Chapter 6

Relationships among Knowledge Structures: Vocabulary Integration within a Subject Domain

Olivier Bodenreider
National Library of Medicine, Bethesda, MD, USA

Carol A. Bean
School of Information Sciences, University of Tennessee, Knoxville, TN, USA

Abstract:
The structure of terminology systems can be seen as one way to organize knowledge. This paper focuses on three types of relationships among terms: synonymy, hierarchical relationships, and explicit mapping relationships. Examples drawn from various medical vocabularies illustrate each type of relationship. The integration of disparate terminological knowledge structures in the Unified Medical Language System is presented and discussed.

1. INTRODUCTION

There are a large number of terminology systems used in medicine. Recent reviews present the scope and the structure of the major medical vocabularies (Cimino, 1996), and evaluate their content coverage (Chute et al., 1996) or their features (Campbell et al., 1997). While some vocabularies have been used for more than a century (e.g., the International Classification of Diseases), others are still very much works in progress (e.g., GALEN, SNOMED-RT).[1] Often vocabularies are designed to serve one particular purpose: For example, the U.S. National Library of Medicine (NLM) develops and uses the Medical Subject Headings (MeSH) as its controlled vocabulary for subject cataloging and to index articles from medical journals. Conversely, the International Classification of Diseases (ICD) is not only used world-wide to record causes of death or to register diseases in health statistics, but many adaptations of it (e.g., ICD-9-CM) are also used to record diagnoses or contact with health services for billing purposes.

Despite recently-formed partnerships between the producers of some major vocabularies (e.g., between SNOMED and LOINC, or between SNOMED-RT and Clinical Terms Version 3[2]), most vocabularies are usually developed independently from one another. Several studies have examined principles for the construction of medical vocabularies (Chute, Cohn, & Campbell, 1998; Cimino, 1998; Evans et al., 1994; Rada et al., 1993; Rossi Mori et al., 1993). Nonetheless, emerging standards such as those defined by the European Committee for Standardization (CEN) have not yet been widely adopted.

In: Bean CA, Green R, editors. Relationships in the organization of knowledge: Kluwer; 2001. p. 81-98.

(For an overview, see Rossi Mori, Consorti, & Galeazzi, 1998).

No single vocabulary offers both a coverage broad enough to encompass the whole biomedical domain and a granularity suitable for the description of patient conditions in applications such as electronic patient records (Chute et al., 1996). In the last fifteen years, two major projects[3] using different approaches have been developed towards such a goal.

A top-down approach has been used in the European Union GALEN project. GRAIL, the "GALEN Representation and Integration Language," was designed prior to defining the CORE model for the representation of medical concepts (Rector & Nowlan, 1994). Putting such an emphasis on the conceptual model has enabled GALEN's success in developing language-independent terminology services to exploit the knowledge representation (Rector et al., 1995), but it still lacks broad coverage (Rector et al., 1998).

The Unified Medical Language System (UMLS) was developed at the U.S. NLM using a bottom-up approach. It provides a common interface to about 40 existing medical vocabularies and reduces the ambiguity inherent in large bodies of content (Humphreys et al., 1998; Lindberg, Humphreys, & McCray, 1993). The structure of a semantic network strengthens the limited knowledge model inherited from each vocabulary and refined by the UMLS editors. With more than 600,000 medical concepts, the UMLS now has reasonably broad coverage, but its knowledge representation is weaker than GALEN's.

The role played by terms is very different in the two systems: The UMLS can be described as a system that organizes terms, while terms are a by-product of the GALEN system. In other words, the UMLS makes heavy use of lexical knowledge to link precoordinated terms together, while terms are generated from the combination of atomic concepts under the GALEN model.

