic-type groups of meta-semantic types A and B, respectively. Furthermore, let $\mathbf{A}_r$ be the root of A and let there exist a semantic relationship *rel* from $\mathbf{A}_r$ to **B**. Then in the metaschema, there exists a meta-relationship `rel`[2] directed from A to B. Note that the semantic type **B** does not need to be the root of its meta-semantic type. Only the source of the relationship *rel* (i.e., $\mathbf{A}_r$) has to be a root in order for a new meta-relationship *rel* to be induced in the metaschema. The derivation of meta-relationships is also motivated in [49].

5.2.3 Evaluation Techniques

5.2.3.1 Cumulative Metaschemas Based on Experts' Responses

An important assumption underlying the construction of the lexical metaschema is that even though the lexical partition is the result of an algorithmic process using string matching, it still effectively yields subject areas of the SN similar to those an expert might choose. A study to evaluate the validity of this assumption was conducted. A group of experts with reputations in UMLS research are selected and sent two pages with diagrams of the SN's IS-A hierarchy, i.e., the two trees rooted at **Event** and **Entity**. Each participant received a page of instructions as follows:

1 Start marking by a star the root node of the tree and continue to scan the semantic types downwards.

2 While scanning, mark by a star semantic types, which you judge as IMPORTANT AND QUITE DIFFERENT from their parent semantic types.

[2]Meta-relationships will be written in a courier font in this chapter.

3 There is one exception: Don't mark semantic types which have no children. Thus, you only need to consider the 45 semantic types with children.

4 The star markings of each participant will be used to define a metaschema where each semantic type marked by a participant names a meta-semantic-type. The metaschema will be compared with the results of other respondents and with the algorithmically derived metaschema.

The instructions heavily utilize the one-to-one correspondence between the semantic-type groups underlying the meta-semantic types, and their root semantic types. By selecting a set of semantic types that are "important and quite different" from their parents, a participating expert induces a partition of the SN and a corresponding metaschema. Each such metaschema is referred to as an "expert metaschema."

Note that although the instructions seem quite elaborate, they only define structural limitations, such as "don't mark semantic types which have no children." These limitations are necessary to make the computation of a valid comparison score between the metaschemas of the participants and the algorithmically obtained lexical metaschema possible. On the other hand, the instructions do not limit the semantic decisions of the participants, who still have the complete freedom to mark semantic types of their choice. The definitions of the semantic types were not provided with the partitoning, but they were available at the NLM Website. Most participants probably relied on their understanding of the semantic types derived from the types' names and positions in the SN's IS-A hierarchy.

It is quite interesting in quantifying the variability of the experts' responses. Towards this, the X-by-X agreement matrix (assuming X participating experts) between

participants is computed to examine the agreement between any two experts. In the agreement matrix, the number in row i and column j indicates how many meta-semantic types participant i and participant j agree on.

It is to be expected that some choices will be repeated by many participating experts. For the study, what is more interesting are metaschemas that represent a kind of aggregation of the experts' responses rather than the expert metaschemas of the individuals. In particular, a sequence of cumulative metaschemas are constructed, each of which reflects a specific level of aggregation of the experts. Suppose there are X experts' responses. A threshold value N is defined in the range (1, X) to represent the level of aggregation. The cumulative metaschema for a given N is constructed as follows. For each semantic type marked by at least N participating experts, a meta-semantic type is defined and given the name of the semantic type. Then *meta-child-of*'s and meta-relationships are derived as described in Section 5.2.2. The cumulative metaschema with the threshold value representing a simple majority [28] of the experts (i.e., $N = \lceil X/2 \rceil$) is referred to as the *consensus metaschema.*

5.2.3.2 Analysis Approach

As noted, the assumption is that the lexical technique can help to capture subject areas of the SN similar to those derived by domain experts. Therefore, it is necessary to evaluate to what degree the lexical metaschema is similar to each expert's choice, and to what degree the lexical metaschema is similar to each cumulative metaschema. In particular, it is good to know how similar the lexical metaschema is to the consensus metaschema representing the simple majority of experts.

A gold standard is created based on the majority vote of the X participating experts (in the study, $X = 11$) on the 45 candidate non-leaf semantic types. To assess the reliability of the gold standard generated by the X experts, Cronbach's α [13] is calculated, which should ideally be greater than or equal to 0.7.

Performance of the experts is calculated using a gold standard composed of the other $X - 1$ (10 in the study) experts. The majority vote with a random decision for ties is used. The agreement between the algorithmic lexical approach and each expert's choice is obtained to show the similarity between the lexical metaschema and each expert metaschema.

Performance of the algorithmic lexical approach is measured in terms of accuracy, sensitivity (recall R), specificity, precision (P), receiver operating characteristic curve trapezoidal area [25], and Rijsbergen's F measure with equal weighting of recall and precision [52]:

$$F = 2PR/(P + R)$$

The performance of the lexical algorithm to the average performance of the experts is computed [29], and confidence intervals and p-values are also calculated using bootstrap [14] estimates of variance.

To verify that majority vote rather than another threshold (e.g., 8 out of 11 experts) should have been used to define the consensus metaschema, the performance of the algorithm for different values of N is also assessed. P, R, and Rijsbergen's F measure of the lexical metaschema are calculated relative to the corresponding cumulative metaschema, using N as an independent variable. The F measure, dependent symmetrically on P and

R, is used as a typical benchmark to evaluate the similarity between the lexical metaschema and the cumulative metaschemas.

5.3 Results

5.3.1 Lexical Metaschema

Applying the "AllMatches" algorithm to the 133 CP-pairs results in string matches involving 88 CP-pairs. The string match (**Plant**; **Organism**; "organism") is one of them. Hence, about 70% of the children in the SN refer to the name or part of the name of their respective parents in their definitions. Therefore, there are 88 lexically related CP-pairs and 45 that are not lexically related.

In total, there are 47 lexically independent semantic types (including **Entity** and **Event**), among which 21 are non-leaf semantic types, and 26 are leaves. For example, (**Organism**, **Physical Object**) is not lexically related, and **Organism** is a non-leaf semantic type. (**Human**, **Mammal**) is not lexically related either, but **Human** is a leaf. Table 5.1 displays all 47 lexically independent semantic types.

Step 4 of Section 5.2.1 yields 21 semantic-type groups for the lexical partition. Each of the 21 non-leaf, lexically independent semantic types starts a new semantic-type group. Each of the 26 lexically independent leaves is assigned to the group of its respective parent. The constituent semantic types of each of the 88 lexically related CP-pairs are assigned to the same groups.

For example, **Organism** is a non-leaf, lexically independent semantic type; its child **Archaeon** is a lexically independent leaf; and the CP-pair (**Organism**, **Plant**) is lexically

Table 5.1 47 Lexically Independent Semantic Types

Non-leaves	Leaves
T001 Organism T017 Anatomical Structure T032 Organism Attribute T033 Finding T039 Physiologic Function T046 Pathologic Function T051 Event T052 Activity T067 Phenomenon or Process T071 Entity T072 Physical Object T078 Idea or Concept T082 Spatial Concept T085 Molecular Sequence T090 Occupation or Discipline T092 Organization T109 Organic Chemical T119 Lipid T121 Pharmacologic Substance T123 Biologically Active Substance T167 Substance	T016 Human T022 Body System T024 Tissue T026 Cell Component T034 Laboratory or Test Result T037 Injury or Poisoning T048 Mental or Behavioral Dysfunction T050 Experimental Model of Disease T059 Laboratory Procedure T060 Diagnostic Procedure T061 Therapeutic or Preventive Procedure T083 Geographic Area T110 Steroid T111 Eicosanoid T114 Nucleic Acid, Nucleoside, or Nucleotide T116 Amino Acid, Peptide, or Protein T118 Carbohydrate T125 Hormone T126 Enzyme T131 Hazardous or Poisonous Substance T171 Language T184 Sign or Symptom T185 Classification T191 Neoplastic Process T192 Receptor T194 Archaeon

related. Hence, **Organism** starts a new semantic-type group; **Archaeon** and **Plant** are also assigned to this group. The chart in Figure 5.4 shows the distribution of the numbers of semantic-type groups according to their sizes. For example, there are four semantic-type groups of size six.

