Chapter 11

A Conceptual Framework for the Biomedical Domain

Alexa T. McCray & Olivier Bodenreider
National Library of Medicine, Bethesda, Md, USA

Abstract:

Specialized domains often come with an extensive terminology, suitable for storing and exchanging information, but not necessarily for knowledge processing. Knowledge structures such as semantic networks, or ontologies, are required to explore the semantics of a domain. The UMLS project at the National Library of Medicine is a research effort to develop knowledge-based resources for the biomedical domain. The Metathesaurus is a large body of knowledge that defines and inter-relates 730,000 biomedical concepts, and the Semantic Network defines the semantic principles that apply to this domain. This chapter presents these two knowledge sources and illustrates through a research study how they can collaborate to further structure the domain. The limits of the approach are discussed.

1. INTRODUCTION

The Unified Medical Language System® (UMLS®) project at the U.S. National Library of Medicine (NLM) is a large-scale research effort to develop knowledge-based tools and resources to compensate for differences in the way in which concepts are expressed in the field of biomedicine. Since 1990, a continually evolving set of UMLS knowledge sources has been released annually to the research community for experimentation and use in a wide range of applications (Lindberg, Humphreys, & McCray, 1993). Each year the knowledge sources are expanded and enhanced both with additional content and with additional research tools. More than 1000 institutions and individuals around the world use the UMLS in research and application.

The UMLS knowledge sources are the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon. The Metathesaurus integrates vocabularies from the biomedical domain, and the Semantic Network is the network of general semantic categories, or types, to which all Metathesaurus concepts are assigned (McCray & Nelson, 1995). The heterogeneity in the nature, scope, and quality of the vocabularies that comprise the Metathesaurus makes it a particularly complex structure. A principal reason for developing the Semantic Network was to bring semantic coherence to this somewhat unwieldy structure. Together, the Metathesaurus and the Semantic Network express and classify a significant portion of the biomedical vocabulary.

The SPECIALIST lexicon and related lexical programs, which have been developed for natural language processing applications, are UMLS resources for managing the high

In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships:
an interdisciplinary perspective. Boston: Kluwer Academic Publishers;
2002. p. 181-198.

degree of linguistic variation in natural language and in the terminologies themselves (McCray, Srinivasan, & Browne, 1994). The lexicon and lexical programs, which contain and manage not only biomedical terminology, but also a good portion of the general English vocabulary, may be used together with the other UMLS knowledge sources, but they may also be used independently in natural language processing applications. The UMLS Knowledge Source Server makes all UMLS resources available over the Internet through a Web-based interface, as well as through an application programming interface (McCray, Razi, Bangalore, Browne, & Stavri, 1996).

Building and revising the UMLS knowledge sources on an annual basis is a labor intensive process. A combination of automated and semi-automated methods is used, and this is followed by human review. Since human review is subject to human error, we have developed a variety of automated techniques to validate the data, and we correct errors as they are detected. Several investigators have explicitly analyzed the correctness, completeness, and usefulness of the UMLS (Bodenreider et al., 1998; Cimino, 1998; Pisanelli, Gangemi, & Steve, 1998; Srinivasan, 1999). Cimino, for example, has developed and implemented methods to "audit" the UMLS, using a number of interesting semantically-based techniques. The comments and analyses of those who actively use the UMLS have lead to significant improvements in the knowledge sources, and inform us as we continue to extend them.

2. CONCEPTS, CATEGORIES, AND RELATIONSHIPS

2.1 UMLS Metathesaurus

The most extensive of the three UMLS knowledge sources is the Metathesaurus. The first edition in 1990 contained approximately 30,000 concepts, representing a handful of vocabularies. The eleventh edition in 2000 contains over 730,000 concepts, representing more than 1,500,000 strings in over fifty vocabularies. These vocabularies include broad coverage biomedical terminolgies, such as the NLM's Medical Subject Headings (MeSH), disease specific terminologies such as the National Cancer Institute's PDQ vocabulary, drug terminologies such as the National Drug Data File, and medical specialty vocabularies such as the Classification of Nursing Diagnoses and the Current Dental Terminology. The vocabularies have in most cases been developed for differing purposes. MeSH, for example, is a thesaurus that is used to index the biomedical literature. The World Health Organization's International Classification of Diseases (ICD) was originally developed for the analysis and comparison of morbidity and mortality data throughout the world. In addition, however, a clinical modification of this terminology is used broadly in the U.S. for hospital and office visit billing purposes. Clinical Terms Version 3 (formerly the Read Codes) is a comprehensive clinical vocabulary that is currently used throughout the British health care system. Other terminologies have been developed for use primarily in computer-based patient records and in hospital information systems. In addition to differences in purpose and scope of coverage, the terminologies differ widely in size, with some having hundreds of terms, and others having tens of thousands. Some of the vocabularies are simply lists of terms, sometimes including synonyms, while others are fully structured thesauri (NISO, 1993), whose terms are interrelated in a variety of ways,

and at least one source claims to be an ontology for a well-defined subdomain of medicine. Occasionally the relationships between the terms in the vocabulary are made explicit, but most often they are implicit.