The identification of relationships among knowledge structures inherited from medical vocabularies was an early goal in the UMLS project and has been long been recognized for contributing added value in the UMLS Metathesaurus (Bodenreider, Nelson, et al., 1998; Cimino et al., 1993; Dessena, Rossi Mori, & Galeazzi, 1999). It is thus quite natural to use the UMLS to illustrate how these relationships are discovered through lexical knowledge, heuristics, and the knowledge of human editors. Numerous journal articles and presentations at international conferences have already described the structure (Nelson et al., 1992) and formal properties (Tuttle et al., 1994) of the UMLS Metathesaurus, as well as the methodology used for its creation and maintenance (Sherertz et al., 1990; Sperzel et al., 1992; Suarez-Munist et al., 1996; Tuttle et al., 1995); interested readers are referred to this literature. However, key elements of UMLS Metathesaurus construction and editorial processes will be briefly discussed as needed to illuminate the relationships among knowledge structures in this particular context.

To show how underlying knowledge structures may be connected through relationships, this chapter focuses on three types of relationships among terms: synonymy, hierarchical relationships, and explicit mapping relationships. Background information and examples from various medical vocabularies are provided for each type of relationship, and specific implications for integration among knowledge structures are discussed.

2. SYNONYMY

2.1 Vocabulary Terms and UMLS Concepts

Except for systems that focus mainly on knowledge representation such as GALEN and SNOMED-RT (and to some extent SNOMED International and Clinical Terms Version 3), the design of medical vocabularies, including the UMLS Metathesaurus, is enumerative rather than compositional. Enumerative terminologies represent each concept by one or more term, regardless of the concept's complexity. Enumerative description is independent of language-surface forms and results in lists of precoordinated terms whose validity and consistency are difficult to test computationally. In contrast, compositional models produce a formal and often complex representation for the concepts that is suitable for manipulation by computer programs. They are usually more difficult to design and labor-intensive to populate (Rassinoux et al., 1997).

In the UMLS, concepts are defined by extension: that is, by a list of terms that are equivalent in meaning. The concept is a sort of virtual entity, identified by a unique identifier (CUI). The concept has no name directly associated with it: By convention, a term is selected from the list of preferred terms in each vocabulary to be the preferred name for this concept, according to a precedence table based on the source (Campbell, Oliver, et al., 1998; McCray & Nelson, 1995). Terms in languages other than English, translated from one vocabulary already integrated in the UMLS, are part of the same concept as their English source. The 1999 edition of the UMLS Metathesaurus includes 1,134,891 terms corresponding to 626,313 concepts. UMLS concepts are of varying complexity and granularity.

Numerous concepts are named using but a few words (e.g., "Head," "Allergic reaction," or "Screening for diabetes"). However, other concepts bear long names resulting from verbose descriptions of medical procedures (e.g., "Electrocardiographic monitoring for 24 hours by continuous computerized monitoring and non-continuous recording, and real-time data analysis utilizing a device capable of producing intermittent full-sized waveform tracings, possibly patient activated; physician review and interpretation," from the Physicians' Current Procedural Terminology) or complex structures such as chemical compounds (a name for the anti-asthmatic drug called "Theophylline" is "3,7-Dihydro-1,3-dimethyl-1H-purine-2,6-dione").

Short names may hide complex concepts. "Transurethral prostatectomy," although a fairly simple name, describes a surgical procedure where prostatic tissue surrounding the urethra is removed using a special kind of endoscope inserted through the urethra. An "Open prostatectomy," on the other hand, differs from the former by more than just one qualifier: In this surgical procedure, an incision is made in the lower abdomen through which the whole prostate is removed by means of surgical instruments.

2.2 Synonyms

Synonymy is based on equivalence in the meaning of terms, so that one term can be interchanged with another, with no change in meaning. Formal definitions of synonymy involve the mutual entailment of sentences containing synonym terms. For example, “Pyrosis” and “Heartburn” are synonyms, both referring to the retrosternal sensation of burning often associated with the reflux of the acid stomach contents into the oesophagus.

In practice, however, such a strict definition is rarely used, and looser definitions are preferred. The UMLS Metathesaurus uses such a loose definition for practical reasons, so that closely related terms are considered synonyms, even though they don’t necessarily have the formal properties of strict synonyms (McCray & Nelson, 1995). For example, “Renal cell carcinoma” (RCC) and “Kidney cancer” are considered synonyms, which might reflect that RCC is the most common form of kidney cancer in adults. “Kidney cancer,” however, is actually broader in meaning than RCC since it also includes, among others, the most common form of kidney cancer in children (nephroblastoma), and kidney metastases.