In Table 5.2 each row shows a root of a semantic-type group, the group's size, and the complete list of the semantic types in the group. For example, the semantic-type group rooted at **Organism** has 17 semantic types which are listed in the first row of the table. The groups are listed according to the order of their roots in Table 5.1.

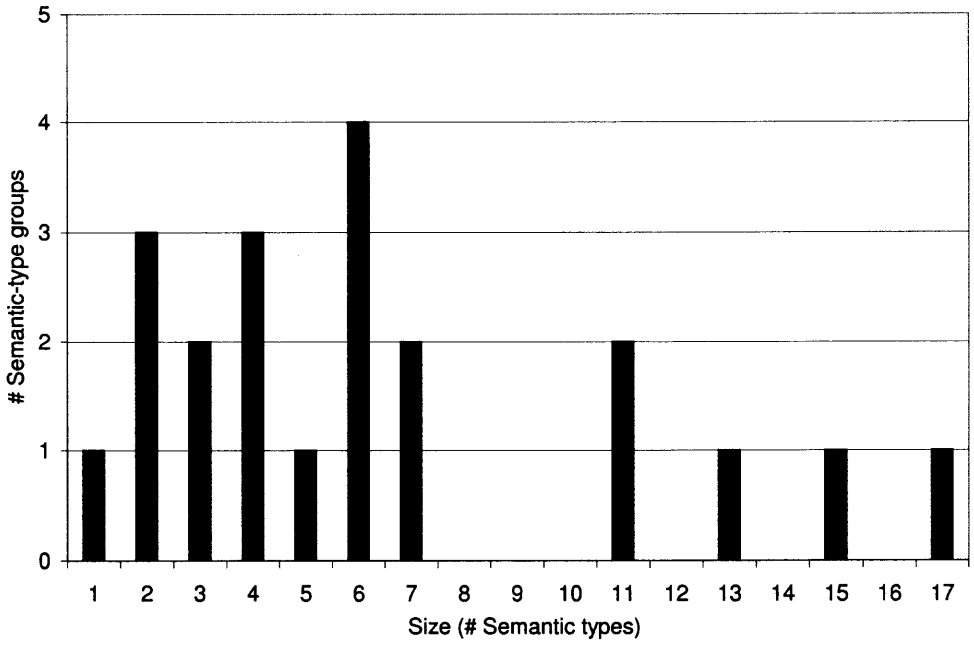


Figure 5.4 Size distribution of semantic-type groups.

With the lexical partition in place, the lexical metaschema can be derived. For example, a meta-semantic type PATHOLOGIC FUNCTION is defined to represent the semantic-type group rooted at **Pathologic Function**. PATHOLOGIC FUNCTION has six constituent semantic types. **Pathologic Function** is the root of PATHOLOGIC FUNCTION, and **Biologic Function** is in the group represented by PHENOMENON OR PROCESS. Since **Pathologic Function** IS-A **Biologic Function** in the SN, there exists a *meta-child-of* directed from PATHOLOGIC FUNCTION to PHENOMENON OR PROCESS. Meanwhile, the four semantic relationships, `co_occurs_with`, `complicates`, `manifestation_of`, and `occurs_in`, are defined from **Pathologic Function**, the root of PATHOLOGIC FUNCTION, to **Injury or Poisoning**, which is in PHENOMENON OR PROCESS. Therefore, four meta-relationships, `co_occurs_with`, `complicates`, `manifestation_of`, and `occurs_in`, are defined from PATHOLOGIC FUNCTION to PHENOMENON OR PROCESS.

The hierarchy of the lexical metaschema is shown in Fig. 5.5. The size of a meta-

Table 5.2 Lexical Partition of the SN

Root of Semantic-Type Group	Size	Semantic Types in Group
Organism	17	Organism; Plant; Alga; Archaeon; Virus; Animal; Invertebrate; Vertebrate; Mammal; Human; Reptile; Fish; Bird; Amphibian; Bacterium; Fungus; Rickettsia or Chlamydia
Anatomical Structure	11	Anatomical Structure; Embryonic Structure; Fully Formed Anatomical Structure; Body Part, Organ, or Organ Component; Tissue; Cell; Cell Component; Anatomical Abnormality; Acquired Abnormality; Congenital Abnormality; Gene or Genome
Organism Attribute	2	Organism Attribute; Clinical Attribute
Finding	3	Finding; Sign or Symptom; Laboratory or Test Result
Physiologic Function	7	Physiologic Function; Organ or Tissue Function; Organism Function; Mental Process; Molecular Function; Genetic Function; Cell Function
Pathologic Function	6	Pathologic Function; Experimental Model of Disease; Disease or Syndrome; Neoplastic Process; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction
Event	1	Event
Activity	15	Activity; Behavior; Individual Behavior; Social Behavior; Daily or Recreational Activity; Machine Activity; Occupational Activity; Health Care Activity; Laboratory Procedure; Diagnostic Procedure; Therapeutic or Preventive Procedure; Governmental or Regulatory Activity; Educational Activity; Research Activity; Molecular Biology Research Technique
Phenomenon or Process	6	Phenomenon or Process; Human-caused Phenomenon or Process; Environmental Effect of Humans; Natural Phenomenon or Process; Biologic Function; Injury or Poisoning;
Entity	13	Entity; Conceptual Entity; Group Attribute; Language; Intellectual Product; Classification; Regulation or Law; Group; Professional or Occupation Group; Population Group; Family Group; Age Group; Patient or Disabled Group
Physical Object	6	Physical Object; Manufactured Object; Research Device; Medical Device; Medical Delivery Device; Clinical Drug;
Idea or Concept	6	Idea or Concept; Functional Concept; Body System; Temporal Concept; Qualitative Concept; Quantitative Concept
Spatial Concept	4	Spatial Concept; Geographic Area; Body Location or Region; Body Space or Junction
Molecular Sequence	4	Molecular Sequence; Amino Acid Sequence; Carbohydrate Sequence; Nucleotide Sequence
Occupation or Discipline	2	Occupation or Discipline; Biomedical Occupation or Discipline
Organization	4	Organization; Professional Society; Health Care Related Organization; Self-help or Relief Organization
Lipid	3	Lipid; Steroid; Eicosanoid
Pharmacologic Substance	2	Pharmacologic Substance; Antibiotic
Biologically Active Substance	7	Biologically Active Substance; Receptor; Vitamin; Enzyme; Hormone; Neuroreactive Substance or Biogenic Amine; Immunologic Factor
Substance	11	Substance; Body Substance; Food; Chemical; Chemical Viewed Functionally; Hazardous or Poisonous Substance; Biomedical or Dental Material; Indicator, Reagent, or Diagnostic Aid; Chemical Viewed Structurally; Inorganic Chemical; Element, Ion, or Isotope
Organic Chemical	5	Organic Chemical; Amino Acid, Peptide, or Protein; Organophosphorus Compound; Nucleic Acid, Nucleoside, or Nucleotide; Carbohydrate

semantic type is displayed in parentheses following its name. Fig. 5.6 shows the metaschema including all *meta-child-of*'s and meta-relationships. Overall, the metaschema contains 21 meta-semantic types, 19 *meta-child-of*'s, and 86 meta-relationships. The average size of a meta-semantic type is close to six.