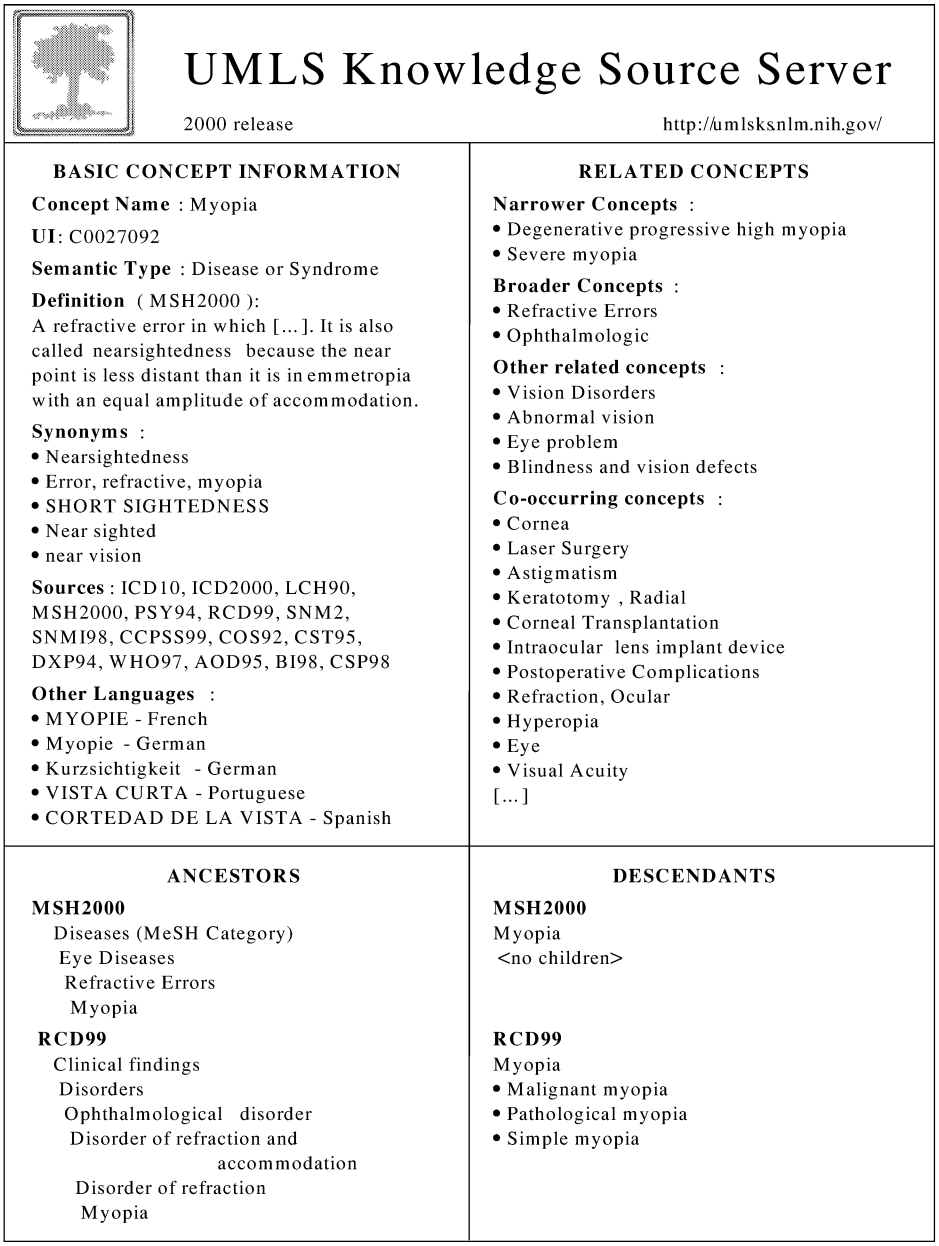
Each Metathesaurus concept may be thought of as a cluster of synonyms. As a new vocabulary is added, lexical and other techniques are used to map it into Metathesaurus concepts. Thus, for example, if the concept "otitis media" already exists in the Metathesaurus and if a new vocabulary contains the term "middle ear infection", then as the new vocabulary is integrated, "middle ear infection" will become part of the synonym class that defines the concept "otitis media".

Metathesaurus concepts are related to multiple other concepts through a small number of broadly defined relationships. These include the 'child' and 'parent' relationships that are found in the constituent vocabularies. Additionally, during Metathesaurus construction, a pair of concepts may be identified as being related by a 'narrower than', 'broader than', or 'other' (saliently related to) relationship. Further, in some cases, the precise nature of the relationship is made explicit by adding a relationship "attribute" label. For example, if a pair of concepts is in the child/parent relationship, the attribute might be 'isa', or it might be 'part of'. This latter could be true, for example, in an anatomy hierarchy. Or, it might be that an 'other' relationship can be made more precise. For example, "middle ear" would be related to "otitis media" by the 'other' relationship in the Metathesaurus, but more specifically, it might also be given the attribute 'location of'. Currently, only about 25% of Metathesaurus concepts have specific relationship attributes. These attributes correspond to relationships specified in the UMLS Semantic Network. Finally, many concepts are inter-related because they co-occur in MEDLINE® bibliographic citation records. In the study we describe in Section 3 below, we take advantage of all of these inter-concept relationships as we build a more robust conceptual structure for the use of the UMLS through the Semantic Network.

As the Metathesaurus is built, the content and structure of each of the source vocabularies is preserved, and additional information is added at the concept level. This process, which is done annually, is a combination of computational techniques and human review. The SPECIALIST lexical tools are used to suggest likely related terms and these are then reviewed for correctness by UMLS editors, all of whom are specialists in the medical domain. The editors add a variety of information, including additional synonyms if these are available, and they add relationships to existing concepts. They categorize the concepts by assigning to each of them one or more semantic types from the Semantic Network. They review definitions and make sure that the cluster of synonyms accurately reflects the meaning of the concept.

Figure 1 is a sample of the type of information that is available for Metathesaurus concepts. Searching the UMLS Knowledge Source Server for the term "nearsightedness", would retrieve the information shown in the figure. The top left shows basic concept information, including the preferred concept name "myopia" and the concept unique identifier (C0027092). As noted, each concept is assigned a semantic type from the Semantic Network, and in this case it is "Disease or Syndrome". Further, since a definition is available, this is also displayed.

The synonyms that make up a concept are often drawn from multiple vocabularies. For this concept, this includes two versions of the International Classification of Diseases (ICD), the Medical Subject Headings (MSH), the Systematized Nomenclature of Medicine



UMLS Knowledge Source Server

2000 release http://umlsks.nlm.nih.gov/

BASIC CONCEPT INFORMATION

Concept Name : Myopia

UI: C0027092

Semantic Type : Disease or Syndrome

Definition (MSH2000):
A refractive error in which [...]. It is also called nearsightedness because the near point is less distant than it is in emmetropia with an equal amplitude of accommodation.

Synonyms :
- Nearsightedness
- Error, refractive, myopia
- SHORT SIGHTEDNESS
- Near sighted
- near vision

Sources : ICD10, ICD2000, LCH90, MSH2000, PSY94, RCD99, SNM2, SNMI98, CCPSS99, COS92, CST95, DXP94, WHO97, AOD95, BI98, CSP98

Other Languages :
- MYOPIE - French
- Myopie - German
- Kurzsichtigkeit - German
- VISTA CURTA - Portuguese
- CORTEDAD DE LA VISTA - Spanish

RELATED CONCEPTS

Narrower Concepts :
- Degenerative progressive high myopia
- Severe myopia

Broader Concepts :
- Refractive Errors
- Ophthalmologic

Other related concepts :
- Vision Disorders
- Abnormal vision
- Eye problem
- Blindness and vision defects

Co-occurring concepts :
- Cornea
- Laser Surgery
- Astigmatism
- Keratotomy , Radial
- Corneal Transplantation
- Intraocular lens implant device
- Postoperative Complications
- Refraction, Ocular
- Hyperopia
- Eye
- Visual Acuity

[...]