In enumerative vocabularies, lexical resemblance is the major technique used to detect possible semantic closeness among lexical items (e.g., McCray, 1998). Lexical matching techniques include case normalization, removal of genitive markers, removal of punctuation, and word sorting among other techniques (McCray, Srinivasan, & Browne, 1994).

Another source of synonyms is the vocabularies themselves. Some medical vocabularies provide a list of synonyms (e.g., SNOMED International). Vocabularies such as MeSH append to each descriptor (or Main Heading) a list of entry terms. Entry terms are not necessarily synonyms of the main heading, but since they are expected to play an identical role in information retrieval, they are closely related to, and are at least possible candidates for, synonymy.

Whether discovered through lexical resemblance techniques or contributed by a source vocabulary, synonymy among terms in the UMLS Metathesaurus is assessed after a review by human editors. Synonymous terms represent the different possible names for a concept.

2.3 Integration Issues Related to Synonymy

2.3.1 Granularity

Synonymous relationships that are valid in the context of one vocabulary, according to its granularity, may become invalid or misleading when several vocabularies of different granularity are used simultaneously or merged. For this reason, the UMLS Metathesaurus may not incorporate all synonyms suggested by the source vocabularies. For example, “Ornithosis” and “Psittacosis” are two clinical forms of the same disease, an infection transmitted by contact with infected birds and marked by a respiratory infection and flu-like symptoms. Although “Ornithosis” and “Psittacosis” are often considered synonyms, they are represented by two distinct concepts in the UMLS Metathesaurus.

The quasi-synonymous relationship between “Renal cell carcinoma” and “Kidney cancer,” presented earlier is found in PDQ, the National Cancer Institute’s cancer database, and has been integrated in the UMLS Metathesaurus. Another example is the synonymous relationship between “Fetal cephalhematoma” and “Cephalohematoma” provided by SNOMED International. While “cephalhematoma” and “cephalohematoma” are spelling variants, the qualifier “fetal” suggests that “Fetal cephalhematoma” is narrower than

"Cephalohematoma." Practically, however, the two terms are synonyms, since cephalhematoma refers to a condition seen almost exclusively in the newborn.

2.3.2 Implicit Contextual Knowledge

As mentioned above, natural language processing techniques are used to compute lexical resemblance among terms as a means of identifying potential synonyms. These techniques assume that terms are both syntactically correct and fully specified entities. While most terms found in medical vocabularies are correct noun phrases (without an initial determiner), some of them are not fully specified, but rather defined by comparison to a parent term. This is especially true of vocabularies that were not designed to be used computationally, such as the International Classification of Diseases (ICD).

In ICD, choices made for the presentation of terms include tabulation and the use of dashes to avoid repeating the part of a term used in several derived terms. For example, the different forms of "Alcoholic hepatic failure" listed below the term include, among others, "- acute," "- chronic," and "- subacute." The alphabetical index can appear even more obfuscated at first sight with terms such as "- - - - cervix." The term "Female infertility of cervical origin" has to be reconstructed by finding the parts corresponding to each dash (here, Infertility / female / associated with / congenital anomaly / cervix), sometimes several pages earlier. This convention makes the index much smaller and therefore somewhat easier to read, but also renders it almost impossible to manipulate computationally.

For the same reasons, the context of a chapter or a group of terms is not always present in every term of this chapter or group. For example, the term "Prostate" (D07.5) doesn't refer to the prostate gland as an organ, but rather to a location for the condition "Carcinoma in situ of other and unspecified genital organs." A fully specified term for D07.5 would be "Carcinoma in situ of prostate."

ICD is by no means the only vocabulary where implicit knowledge of the context is necessary. Such a design is common and is beneficial as long as the vocabulary is not used for natural language processing or knowledge representation. However, since numerous UMLS-based applications take advantage of lexical processing and would be confused by multiple meanings for the same term, Metathesaurus editors often restore meaningful terms from the context prior to integrating them into the UMLS.