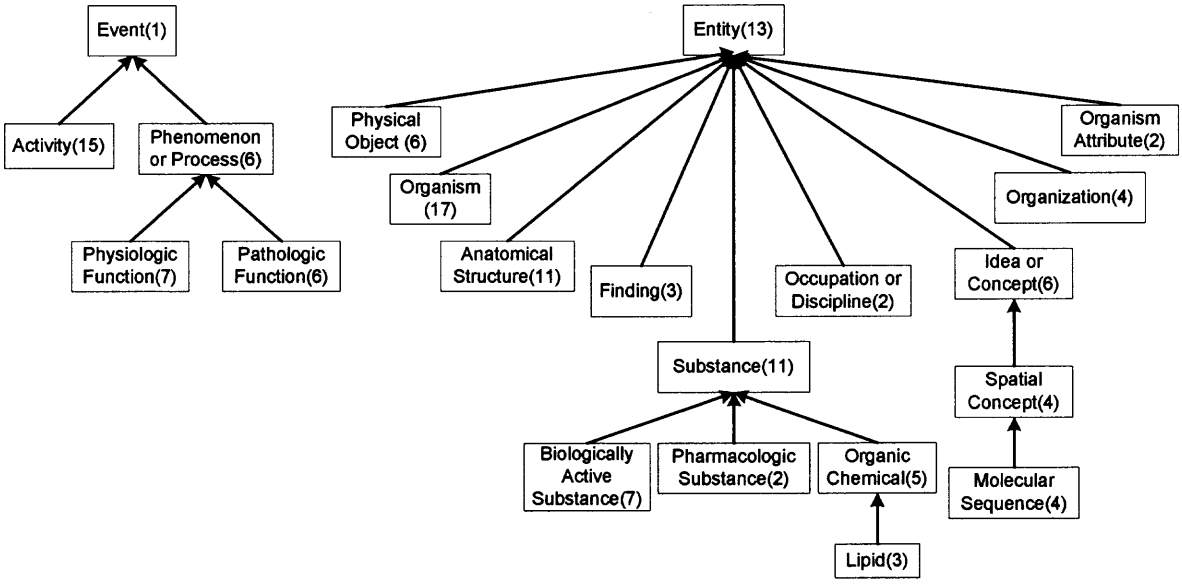


Figure 5.5 Lexical metaschema hierarchy.

5.3.2 Cumulative metaschemas

In the study, eleven responses from eleven experts ($X = 11$) were received and thus eleven cumulative metaschemas were obtained by varying N over the range (1, 11). For $N = 8$, for example, there were 16 semantic types marked by at least eight out of the eleven experts, and so the corresponding cumulative metaschema has 16 meta-semantic types. Table 5.3 shows the number of semantic types marked for each N. Obviously, the larger the value of N, the smaller the common number of meta-semantic types.

As in the table, the number of meta-semantic types varies from two (for $N = 11$)

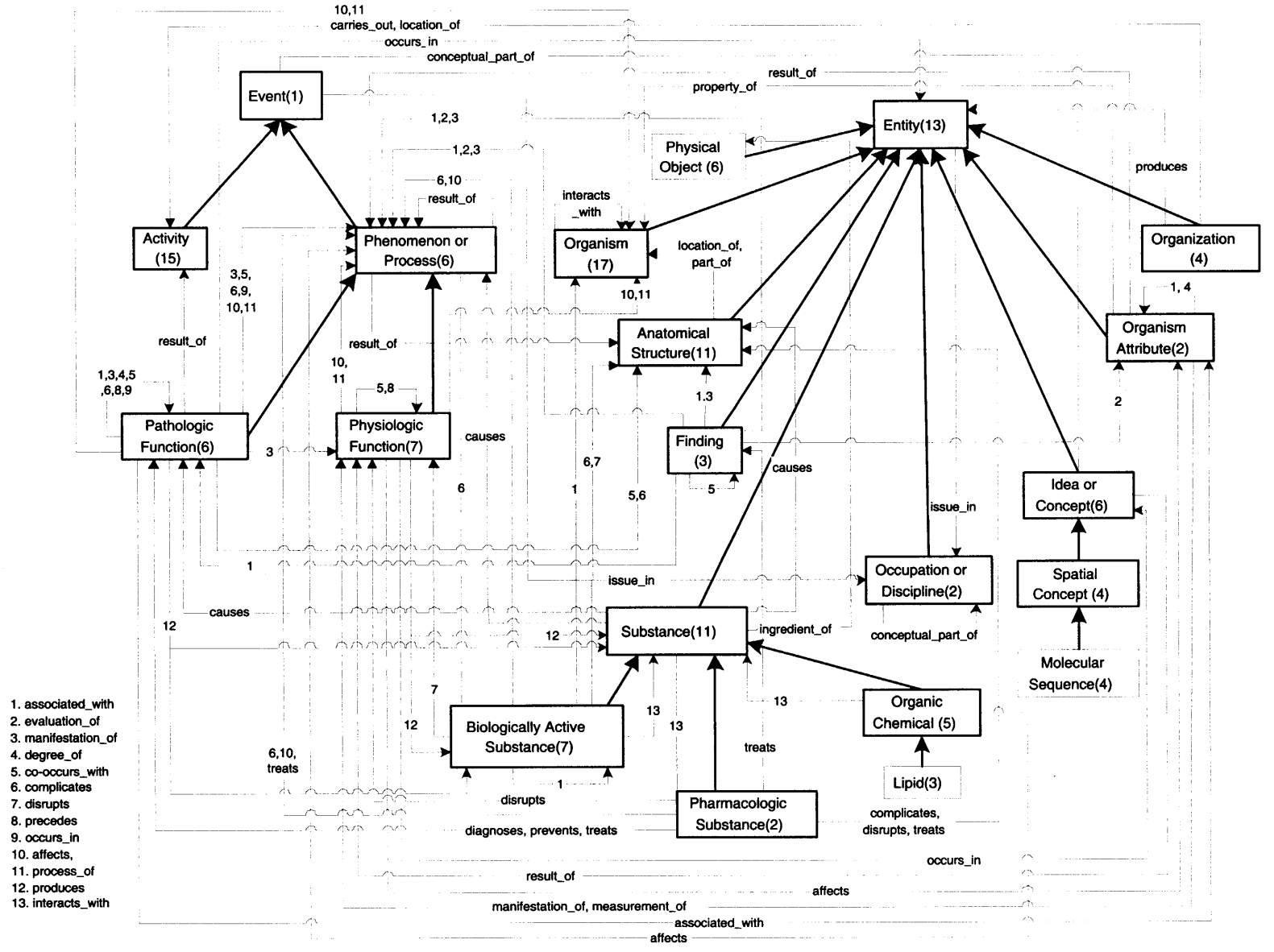


Figure 5.6 Entire lexical metaschema.

Table 5.3 Threshold Value N and Number of Semantic Types Marked (= # meta-semantic types chosen)

Threshold (N)	1	2	3	4	5	6	7	8	9	10	11
# meta-semantic types	45	45	45	42	36	26	20	16	10	7	2

to 45 (for $N = 1, 2$, and 3). The corresponding metaschema for the first case contains two meta-semantic types ENTITY and EVENT, each spanning the whole corresponding tree of the SN. For the latter cases, each non-leaf semantic type names a meta-semantic type. The metaschema that emerges in those cases is effectively just the SN itself. No real grouping of related semantic types occurs. Obviously such extreme metaschemas are not interesting. The consensus metaschema ($N = 6$) contains 26 meta-semantic types. Its hierarchy is shown in Figure 5.7.