ANCESTORS

MSH2000
Diseases (MeSH Category)
Eye Diseases
Refractive Errors
Myopia

RCD99
Clinical findings
Disorders
Ophthalmological disorder
Disorder of refraction and accommodation
Disorder of refraction
Myopia

DESCENDANTS

MSH2000
Myopia
<no children>

RCD99
Myopia
- Malignant myopia
- Pathological myopia
- Simple myopia

Figure 1. Concept information in the UMLS for “myopia”

(SNM), and the World Health Organization (WHO) adverse drug reaction terminology. Translations in several languages are also available. The lower portion of the figure shows the hierarchical contexts of the concept in the various vocabularies in which it appears. In MeSH, myopia is a child of the term "refractive errors", while in the British Read (RCD) vocabulary it is a child of the term "disorder of refraction". (These two parent terms, while expressed somewhat differently, are actually the same concept and are represented as such in the Metathesaurus.) The top right of the figure shows that myopia is related to several other concepts in the Metathesaurus. Two concepts, "degenerative progressive high myopia" and "severe myopia", are listed as being narrower in meaning than myopia, and two are broader in meaning. Through the explicit 'other' relationship, myopia is closely related to several additional concepts, including "abnormal vision" and "eye problem". It frequently co-occurs with many other concepts in MEDLINE. Interesting implicit relationships hold between these co-occurring concepts. One might, for example, surmise that myopia 'has location' eye, 'is treated' by laser surgery, and 'affects' visual acuity. These co-occurring concepts are indicators of what is being written about in the biomedical literature, but, perhaps more interesting in this context, they are indicators of powerful associations among biomedical concepts, creating a latent semantic space for the domain (Landauer & Dumais, 1997).

2.2 UMLS Semantic Network

The UMLS Semantic Network has in common with most semantic networks that it consists of a collection of basic semantic types, which are the nodes in the network, and a set of relationships, which are its links (Brachman, 1979; Greenhill & Venkatesh, 1998; Lehmann, 1992; Quillian, 1968; Ruan, Burkle, & Dudeck, 2000; Sowa & Borgida, 1991; Woods, 1985). Quillian is most often credited with first developing the notion of semantic networks, and his work, though actually developed in the context of a psychological theory, heavily influenced subsequent work in knowledge representation. Brachman was among the first to critically evaluate the formal semantics of semantic networks and to suggest the ways in which networks might be used to build knowledge representation languages. More recent work, such as that of Greenhill & Venkatesh and Ruan et al., has emphasized the power of semantic networks to navigate complex knowledge spaces, particularly through robust visualization tools.

The UMLS Semantic Network was developed in order to provide a high level semantic structure for organizing the biomedical domain. It has the potential, through its 134 semantic types and 54 relationships, to simplify and bring coherence to a very large semantic space. There are two basic type hierarchies, one for entities and the other for events. Semantic types for organisms, anatomical structures, chemicals, concepts or ideas, behaviors, and physiologic and pathologic functions are included. There are also two categories of relationships. The first is the 'isa' relationship and the other comprises the non-hierarchical associative relationships. These latter are divided into five additional categories, including physical, spatial, functional, temporal, and conceptual relationships. (See table 1 for the full list of types and relationships.) While many of the semantic types are specific to the biomedical domain, most of the relationships are equally applicable in domains outside of medicine.

Entity Physical Object Organism Plant Alga Fungus Virus Rickettsia or Chlamydia Bacterium Archaeon Animal Invertebrate Vertebrate Amphibian Bird Fish Reptile Mammal Human Anatomical Structure Embryonic Structure Anatomical Abnormality Congenital Abnormality Acquired Abnormality Fully Formed Anatomical Structure Body Part, Organ, or Organ Component Tissue Cell Cell Component Gene or Genome Manufactured Object Medical Device Research Device Clinical Drug	[Entity] (continued) [Physical Object] (continued) Substance Chemical Chemical Viewed Functionally Pharmacological Substance Antibiotic Biomedical or Dental Material Biologically Active Substance Neuroreactive Substance or Biogenic Amine Hormone Enzyme Vitamin Immunologic Factor Receptor Indicator, Reagent, or Diagnostic Aid Hazardous or Poisonous Substance Chemical Viewed Structurally Organic Chemical Nucleic Acid, Nucleoside, or Nucleotide Organophosphorous Compound Amino Acid, Peptide, or Protein Carbohydrate Lipid Steroid Eicosanoid Inorganic Chemical Element, Ion, or Isotope Body Substance Food

Table 1a. The UMLS Semantic Network: ‘isa’ relations between semantic types

[Entity] (continued)	Event
Conceptual Entity	Activity
Idea or Concept	Behavior
Temporal Concept	Social Behavior
Qualitative Concept	Individual Behavior
Quantitative Concept	Daily or Recreational Activity
Functional Concept	Occupational Activity
Body System	Health Care Activity
Spatial Concept	Laboratory Procedure
Body Space or Junction	Diagnostic Procedure
Body Location or Region	Therapeutic or Preventive Procedure
Molecular Sequence	Research Activity
Nucleotide Sequence	Molecular Biology Research Technique
Amino Acid Sequence	Governmental or Regulatory Activity
Carbohydrate Sequence	Educational Activity
Geographic Area	Machine Activity
Finding	Phenomenon or Process
Laboratory or Test Result	Human-caused Phenomenon or Process
Sign or Symptom	Environmental Effect of Humans
Organism Attribute	Natural Phenomenon or Process
Clinical Attribute	Biologic Function
Intellectual Product	Physiologic Function
Classification	Organism Function
Regulation or Law	Mental Process
Language	Organ or Tissue Function
Occupation or Discipline	Cell Function
Biomedical Occupation or Discipline	Molecular Function
Organization	Genetic Function
Health Care Related Organization	Pathologic Function
Professional Society	Disease or Syndrome
Self-help or Relief Organization	Mental or Behavioral Dysfunction
Group Attribute	Neoplastic Process
Group	Cell or Molecular Dysfunction
Professional or Occupational Group	Experimental Model of Disease
Population Group	Injury or Poisoning
Family Group	
Age Group	
Patient or Disabled Group	