2.3.3 Evolution over Time

Synonyms in the loose definition often change over time; this is especially true for synonyms across knowledge structures (Cimino & Clayton, 1994). Some terms once considered synonyms may be split into several distinct concepts, such as what occurs when a finer grained vocabulary is encountered (refinement), or when terms showing the same surface form actually have different meanings (disambiguation). Conversely, terms originally not considered to be synonyms and assigned to different concepts may be merged into one concept, with one or more concepts being deleted. The CONCORDIA model (Oliver et al., 1999) addresses the issue of such changes in medical terminologies.

The UMLS Metathesaurus keeps track of merges, splits, and deletions. These vocabulary maintenance issues make it difficult for data encoded using one version of the Metathesaurus to be used consistently with later versions.

Concepts deleted following a merge process must be given the identifier of the concept they have been merged into. For example, in the UMLS Metathesaurus, the term "Abnormal electrocardiogram" (formerly a name for the C0000752 concept) was merged into the C0522055 concept ("Abnormal electrocardiographic finding") in 1999.

Conversely no simple solution exists for splits. To decide whether the original concept C (named by term T), now split into C_1 (named by term T_1) and C_2 (named by term T_2), should be coded C_1 rather than C_2 would require additional information about its original meaning. The original concept C should be retained and be renamed "T_1 or T_2" to ensure compatibility with older data. For example, an earlier version of the UMLS used a single concept for "Cryptorchidism" and "Ectopia testis." Both terms suggest that the testicle failed to descend into the scrotum. However, in cryptorchidism the testicle is located at some point on its migration path, which is not the case in ectopia testis. Because of this distinction, the treatment for these two conditions can be quite different, and thus the two terms are not synonyms to a urologic surgeon. This was corrected in a subsequent version of the UMLS by removing "Ectopia testis" from the synonyms of "Cryptorchidism," and by creating a new concept for it. As a consequence, the meaning of the original concept drifted from "Cryptorchidism or Ectopia testis" to "Cryptorchidism [only]," making it difficult to compare data coded with different versions of the UMLS (Bodenreider, Burgun, et al., 1998).

This problem, although more likely to occur across heterogeneous data structures, can also occur within a single vocabulary family (e.g., the evolution of the ICD, from the 9th to the 10th revision).

3. HIERARCHICAL RELATIONSHIPS

Hierarchical relationships present a powerful means for structuring knowledge. Three primary structural models are commonly used in medical vocabularies: trees, graphs, and conceptual structures.

Traditional medical classifications are monohierarchical; that is, they have a simple single-tree architecture and use the position in the tree to identify concepts. The ICD is organized according to this architecture.

Other vocabularies allow concepts to have several parent concepts and do not use concept identifiers directly to describe their architecture. Concepts are usually given a unique identifier, while the structure is described either by independent identifiers or by a list of parent-child pairs based on the unique identifiers. MeSH descriptors, for example, have both one unique identifier and one or more tree numbers. Clinical Terms Version 3 and GALEN also use polyhierarchical structures. Such a data structure is called a directed acyclic graph (DAG).

Conceptual graphs (Sowa, 1984) have been used in the biomedical domain to address issues as diverse as clinical concept and data representation, classification systems, information retrieval, and natural language understanding and processing (Volot, Joubert,

& Fieschi, 1998); however, few medical vocabularies actually use them. Medical terminology systems based on conceptual structures and description logic formalisms include GALEN (Rector et al., 1997), using the GALEN Representation and Integration Language (GRAIL), and SNOMED-RT (Spackman, Campbell, & Cote, 1997), using the Knowledge Representation System Specification (KRSS).

3.1 UMLS Metathesaurus

Since it preserves the original structure of its source vocabularies, some of which allow multiple inheritance, the UMLS Metathesaurus has a *de facto* graph structure. Moreover, by combining hierarchies (or contexts) from different sources, the UMLS Metathesaurus not only allows but also favors multiple inheritance. The UMLS Metathesaurus structure is thus compatible with the definition of a directed acyclic graph. UMLS concepts have unique identifiers and pairs of concept identifiers, associated by relationship qualifiers, which are used to describe the structure of the UMLS Metathesaurus.