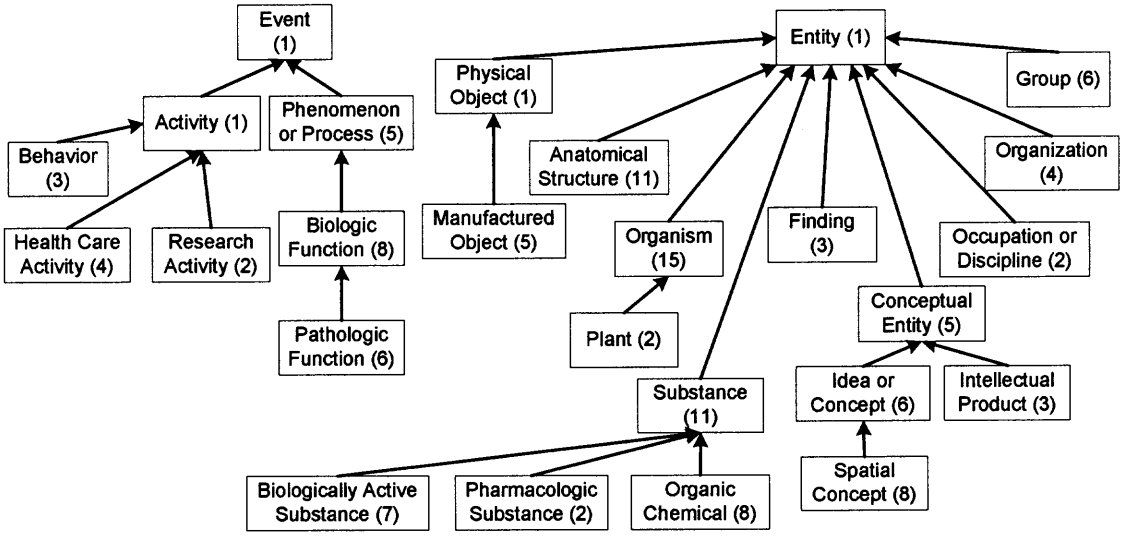


Figure 5.7 Consensus metaschema hierarchy ($N = 6$).

5.3.3 Statistical Evaluation Results

While studying responses from different experts, it is found found that individual participants' responses varied greatly both in the choice of semantic types marked and their number. First, the agreement between the lexical metaschema and each expert metaschema is calculated. For example, expert 1 chose 21 semantic types to name meta-semantic types in his expert metaschema. Among these 21 meta-semantic types, 13 also appear in the lexical metaschema. Table 5.4 shows the number of semantic types that were chosen by both a participating expert and the lexical algorithm. The second row in Table 5.4 shows the number of meta-semantic types for the participants. The number in the third row shows how many semantic types among those marked by this expert (shown in the second row) were also chosen by the lexical algorithm. For example, expert 1 marked 21 semantic types, among which 13 also appear as meta-semantic types in the lexical metaschema since they are roots of semantic-type groups in the lexical partition. The average similarity of the participants with the lexical metaschema is 13.27, with a high of 17 and low of 7. The average number of meta-semantic types marked by a participant is about 26, with minimum and maximum numbers of 12 and 36, respectively. The large variation in the numbers of the expert metaschemas' meta-semantic types raises doubts about the appropriateness of using them to evaluate the lexical metaschema and led to the consideration of aggregating their responses.

To substantiate this, the agreement matrix of all eleven experts (Table 5.5) was constructed to demonstrate the agreement as well as the high variability of participant responses. For instance, participants 2 and 5 both marked 34 semantic types and agree on 27

of them. The average inter-participant agreement is 16.76, with a high of 30 and a low of six. The large range shows the high variability of participant responses and the difficulty of using an individual expert metaschema as a measure to evaluate the lexical metaschema (or for that matter, any other algorithmically derived metaschema), as experts vary so much in their opinions. In other words, what is the sense of comparing the lexical metaschema with the expert metaschemas if each expert creates one that is so different from those of his or her peers?

Table 5.4 Algorithm-participant Similarity

Participant	1	2	3	4	5	6	7	8	9	10	11	Average
# Meta-semantic types (Expert)	21	34	21	35	34	35	25	26	12	15	36	26.73
Expert and algorithm	13	17	13	15	15	16	12	9	7	12	17	13.27

Table 5.5 Inter-participant Agreement Matrix; Average = 16.76

	1	2	3	4	5	6	7	8	9	10	11
1		19	15	16	15	19	12	11	11	12	20
2			18	28	27	27	20	19	12	14	28
3				16	16	17	14	9	10	10	18
4					28	26	23	21	8	10	30
5						27	20	20	8	10	27
6							19	22	10	14	27
7								14	8	7	24
8									6	9	18
9										9	11
10											13
11											

Cronbach's α for the gold standard was 0.62. The performance of the lexical metaschema and of the experts is shown in Table 5.6, with 95% confidence intervals in parentheses. The values (i.e., accuracy, R, etc.) measure the performance of the lexical metaschema compared to the average performance of the experts. There were no statistically significant differences between the lexical algorithm and the experts. The experiment had sufficient power to detect a difference of 0.15 in ROC area and in the F measure.

Table 5.6 Performance Comparison of Lexical Algorithm and Experts

	Lexical Algorithm	Experts
Accuracy	0.71 (0.53 to 0.84)	0.59 (0.50 to 0.66)
R	0.65 (0.44 to 0.84)	0.66 (0.57 to 0.73)
Specificity	0.79 (0.38 to 0.95)	0.51 (0.44 to 0.56)
P	0.81 (0.53 to 0.95)	0.70 (0.56 to 0.79)
ROC area	0.72 (0.58 to 0.85)	0.59 (0.52 to 0.64)
F measure	0.72 (0.53 to 0.87)	0.65 (0.53 to 0.74)

To verify that the simple majority vote, rather than another threshold, should have been used in the consensus metaschema evaluation, the performance of the algorithm for different levels of the threshold was assessed. Table 5.7 shows the results. The second column shows the number of semantic types marked (i.e., number of meta-semantic types chosen) by at least N participants. The third number is the number of semantic types marked by at least N participants that were also identified as roots of groups by the lexical metaschema. For example, the cumulative metaschema with $N = 8$ contains 16 meta-semantic types, among which eleven also appear in the lexical metaschema. Therefore, precision $P = 11/21 = 0.524$, recall $R = 11/16 = 0.688$, and $F = 0.595$.

From the plots in Figure 5.8, it is easy to see that the larger the value of N, the

Table 5.7 Performance Comparison of Lexical Metaschema for Different Values of N

Threshold (N)	Marked (B)	Lexical (C)	P= C/21	R = C/B	F=2PR/(P+R)
11	2	2	0.095	1.000	0.174
10	7	5	0.238	0.714	0.357
9	10	8	0.381	0.800	0.516
8	16	11	0.524	0.688	0.595
7	20	13	0.619	0.650	0.634
6	26	17	0.810	0.654	0.723
5	36	20	0.952	0.556	0.702
4	42	21	1.000	0.500	0.667
3	45	21	1.000	0.467	0.636
2	45	21	1.000	0.467	0.636
1	45	21	1.000	0.467	0.636

smaller the number of semantic types marked by at least N experts, and thus the lower the precision value. Also, the smaller the value of N, the lower the recall. The F measure peaks at $N = 6$, with a high precision and a medium recall. This result indicates that the lexical metaschema is most similar to this cumulative metaschema, which, in fact, is actually the consensus metaschema representing a simple majority of the experts. Out of the 26 meta-semantic types in the consensus metaschema, 17 are also in the lexical metaschema with the recall value of 81%, indicating high similarity between the two metaschemas.