Table 1a. The UMLS Semantic Network: 'isa' relations between semantic types—Cont.

isa
associated with
  physically related to
  part of
  consists of
  contains
  connected to
  interconnects
  branch of
  tributary of
  ingredient of
  spatially related to
  location of
  adjacent to
  surrounds
  traverses
  functionally related to
  affects
  manages
  treats
  disrupts
  complicates
  interacts with
  prevents
  brings about
  produces
  causes
	[associated with]
	  [functionally related to]
	  performs
	  carries out
	  exhibits
	  practices
	  occurs in
	  process of
	  uses
	  manifestation of
	  indicates
	  result of
	  temporally related to
	  co-occurs with
	  precedes
	  conceptually related to
	  evaluation of
	  degree of
	  analyzes
	  assesses effect of
	  measurement of
	  measures
	  diagnoses
	  property of
	  derivative of
	  developmental form of
	  method of
	  conceptual part of
	  issue in

Table 1b. The UMLS Semantic Network: 'isa' relations between relationships

Figure 2 shows a partial view of the Semantic Network, illustrating the kinds of relations that exist. As an example, note that anatomical structure is 'part of' an organism, an organism attribute is a 'property of' an organism, biologic function is a 'process of' an organism, and a plant 'isa' organism. By transitivity, a human also 'isa' organism.

The Semantic Network has a graph structure. Hierarchical (isa) relationships are organized in a single tree structure whereas associative relationships link semantic types from various levels of the hierarchical structure.

A range of information is provided for each semantic type. Each type is assigned a unique identifier and also a number that places it in the hierarchy of semantic types. For example, the unique identifier of "Experimental Model of Disease" is T050, and its tree number is B2.2.1.2.3, and it is a child of the semantic type "Pathologic Function" (B2.2.1.2). A definition is given for each type, and several examples of concepts to which this type may be assigned are also provided. In this case, the definition is "A representation in a non-human organism of a human disease for the purpose of research into its mechanism or treatment." A usage note assists those who are assigning semantic types to

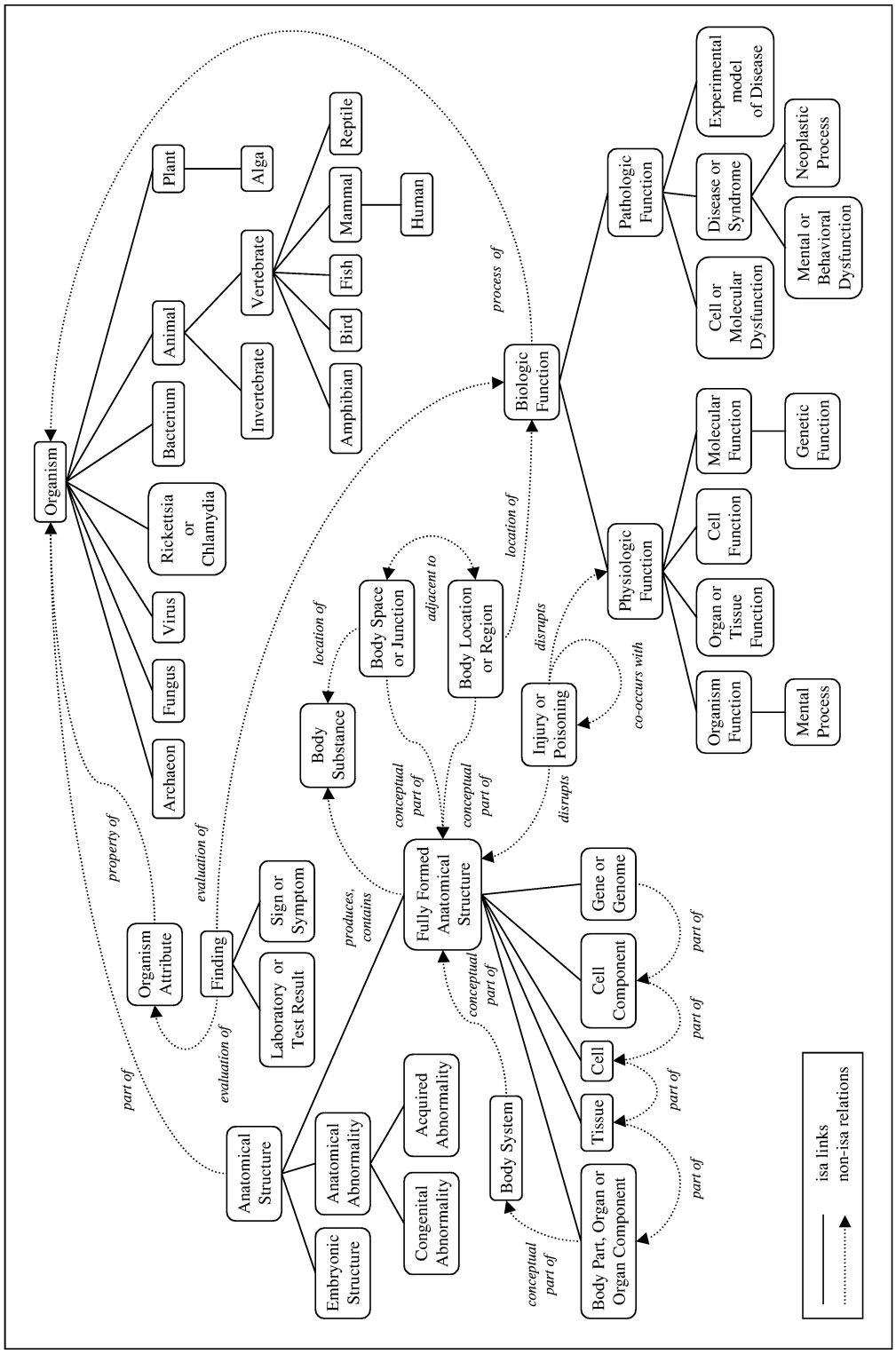


Figure 2. The UMLS Semantic Network (partial representation)

Metathesaurus concepts. In addition, through the UMLS Knowledge Source Server, a UMLS editor may at any time see the other concepts to which this semantic type has been assigned. "Experimental Model of Disease" has been assigned to fifty-three concepts in the current Metathesaurus, including "avian leukosis" and "experimental melanoma".

Analogously, information for relationships includes unique identifiers, the name of the relationship and its inverse, a tree number that places it in the hierarchy of relationships, a definition, and the pairs of semantic types it links. For example, the relationship 'treats' is defined as, "Applies a remedy with the object of effecting a cure or managing a condition." Its tree number is R3.1.2, and its parent in the relationship hierarchy is 'affects' (R3.1), which is in the 'functionally related to' (R3) hierarchy. The treats relationship links, for example, pathologic functions and injuries to pharmacologic substances, therapeutic procedures, and medical devices. The relationships are stated between high level semantic types and are generally inherited by all the descendants of those types. In this case, then, drugs are linked by the 'treats' relationship not only to pathologic functions, but also to all of the descendants of pathologic function, including diseases, mental or behavioral dysfunctions, and neoplastic processes. It is important to note that the relationships link semantic types to each other, but they do not directly link concepts to one another.

The UMLS Semantic Network has been explored by a number of researchers (Carenini & Moore, 1993; Gu et al., 2000; Joubert, Miton, Fieschi, & Robert, 1995; Volot et al., 1993; Yu, Friedman, Rhzetsky, & Kra, 1999). The focus of some of the work has been to "reuse" the knowledge encoded in the Semantic Network and express it in a variety of different knowledge representation frameworks. For example, Joubert et al. and Volot et al. have reformulated the Semantic Network in the closely related conceptual graph theory, and Gu et al. have represented both the Metathesaurus and the Semantic Network in an object oriented database framework. Among others, Carenini & Moore and Yu et al. have experimented with the Semantic Network in conceptualizing smaller, more focused domains within the broader biomedical domain.

3. BUILDING THE CONCEPTUAL FRAMEWORK: AN EXTENDED EXAMPLE

In order to illustrate the ability and the limits of the UMLS Semantic Network to provide a conceptual framework for the biomedical domain, we designed the following study. Starting from a given concept, we gathered the concepts that constitute its semantic neighborhood by exploiting a set of inter-concept relationships represented in the UMLS Metathesaurus. For each pair of related concepts from this set, we calculated the possible relationships between the concepts using the semantic links defined in the UMLS Semantic Network between the semantic types that had been assigned to these concepts. Besides revealing the semantic structure in this set of concepts, other expected results included qualifying broadly defined relationships in the Metathesaurus, assessing already defined ones, and, more generally, by enforcing semantic rules, detecting inconsistencies in the Metathesaurus or in the Semantic Network itself.

The top part of figure 3 represents the semantic types and the relationships between them, as defined in the Semantic Network. The Metathesaurus, a set of concepts linked by inter-concept relationships, is represented in the bottom part of the figure. The two

structures are related by means of the semantic types assigned to the concepts in order to categorize them. Therefore, inter-concept relationships can be inferred, validated, or rejected by comparison to the relationships defined between the semantic types assigned to the concepts.

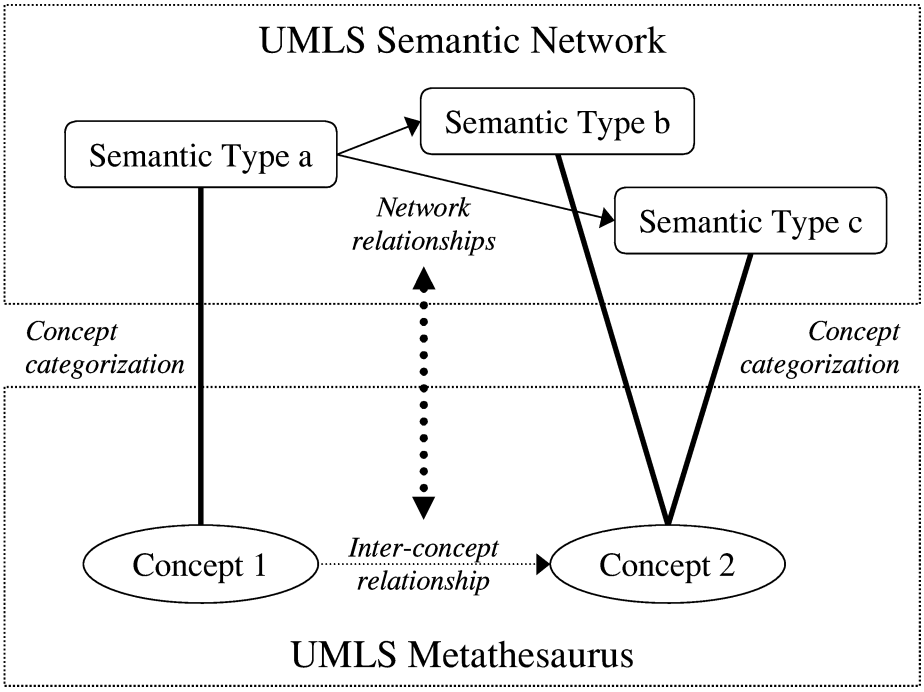


Figure 3. Defining the semantic structure for the domain.

3.1 Materials and Methods

3.1.1 Selecting a Set of Concepts for Experimentation

Starting from the concept "heart" (unique identifier: C0018787), we used the information available in the 1999 edition of the Metathesaurus to discover the concepts related to it:

- Concepts that are hierarchically related to "heart" through the 'parent', 'broader than', 'child', and 'narrower than' relationships were extracted. Hierarchically related concepts were extracted all the way to the top and bottom of the hierarchy, so that the set contains all the ancestors and descendants of "heart".
- Concepts that are related to "heart" through the 'other' (associative) relationship were extracted. In this case, only concepts directly associated with "heart" were selected.
- Concepts co-occurring with "heart" at least four times in MEDLINE were extracted. This selection represents 90% of all co-occurrences for "heart".

Using this methodology, we discovered 3764 closely related concepts. Figure 4 shows the partitioning of the concepts with respect to their relationship to "heart". Most concepts are related by only one type of relationship: Sixty-six are ancestors, 1952 are descendants, 242 are related through the 'other' relationship, and 1412 are co-occurring concepts. Some concepts are related by multiple relationships: One ancestor and twenty-seven descendants are also in the 'other' relationship to "heart". Two ancestors and thirty-two descendants also co-occur with "heart". Finally, twenty-five concepts in the 'other' relationship also co-occur with "heart", and five concepts simultaneously co-occur with, are descendants, and are in the 'other' relationship to "heart". These latter two groups are not shown in the figure.