5.4 Discussion

While the value of 0.62 obtained for Cronbach's α is lower than the target of 0.7 [13], it is not unreasonable. Future studies might benefit from using, say, 15 rather than eleven experts.

Table 5.6 compares the performance of the lexical metaschema to the average ex-

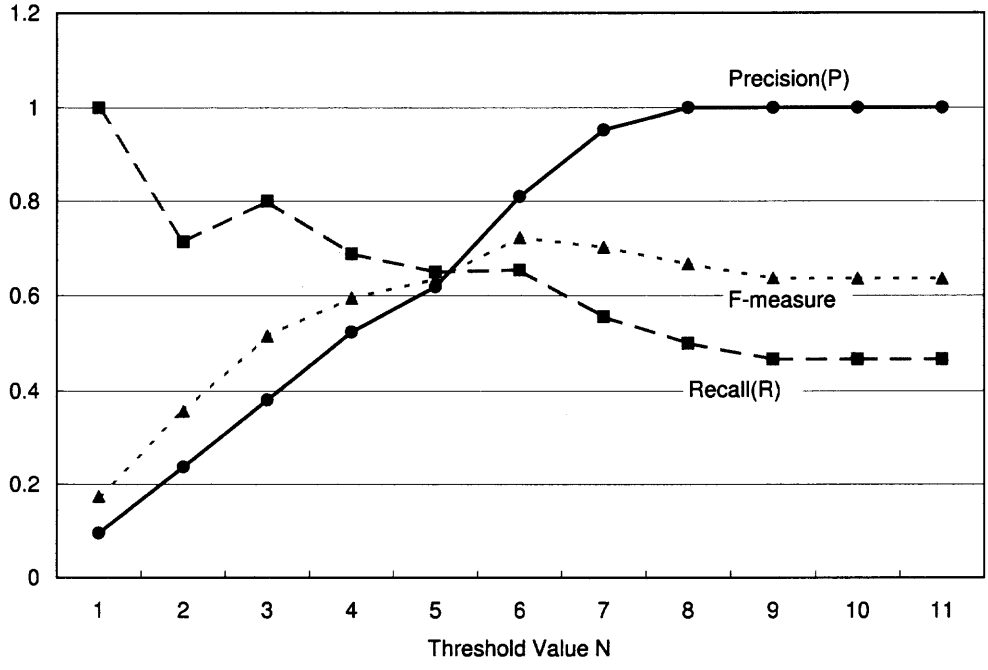


Figure 5.8 P, R, and F values for different thresholds N.

perts' performance. It shows that while there appears to be a trend of the lexical approach outperforming the experts, none of the differences were statistically significant. One can at least conclude that the algorithmic technique did not grossly underperform the experts.

The results in Table 5.7 show that the lexical metaschema is quite similar to the consensus metaschema. Thus, it can be concluded that the lexical metaschema can capture subject areas in the SN similar to the ones picked by a simple majority of the experts. While most of the results bear this out, some do not. Consider, for example, **Plant** in the lexically related CP-pair (**Plant**, **Organism**). As such, it is part of the meta-semantic type ORGANISM in the lexical metaschema. But in the consensus metaschema, PLANT is a separate meta-semantic type, probably due to the difference from other semantic types in the ORGANISM group.

In the comparison, only the meta-semantic types' names are considered without

taking into account the underlying semantic-type groups represented by the meta-semantic types. Although the chosen semantic types determined the whole metaschema, it is good to compare the two metaschemas in more detail. Now the lexical metaschema and the consensus metaschema are compared by their underlying semantic-type groups. To support the comparison, some definitions are given.

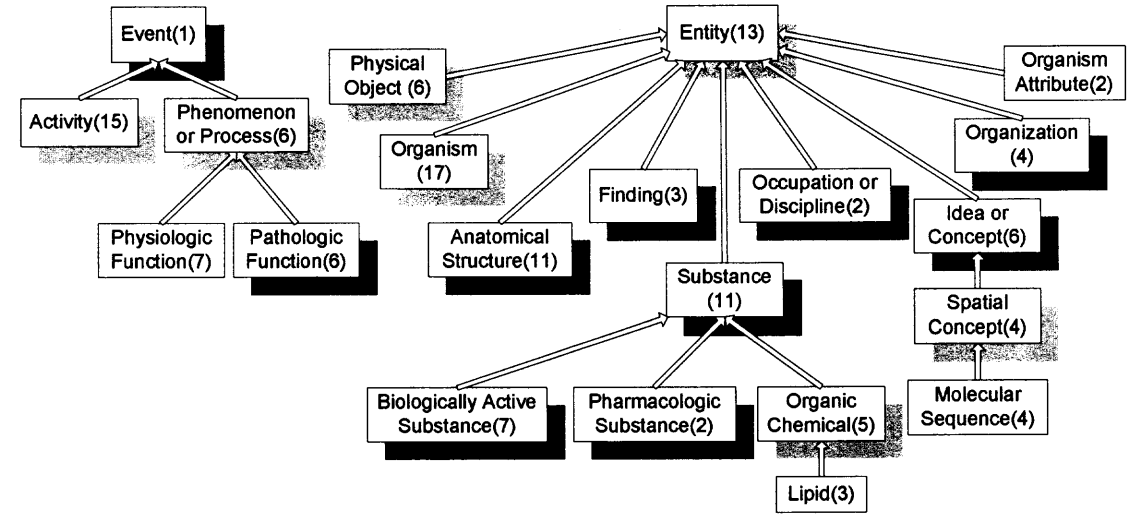
Let M_1 and M_2 be two metaschemas of the SN.

Definition (Identical): A meta-semantic type A in M_1 is *identical* to a meta-semantic type B in M_2 if both meta-semantic types have the same underlying semantic-type group. □

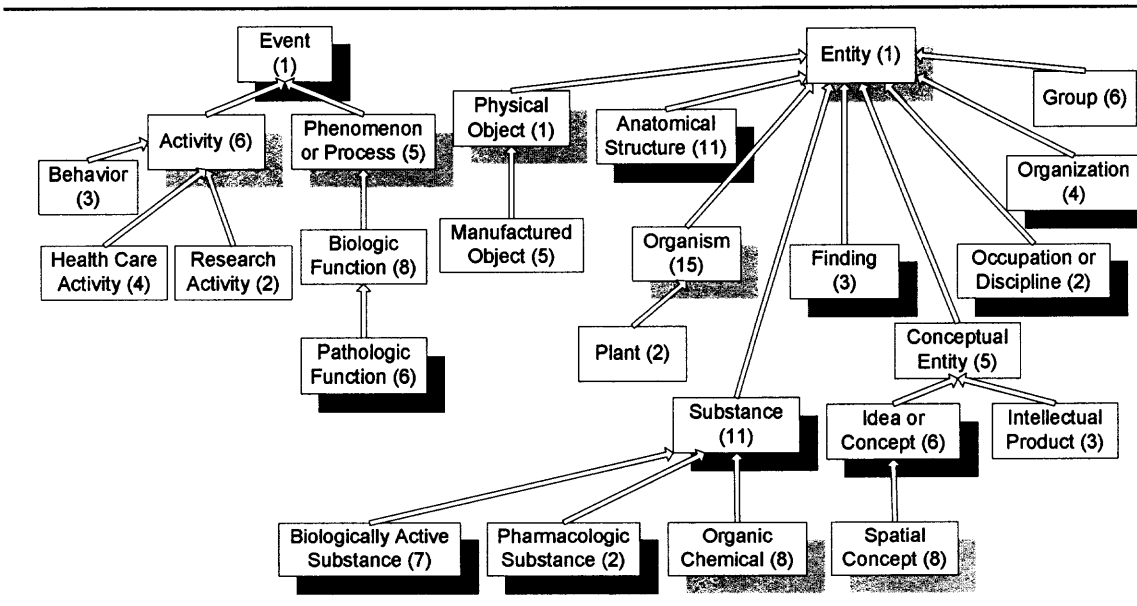
Definition (Similar): A meta-semantic type A in M_1 is *similar* to a meta-semantic type B in M_2 if the roots of their underlying semantic-type groups are the same. □

To facilitate the comparison between the lexical and consensus metaschemas, their hierarchies are shown together in Fig 5.9. Identical meta-semantic types are indicated by black shadows. Similar meta-semantic types are denoted by gray shadows.