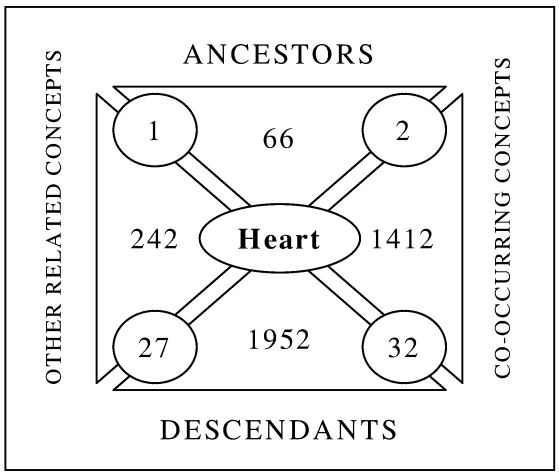


Figure 4. Origin of the concepts in the set of concepts related to "heart"

3.1.2 Preparing the Semantic Links

In order to facilitate knowledge processing, the UMLS provides a file of all semantic links resulting from the transitive closure of the Semantic Network graph. For example, no direct link is specified between the two semantic types "Disease or Syndrome" and "Body Part, Organ, or Organ Component". The 'location of' relation, however, may be inferred from the link between "Biologic Function" and "Fully Formed Anatomical Structure" since "Disease or Syndrome" is a descendant of "Biologic Function" and "Body Part, Organ, or Organ Component" is a child of "Fully Formed Anatomical Structure". Figure 5 shows an example of some additional semantic links that are calculated by the transitive closure of the graph.

For practical purposes, we complemented the list of explicit links between semantic types using two pieces of information implicit in the Semantic Network:

- We systematically added a reflexive 'isa' link (e.g. "Disease or Syndrome" isa "Disease or Syndrome") to reflect inter-concept relationships such as "myocardial infarction" 'isa' "cardiovascular diseases".

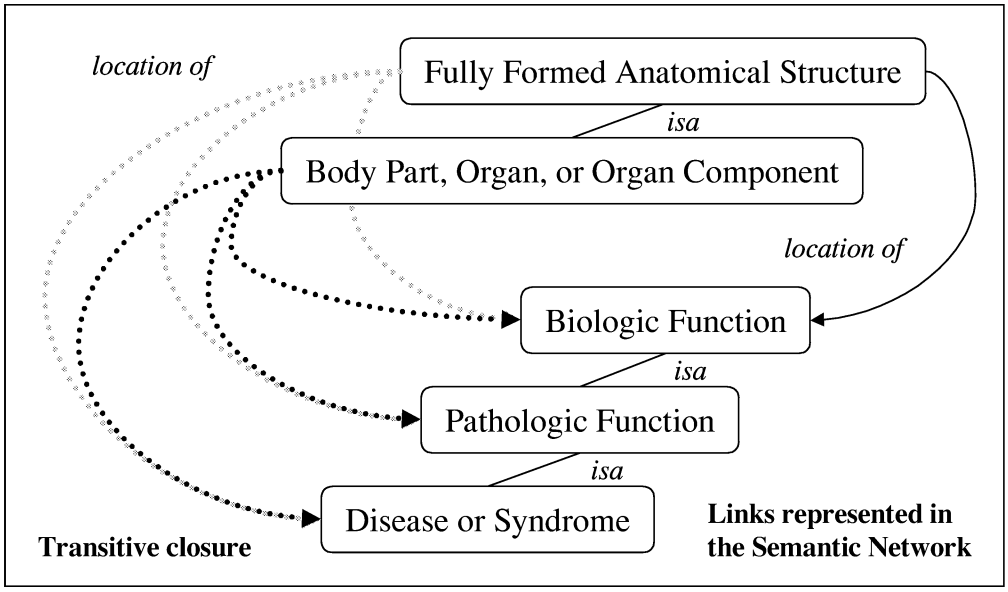


Figure 5. Additional semantic links calculated by transitive closure

- We added the symmetric inverse link for each pair of semantic types (e.g. "Body Part, Organ, or Organ Component" 'has location' "Disease or Syndrome", or "Disease or Syndrome" 'inverse isa' "Disease or Syndrome").

Once augmented, this resulted in 13,590 links (direct, inverse, and reflexive) among the 134 semantic types in the UMLS Semantic Network.

3.1.3 Revealing the Semantic Structure

Our set of 3764 concepts represents a total of 6894 pairs of related concepts. These numbers represent all the concepts directly related to "heart" as well as the concepts in the full set that are related to each other. The Metathesaurus may stipulate that a relationship exists between a pair of concepts, but it does not necessarily stipulate the precise nature ("attribute") of that relationship. To determine specific relationships between pairs of concepts, we followed several steps:

We abstracted away from the concepts themselves and compared their semantic types to the allowable Semantic Network links.

- If there was only one possible relationship for the semantic type pair, this one was tentatively chosen as the inter-concept relationship.
- If there were multiple possible relationships between the semantic type pair, then the Metathesaurus was checked for the original broad relationship between the concepts. In this case,
  - If the original broad relationship was a hierarchical one ('child', 'parent', 'broader than' or 'narrower than'), and if 'isa' was an allowable relationship between the semantic type pair, then 'isa' was tentatively chosen as the inter-concept relationship.

  - If the original broad relationship was associative ('other') or if the concepts were derived from co-occurrence, then if there was only one allowable associative relationship between the semantic type pair, it was tentatively chosen.
  - If there was more than one associative relationship, then it was not possible to make a tentative assignment, and the pair was ambiguous.
- If no relationship existed between the semantic type pair, then no inter-concept relationship could be inferred.

All tentative and ambiguous inter-concept assignments were then checked in the Metathesaurus to see if there were relationship attributes listed for the specific concept pair. In this case,

- If the attribute was compatible with an allowable relationship between the semantic type pair, it was chosen as the inter-concept relationship.
- If there was no attribute, or if the attribute was incompatible, the inter-concept relationship could not be resolved.