The lexical metaschema contains 21 meta-semantic types, while the consensus metaschema contains 26 meta-semantic types. There are ten identical meta-semantic types between the two metaschemas. For example, FINDING is a meta-semantic type appearing in both metaschemas and representing the same underlying semantic-type group containing three semantic types. Therefore, FINDING in the lexical metaschema is identical to FINDING in the consensus metaschema. Table 5.8 lists all the identical meta-semantic types and their sizes. This means both metaschemas agree that these ten meta-semantic types represent important subject areas in the SN. Altogether, they cover 53 semantic types (i.e., 39.3% of the SN).



Lexical Metaschema



Consensus Metaschema

Figure 5.9 Hierarchies of lexical and consensus metaschemas.

Table 5.8 Identical Meta-semantic Types in Lexical and Consensus Metaschemas

Meta-semantic type	Size
ANATOMICAL STRUCTURE	11
BIOLOGICALLY ACTIVE SUBSTANCE	7
EVENT	1
FINDING	3
IDEA OR CONCEPT	6
OCCUPATION OR DISCIPLINE	2
ORGANIZATION	4
PATHOLOGIC FUNCTION	6
PHARMACOLOGIC SUBSTANCE	2
SUBSTANCE	11

There are seven similar meta-semantic types. For example, SPATIAL CONCEPT in the lexical metaschema represents an underlying semantic-type group with four semantic types, while SPATIAL CONCEPT in the consensus metaschema represents a semantic-type group with eight semantic types. Hence, SPATIAL CONCEPT in the lexical metaschema is similar, but not identical, to SPATIAL CONCEPT in the consensus metaschema. Table 5.9 shows these similar meta-semantic types along with their sizes in each of the two metaschemas. In the lexical metaschema, these seven cover 66 semantic types, which is about 48.9% of the SN. In the consensus metaschema, these seven cover 44 semantic types, which is about 32.6%.

To better understand the difference between pairs of similar meta-semantic types, note that in some cases the difference reflects various levels of granularity in the partition, rather than major disagreements between the metaschemas. To be more specific, it is found that some meta-semantic types in the lexical metaschema are split into several separate

Table 5.9 Similar Meta-semantic Types in Lexical and Consensus Metaschemas

Meta-semantic type	Size in lexical metaschema	Size in consensus metaschema
ACTIVITY	15	6
ENTITY	13	1
ORGANIC CHEMICAL	5	8
ORGANISM	17	15
PHENOMENON OR PROCESS	6	5
PHYSICAL OBJECT	6	1
SPATIAL CONCEPT	4	8

meta-semantic types in the consensus metaschema, or vice versa.

To be formal, "refinement" can be defined as follows. Let $G_M(\mathrm{A})$ denote the semantic-type group represented by the meta-semantic type A in the metaschema M.

Definition (Refinement): Let A be a meta-semantic type in metaschema M_1. If there exists a set of meta-semantic types $\{\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_k\}$ $(k \geq 2)$ in metaschema M_2 such that $G_{M_1}(\mathrm{A}) = \cup_{i=1}^{k} G_{M_2}(\mathrm{B}_i)$, then the set $\{\mathrm{B}_1, \mathrm{B}_2, \ldots, \mathrm{B}_k\}$ is called a *refinement* of A. □

As an example, the meta-semantic type ACTIVITY in the lexical metaschema represents a semantic-type group containing 15 semantic types. These 15 semantic types are split into four semantic-type groups represented by ACTIVITY, BEHAVIOR, HEALTH CARE ACTIVITY, and RESEARCH ACTIVITY in the consensus metaschema. Therefore, {ACTIVITY, BEHAVIOR, HEALTH CARE ACTIVITY, RESEARCH ACTIVITY} in the consensus metaschema is a refinement of ACTIVITY in the lexical metaschema.

Table 5.10 shows the cases of refinement from the lexical metaschema to the consensus metaschema. The size of a meta-semantic type is displayed in parentheses following the name. This kind of refinement covers 38 semantic types.

Table 5.10 Refinements in Consensus Metaschema

Meta-semantic type in lexical metaschema	Refinement in the consensus metaschema
ACTIVITY (15)	{ACTIVITY (6), BEHAVIOR (3), HEALTH CARE ACTIVITY (4), RESEARCH ACTIVITY (2)}
PHYSICAL OBJECT (6)	{PHYSICAL OBJECT (1), MANUFACTURED OBJECT (5)}
ORGANISM (17)	{ORGANISM (15), PLANT (2)}

Table 5.11 Refinements in Lexical Metaschema

Meta-semantic type in consensus metaschema	Refinement in the lexical metaschema
ORGANIC CHEMICAL (8)	{ORGANIC CHEMICAL (5), LIPID (3)}
SPATIAL CONCEPT (8)	{SPATIAL CONCEPT (4), MOLECULAR SEQUENCE (4)}

On the other hand, there are also some refinements in the other direction from consensus metaschema to lexical metaschema. For example, {ORGANIC CHEMICAL, LIPID} in the lexical metaschema is a refinement of ORGANIC CHEMICAL in the consensus metaschema. Table 5.11 shows all such refinement cases. This kind of refinement covers 16 semantic types.

Note that if there is a refinement case, then there is always a meta-semantic type in one metaschema that is similar to one of the meta-semantic types in the refinement. For example, {ORGANIC CHEMICAL, LIPID} in the lexical metaschema is a refinement of ORGANIC CHEMICAL in the consensus metaschema, where the ORGANIC CHEMICAL meta-semantic types in both metaschemas are similar. However, not every case of similar meta-semantic types is a refinement. For example, ENTITY and PHENOMENON OR PROCESS are both cases of similarity, but they do not have refinements. The total number of

semantic types covered by refinements in either direction is 54 (about 40%).

Besides the identical meta-semantic types, the similar meta-semantic types, and the meta-semantic types appearing in refinements, there are two meta-semantic types that appear exclusively in the lexical metaschema; these are PHYSIOLOGIC FUNCTION and ORGANISM ATTRIBUTE. There are also four meta-semantic types that appear exclusively in the consensus metaschema; these are BIOLOGIC FUNCTION, CONCEPTUAL ENTITY, INTELLECTUAL PRODUCT, and GROUP.

If, as in the previous section, only the meta-semantic type names and not the underlying semantic-type groups are considered, then 17 out of the 21 meta-semantic types in the lexical metaschema also appear in the consensus metaschema (about 81%). At the same time, the semantic types covered by identical meta-semantic types and refinements together are 107 (about 79%). Both measures show the high similarity between the two metaschemas. In other words, the lexical metaschema provides a good approximation for a partition of meaningful subject areas in the SN, when compared to the consensus metaschema capturing the aggregation of a simple majority of the human experts' opinions.

5.5 Summary

In this chapter, the lexical metaschema derived via an algorithmic lexical partitioning approach is presented. A sequence of cumulative metaschemas are also built as aggregations of the opinions of eleven UMLS experts participating in an evaluation study. Of particular interest is the consensus metaschema representing a simple majority aggregation of

the experts' opinions. The cumulative metaschemas is used to evaluate the quality of the lexical metaschema. From the evaluation, it can be concluded that the result of the lexical algorithmic approach was similar to the consensus metaschema, within the limits of the experiment. It is interesting to note that among all cumulative metaschemas, the lexical metaschema is closest to the consensus metaschema. Note that this is a coincidental result. It is not always expected.

A metaschema is a compact, abstract view of the SN. Various metaschemas are possible. In previous work [49], the cohesive metaschema derived according to purely structural considerations is presented. In that metaschema, each meta-semantic type represented a group of semantic types with the same (or almost the same) relationships. A natural question is: which of these three metaschemas, cohesive, lexical, or consensus, is better than the others in supporting user orientation to the SN? To answer this question, it is needed to find a way to measure the overall quality of a given metaschema. As can be expected, each metaschema has its advantages and disadvantages. This observation leads to a natural question: is it possible to construct a metaschema that incorporates the "good parts" of each of the above metaschemas while avoiding their pitfalls? These issues will be addressed in future research.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

In this dissertation, the UMLS's Semantic Network (SN) hierarchy was enhanced from a two-tree structure to a DAG structure by adding new IS-A links and new semantic types to accommodate multiple parents. The resulting Enriched Semantic Network contains 139 semantic types and 150 IS-A relationships. The ESN contains cases of multiple subsumption for several semantic types. The semantic relationship distribution in the ESN is more complex than that of the SN due to the new multiple-parent IS-A hierarchy. In this arrangement, relationships can be inherited from more than one source. The relationship distribution of the ESN was derived based on that of the SN. The ESN consists of a total of 7,303 semantic relationships. The concept configuration of the ESN was derived from that of the SN through a mapping function that prevents any redundant categorizations. The function ensured that a concept is only assigned to the most specialized semantic types that are appropriate. The resulting concept configuration of the ESN in total has 1,013,876 concept assignments with an average of 7,294 per semantic type, 26,950 fewer assignments than that of the SN. Compared to the SN, the ESN serves as an extended and more refined abstraction of the UMLS's META.

During the enrichment of the SN, a connected partition of the ESN comprising 19 groups was derived; each group in the partition exhibits connectivity and semantic uniformity. The previously developed metaschema notion was extended to be applicable for a network with a DAG structure such as the ESN. Based on this new partition, a "qualified

metaschema" was derived as a higher-level abstraction of the ESN. Additionally, a "cohesive metaschema" of the ESN was also derived from a partition in which semantic types with the same relationship structure were grouped together. The two metaschemas and their underlying partitions were compared. Each metaschema can be used as a compact abstract layer of the ESN to help in its comprehension.

Besides the metaschemas for the ESN, a new lexical partitioning technique was introduced for the SN. In this partition, semantic types that are lexically related were grouped in the same semantic-type group. Based on this lexical partition, the lexical metaschema of the SN was derived. A sequence of cumulative metaschemas were built as aggregations of the opinions of eleven UMLS experts participating in an evaluation study. Of particular interest is the consensus metaschema representing a simple majority aggregation of the experts' opinions. The cumulative metaschemas was used to evaluate the quality of the lexical metaschema. From the evaluation, it comes the conclusion that the result of the lexical algorithmic approach was similar to the consensus metaschema, within the limits of the experiment.

Applications of metaschemas that can help user orientation were described in this dissertation. A metaschema can be used to provide a user with a partial graph of the SN containing a specific subject area that is of interest to him. A metaschema is also helpful in studying the relationships existing between two different subject areas. Moreover, a metaschema can help in detecting the UMLS's concept categorization errors.

6.2 Future Work

In this dissertation, two methodologies were presented to enrich the SN from a two-tree structure to a DAG structure. Two possible research issues arise as a result. The first issue is: Are there any other methods to identify extra valid IS-A links omitted in the SN? The second issue is: How do we use quantitative methods to prove that the ESN is better than the SN in capturing and modeling medical knowledge more accurately? This dissertation showed that the ESN is an extension of the SN in capturing and modeling current medical knowledge. But there is no quantitative method developed for this evaluation.

Another important future work is to study new partitioning technique for the SN or the ESN. As a result, several different metaschemas will be derived. It is important to develop an efficient evaluation measurement to evaluate the quality of each metaschema. In this dissertation, primary statistical techniques were used for metaschema evaluation. Is it possible to develop more advanced evaluation techniques?

As was mentioned in this dissertation, a metaschema is a compact, abstract view of the SN. Various metaschemas were derived for the SN or the ESN based on different partitioning techniques. As expected, each metaschema has its advantages and disadvantages. Another promising future study is to construct a consolidated metaschema (or "ideal metaschema") that incorporates the "good parts" of each metaschema while avoiding their pitfalls. Hence, it is necessary to define the criteria which a metaschema needs to satisfy in order to be an "ideal metaschema." With the consolidated metaschema, it is possible to define the distance between each algorithmic metaschema and the consolidated metaschema and use the distance as an measurement to evaluate a metaschema's quality.

REFERENCES

[1] O. Bodenreider. An object-oriented model for representing semantic locality in the UMLS. *Proc. Medinfo. 2001*, 10(1):161–165, 2001.

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

[3] K. E. Campbell, S. P. Cohn, C. G. Chute, E. H. Shortliffe, and G. Rennels. Scalable methodologies for distributed development of logic-based convergent medical terminology. *JAMIA*, 37(4-5):426–439, Nov 1998.

[4] K. E. Campbell, A. K. Das, and M. Musen. A logical foundation for representation of clinical data. *JAMIA*, 1(3):218–232, May/Jun 1994.

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

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

[7] C. G. Chute. The copernican era of healthcare terminology: a re-centering of health information systems. In *Proc. 1998 AMIA Annual Symposium*, pages 68–73, 1998.

[8] J. J. Cimino. Auditing the unified medical language system with semantic methods. *JAMIA*, 5:41–51, 1998.

[9] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394–403, Nov 1998.

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

[11] J. J. Cimino and S. B. Johnson. Use of the Unified Medical Language System in patient care. *Methods of Information in Medicine*, 34(1/2):158–184, 1995.

[12] S. Dessena, A. Rossi-Mori, and E. Galeazzi. Building cross-thesauri with the support of UMLS. In C. B, editor, *MEDINFO 98*, pages 654–659. Amsterdam:IOS Press, 1998.

[13] G. Dunn. *Design and Analysis of Reliability Studies*. New York: Oxford University Press, 1989.

[14] B. Efron, R. J. Tibshirani, B. Efron, and R. J. Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.

[15] M. J. Greenacre. *Theory and Application of Correspondence Analysis*. Academic Press, London, 1984.

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

[17] H. Gu, H. Min, Y. Peng, L. Zhang, and Y. Perl. Using the metaschema to audit UMLS classification errors. In *Proc. 2002 AMIA Annual Symposium*, pages 310–314, San Antonio, TX, Nov. 2002.