3.2 Results

Among the 6894 pairs of related concepts, we obtained the following results:

- In 4496 cases (65%), a semantic relation could be inferred unambiguously from the Semantic Network. The semantic relation inferred allowed us to determine inter-concept relationships whose attribute was not defined in 2515 of these cases, and to confirm the validity of the relationship attribute in 1981 of these cases.
- In 1491 cases (22%), multiple semantic links existed between the semantic types of the two concepts, leading to several possible attributes for these inter-concept relationships.
- In the remaining 907 pairs (13%), the inter-concept relationships represented a violation of the Semantic Network. In 372 pairs, there was no semantic link between the semantic types of the two concepts. In 415 pairs, the inter-concept relationship was not compatible with that of the corresponding Semantic Network relationship. Finally, in 120 pairs, the attribute of the inter-concept relationship was not compatible with the semantic relationships allowed between the semantic types of the two concepts.

Although the relevance of the semantic relations inferred from the Semantic Network is difficult to evaluate without systematic human review, their compatibility with the relationship attribute, when specified, might provide an estimate. In 89% of the cases, the semantic relationship inferred from the Semantic Network (unambiguously or not) is compatible with the relationship attribute.

Figure 6 shows a portion of the conceptual framework we built for cardiology. The most common relationships in this set of concepts are 'isa', organizing diseases, and 'location of', linking diseases or procedures to the corresponding "Body Part, Organ, or Organ Component". Various other associative relationships help structure the domain. The attribute of inter-concept relationships is accurately and unambiguously inferred in

most cases. However, in some cases attributes are incorrectly inferred. For example, 'isa' is inferred between "heart" and "heart valves", while it should be 'part of'. Finally, the relationship of the sign "systolic murmur" to the disease "mitral valve insufficiency" cannot be selected automatically since three possible relationships are defined between their semantic types.

3.3 Discussion

The Semantic Network is able to suggest a semantic structure from a set of concepts in the Metathesaurus. Additional knowledge about inter-concept relationships is not technically required for inferring semantic relations from the Network. However, since the Semantic Network defines relationships between very broad categories, these may not hold between concepts to which these categories are assigned. For example, "heart" 'location of' "intracranial pressure" is valid according to the Semantic Network, since a "Body Part, Organ or Organ Component" is the location of an "Organ or Tissue Function". However, this is an incorrect inference. Therefore, the Semantic Network is best used in collaboration with the Metathesaurus, where the existence of a particular relationship, hierarchical or associative, between concepts is described, even though broadly. Using the methods we have described here, the Semantic Network is able to suggest the appropriate, more specific, relationship. The results of this study do not suggest an entirely automated approach for organizing biomedical concepts, in particular since about 20% of the semantic relationships could not be inferred unambiguously. However, the method can be used to suggest possible relationships to the human editors of the Metathesaurus. For example, the relationship of "heart" ("Body Part, Organ, or Organ Component") to "cobra venoms" ("Biologically Active Substance") may be either 'produces' or 'disrupted by'. A human editor will easily select the latter when prompted with this selection.

This method also helps detect discrepancies between the semantics of the Metathesaurus and the semantics expressed by the Semantic Network. Automatic detection helps limit the need for human review by focusing on conflicting relationships that violate the semantic rules. One major cause for such a violation is somewhat artificial: Concepts with an abstract semantic type (e.g. "Classification") may have related concepts having a concrete semantic type (e.g. "Body Part, Organ, or Organ Component"). These types would be unrelated in the Semantic Network. The relationship of "heart: general terms" to "right side of heart", for example, violates the Semantic Network for this reason. Another source of problems is that partonymic relationships (part of) are considered associative in the Semantic Network, while in many medical vocabularies they are used hierarchically. Frequently occurring semantic discrepancies may also help identify missing semantic links in the Semantic Network. For example, the relationship of "chest pain" to "thorax" violates the Semantic Network, since the 'location of' relationship has not been defined between a "Body Location or Region" and a "Sign and Symptom".