[18] H. Gu, Y. Perl, G. Elhanan, H. Min, L. Zhang, and Y. Peng. Auditing concept categorizations in the UMLS. *Artificial Intelligence of Medicine*, 2004. To appear.

[19] H. Gu, Y. Perl, J. Geller, M. Halper, L. Liu, and J. J. Cimino. Representing the UMLS as an object-oriented database: modeling issues and advantages. *JAMIA*, 7(1):66–80, Jan-Feb 2000. Selected for reprint in: R. Haux and C. Kulikowski, editors, Yearbook of Medical Informatics: Digital Libraries and Medicine (International Medical Informatics Association), pp. 271-285, Schattauer, Stuttgart, Germany, 2001.

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

[21] H. Gu, Y. Perl, J. Geller, M. Halper, and M. Singh. A methodology for partitioning a vocabulary hierarchy into trees. *Artificial Intelligence in Medicine*, 15(1):77–98, Jan. 1999.

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

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

[24] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco, CA, August 2000.

[25] J. A. Hanley and B. J. McNeil. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.

[26] I. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: a survey. *IEEE Trans Visual Comput Graphics*, 6(1):24–43, 2000.

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

[28] G. Hripcsak, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson, and P. D. Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Ann Intern Med*, 122:681–688, 1995.

[29] G. Hripcsak and A. Wilcox. Reference standards, judges, comparison subjects: roles for experts in evaluating system performance. *JAMIA*, 9:1–15, 2002.

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

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

[32] S. B. Johnson, C. Friedman, J. J. Cimino, B. T. Clark, G. Hripcsak, and P. D. Clayton. Conceptual data model for a central patient database. In *Proc. the Fifteen Annu Symp Comput Appl Med Care*, pages 381–385, New York, NY, 1991.

[33] Q. Li, P. Shilane, N. F. Noy, and M. A. Musen. Ontology acquisition from on-line knowledge sources. In *Proc. 2000 AMIA Annual Symposium*, pages 497–501, Los Angeles, CA, Nov. 2000.

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

[35] L. Liu, M. Halper, J. Geller, and Y. Perl. Controlled vocabularies in OODBs: Modeling issues and implementation. *Distributed and Parallel Databases*, 7(1):37–65, 1999.

[36] L. Liu, M. Halper, J. Geller, and Y. Perl. Using OODB modeling to partition a vocabulary into structurally and semantically uniform concept groups. *IEEE TKDE*, 14(4):850–866, 2002.

[37] A. T. McCray. UMLS Semantic Network. In *Proc. Thirteenth Annual SCAMC*, pages 503–507, Washington, DC, 1989.

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

[39] A. T. McCray, A. Burgun, and O. Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. In *Proc. Medinfo 2001*, pages 171–175, London, UK, Sept. 2001. Amsterdam: IOS Press.

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

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

[42] A. T. McCray, S. Srinivasan, and A. C. Browne. Lexical methods for managing variation in biomedical terminologies. In *Proc. the 18th Annual SCAMC*, pages 235–239, Washington DC, Nov. 1994.

[43] J. Michael, J. L. V. Mejino, and C. Rosse. The role of definitions in biomedical concept representation. In B. S, editor, *Proc. 2001 AMIA Annual Symposium*, pages 463–467, Washington DC, Nov. 2001.

[44] G. Michailidis and J. D. Leeuw. Data visualization through graph drawing. *Computation Stat*, 16(3):435–450, 2001.

[45] S. J. Nelson. The semantic structure of the umls metathesaurus. In *Proc. 1992 Annu Symp Comput Appl Med Care*, pages 649–653, Baltimore, MD, Nov. 1992.

[46] S. J. Nelson, D. D. Sherertz, M. S. Tuttle, and G. Rennels. Using MetaCard: A hypercard browser for biomedical knowledge sources. In *Proc. Annu Symp Comput Appl Med Care*, pages 151–154, 1990.

[47] N. F. Noy, R. W. Fergerson, and M. A. Musen. The knowledge model of protégé 2000: combining interoperability and fexibility. In *Proc. 2th Internat Conf on Knowledge Eng Knowledge Manage(EKAW-2000)*, Juan-les-Pins France: Springer, 2000.

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

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

[50] A. Rector. Thesauri and formal classifications: Terminologies for people and machines. *Methods of Information in Medicine*, pages 501–509, 1998.

[51] A. Rector, J. Rogers, A. Roberts, and C. Wroe. Scale and context: Issues in ontologies to link health- and bio-informatics. In *Proc. 2002 AMIA Annual Symposium*, pages 642–464, San Antonio, TX, Nov. 2002.

[52] C. J. V. Rijsbergen. *Information Retrieval*. London: Butterworth, 1979.

[53] C. Rosse and J. L. V. Mejino. The Digital Anatomist Foundational Model: Principles for defining and structuring its concept domain. In *Proc. 1998 AMIA Annual Symposium*, pages 820–824, Orlando, FL, Nov. 1998.

[54] C. Rosse and J. L. V. Mejino. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *JBI special issue of structural issues in UMLS research*, 36(6):478–500, Dec. 2003.

[55] A. Rossi-Mori, F. Consorti, and E. Galeazzi. Standards to support development of terminological systems for healthcare telematics. *Methods of Information in Medicine*, pages 551–563, 1998.

[56] P. L. Schulyer, W. T. Hole, M. S. Tuttle, and D. D. Sherertz. The UMLS metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc*, 81(2):217–222, Apr. 1993.

[57] M. S. Tuttle, W. G. Cole, D. D. Sherertz, and S. J. Nelson. Navigating to knowledge. *Methods Inf Med*, 34(1-2):214–231, 1995.

[58] M. S. Tuttle, D. D. Sherertz, N. E. Olson, M. Erlbaum, L. F. Fuller, and S. J. Nelson. Using meta-1 – the 1st version of the UMLS metathesaurus. In M. RA, editor, *Proc. the Fourteen Annu Symp Comput Appl Med Care*, pages 131–135, Washington, DC, 1990.

[59] U. S. Dept. of Health and Human Services, National Institutes of Health, National Library of Medicine. Unified Medical Language System (UMLS), 2001.

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

[61] U. S. Dept. of Health and Human Services, National Institutes of Health, National Library of Medicine. Unified Medical Language System (UMLS), 2003.

[62] H. Yu, C. Freidman, A. Rhzetsky, and P. Kra. Representing genomic knowledge in the UMLS Semantic Network. In *Proc. of the 1999 AMIA Annual Symposium*, pages 181–186, Washington, DC, Nov. 1999.

[63] L. Zhang, M. Halper, Y. Perl, J. Geller, and J. J. Cimino. The relationship distribution and concept configuration of the UMLS Enriched Semantic Network. *JAMIA*, 2004. Submitted for journal publication.

[64] L. Zhang, Y. Perl, J. Geller, M. Halper, and J. J. Cimino. Enriching the structure of the UMLS Semantic Network. In *Proc. of the 2002 AMIA Annual Symposium*, pages 939–943, San Antonio, TX, Nov. 2002.

[65] L. Zhang, Y. Perl, J. Geller, M. Halper, and J. J. Cimino. An enriched UMLS Semantic Network with a multiple subsumption hierarchy. *JAMIA*, 2004. To appear.

[66] L. Zhang, Y. Perl, J. Geller, M. Halper, and G. Hripcsak. A lexical metaschema for the UMLS Semantic Network. *Artificial Intelligence of Medicine*, 2004. Submitted for journal publication.

[67] L. Zhang, Y. Perl, M. Halper, and J. Geller. Designing metaschemas for the UMLS Enriched Semantic Network. *Journal of Biomedical Informatics*, 36(6):433 – 449, Dec. 2003.