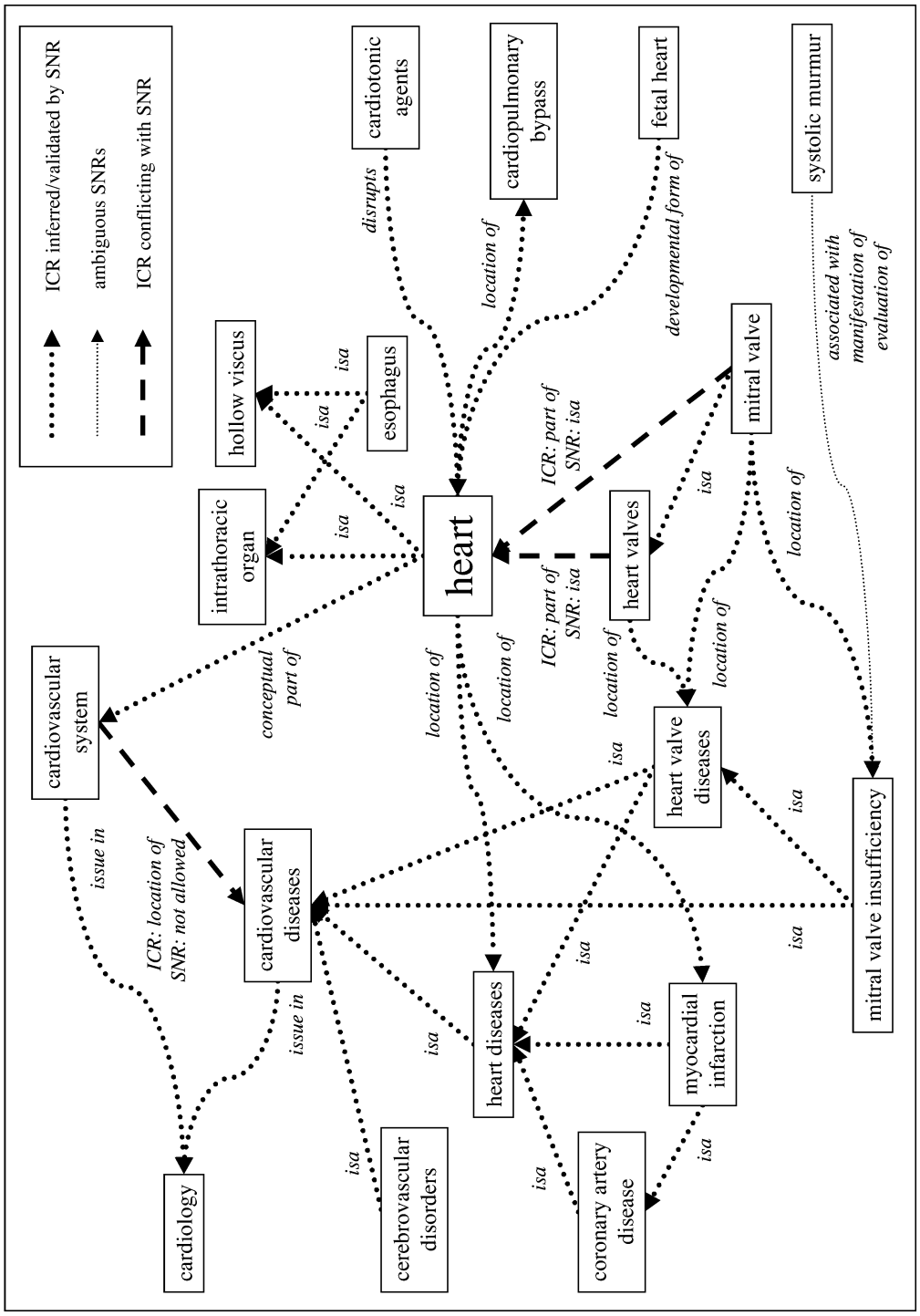


Figure 6. A portion of the conceptual framework for cardiology (attributes of the inter-concept relationships [ICRs] as suggested from semantic network relationships [SNRs])

4. CONCLUSIONS

In the UMLS each concept can be understood and defined by its relationships to other concepts and by the semantic category to which it belongs. This principle, called semantic locality, is an important organizing principle in the UMLS. The semantic structure of the UMLS consists of two related parts: the Semantic Network and the Metathesaurus. The Semantic Network provides a small number of strong semantic rules, by defining relationships among a small number of high-level semantic types. The Metathesaurus provides a large number of inter-related concepts. The semantic typing of all concepts in the Metathesaurus allows it to inherit the semantic rules provided by the Semantic Network. Together, the two knowledge sources define a large portion of the biomedical domain.

Ideally, inter-concept relationships in the Metathesaurus could be limited to representing factual knowledge, relying on the Semantic Network to interpret this knowledge. For example, it might be sufficient to know that "heart" is related to "heart diseases", if the nature of the relationship can be inferred unambiguously at a higher level. In fact, however, because the Semantic Network has a limited number of semantic types of coarse granularity for a domain as broad as biomedicine, it does not allow us to completely achieve such a goal. However, we have shown that the Semantic Network can be used to infer quite accurately the nature of the relationships between concepts in the Metathesaurus in a particular subdomain of medicine. The results suggest that this method can be applied more broadly to the Metathesaurus as a whole.

References

Bodenreider, O., Burgun, A., Botti, G., Fieschi, M., Le Beux, P., & Kohler, F. (1998). Evaluation of the Unified Medical Language System as a medical knowledge source. *Journal of the American Medical Informatics Association,* 5, 76-87.

Brachman, R. J. (1979). On the epistemological status of semantic networks. In N. Findler (Ed.), *Associative Networks: Representation and Use of Knowledge by Computers*, 3-50. New York: Academic Press.

Carenini, G., & Moore, J. D. (1993). Using the UMLS Semantic Network as a basis for constructing a terminological knowledge base: A preliminary report. *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care,* 725-729.

Cimino, J. J. (1998). Auditing the Unified Medical Language System with semantic methods. *Journal of the American Medical Informatics Association,* 5, 41-51.

Greenhill, S., & Venkatesh, S. (1998). Noetica: A tool for semantic data modelling. *Information Processing & Management,* 34, 739-760.

Gu, H. Y., Perl, Y., Geller, J., Halper, M., Liu, L. M., & Cimino, J. J. (2000). Representing the UMLS as an object-oriented database: Modeling issues and advantages. *Journal of the American Medical Informatics Association,* 7, 66-80.

Joubert, M., Miton, F., Fieschi, M., & Robert, J. J. (1995). A conceptual graphs modeling of UMLS components. *Medinfo,* 8, 90-94.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review,* 104, 211-240.

Lehmann, F. (1992). Semantic networks. In F. Lehmann (Ed.), *Semantic Networks in Artificial Intelligence*, 1-50. Tarrytown, NY: Academic Press.

Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. *Methods of Information in Medicine,* 32, 281-291.

McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods of Information in Medicine,* 34, 193-201.

McCray, A. T., Razi, A. M., Bangalore, A. K., Browne, A. C., & Stavri, P. Z. (1996). The UMLS Knowledge Source Server: A versatile Internet-based research tool. *Proceedings of the 1996 AMIA Annual Fall Symposium*, 164-168.

McCray, A. T., Srinivasan, S., & Browne, A. C. (1994). Lexical methods for managing variation in biomedical terminologies. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care,* 235-239.

National Information Standards Organization (NISO). (1993). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri (Developed by the National Information Standards Organization. Approved August 30, 1993, by the American National Standards Institute).*

Pisanelli, D. M., Gangemi, A., & Steve, G. (1998). An ontological analysis of the UMLS Methathesaurus. *Proceedings of the 1998 AMIA Symposium*, 810-814.

Quillian, M. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic Information Processing,* 227-270. Cambridge, MA: MIT Press.

Ruan, W., Burkle, T., & Dudeck, J. (2000). An object-oriented design for automated navigation of semantic networks inside a medical data dictionary. *Artificial Intelligence in Medicine,* 18, 83-103.

Sowa, J. F., & Borgida, A. (1991). *Principles of Semantic Networks : Explorations in the Representation of Knowledge*. San Mateo, CA: Morgan Kaufmann.

Srinivasan, P. (1999). Exploring the UMLS: A rough sets based theoretical framework. *Proceedings of the 1999 AMIA Symposium*, 156-160.

Volot, F., Zweigenbaum, P., Bachimont, B., Ben Said, M., Bouaud, J., Fieschi, M., & Boisvieux, J. F. (1993). Structuration and acquisition of medical knowledge. Using UMLS in the conceptual graph formalism. *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care,* 710-714.

Woods, W. (1985). What's in a link: Foundations for semantic networks. In R. J. Brachman & H. J. Levesque (Eds.), *Readings in Knowledge Representation,* 218-241. Los Altos, CA: Morgan Kaufmann.

Yu, H., Friedman, C., Rhzetsky, A., & Kra, P. (1999). Representing genomic knowledge in the UMLS semantic network. *Proceedings of the 1999 AMIA Symposium*, 181-185.