Compared to that of any given source vocabulary, the context offered by the UMLS Metathesaurus is both broader and deeper. A broader context means that the ancestors of a concept are not necessarily constrained to any single particular representation of the world or ontology. A deeper context means that the granularity of the UMLS Metathesaurus is usually much finer than that of any source vocabulary.

Hierarchical relationships account for roughly half of the relationships represented in the UMLS Metathesaurus, excluding those, such as siblings, that are derived from other relationships. Some hierarchical relationships found in the UMLS Metathesaurus come from the source vocabularies. By convention, these relationships are called parent/child relationships. Even if these relationships were originally defined at the term level (i.e., among terms in a particular vocabulary), they are recorded at the concept level in the UMLS Metathesaurus, in the form of pairs of concept identifiers associated with a PAR (parent) or CHD (child) relationship type.

The UMLS Metathesaurus has another type of hierarchical relationship, called "broader in meaning" and "narrower in meaning," identified by the "RB" and "RN" relationship types. These hierarchical relationships differ from the former only by virtue of their origin. Instead of being inherited form the source vocabularies, the RB/RN relationships are added to the original structure using different methods. A relationship between two terms is first suggested by lexical analysis of the terms, refined through a facts database, and possibly reviewed by human editors (Sperzel et al., 1992). Equivalent strategies have been used outside the UMLS context to build SNOMED-RT (Campbell, Tuttle, & Spackman, 1998) or to merge overlapping terminologies such as SNOMED International and LOINC (Dolin et al., 1998). As with synonymy, hierarchical relationships can also be established by human editors in the absence of any common lexical features (e.g., the relationship of "Hypoadrenalism" to "Severe adrenal insufficiency").

Some of the RB/RN relationships are redundant with their PAR/CHD counterparts (e.g., the relationship of "Adrenal Gland Diseases" to "Adrenal Cortex Diseases" is recorded with both PAR and RB identifiers). However, allowing the term comparison process to be performed independently from the context of a given vocabulary permits the discovery of relationships among concepts coming from different sources that by definition cannot be inherited from the sources. For example, the ICD-10 term “Other disorders of adrenal gland” is considered narrower than the MeSH term “Adrenal Gland Diseases,” although “Adrenal Gland Diseases” does not appear in ICD-10 hierarchies nor does “Other disorders of adrenal gland” in MeSH’s.

Figure 1 provides the hierarchical context (ancestors and descendants) for “Addison’s Disease” in the UMLS Metathesaurus. Although for practical reasons only part of the context is represented, the graph demonstrates some of the following advantages of the UMLS Metathesaurus structure. The granularity in the UMLS Metathesaurus is finer than in any other source vocabulary. For example, the five-level C19 MeSH hierarchy for “Addison’s Disease” expands to ten levels in the UMLS. The structure also shows that an autoimmune disorder is only one possible causal mechanism for Addison’s disease by making “Addison’s disease due to autoimmunity” a child of “Addison’s Disease.” Finally, even the ICD-10 hierarchy, although comprising classification-specific terms with little meaning outside the classification itself (e.g., “Disorders of other endocrine glands”), is linked to meaningful concepts through relationships added by the Metathesaurus editors.

3.2 Nature of Hierarchical Relationships

Hierarchical relationships are based on subsumptive principles and include two major kinds of relationships (McCray & Nelson, 1995). Hyponymy (or the generic relation) is represented by the “isa” relation (is a kind of) or by “narrower than.” “X isa Y” means that X and Y share essential features (called genus), while X has some special feature(s) (called differentia) that makes it different from Y and from other hyponyms of X. The generic relation is transitive. Concepts such as diseases, findings, and procedures can be organized by a generic relation. Meronymy (or the partitive relation) is represented by the “part_of” relation, that is, the part to whole relation. The partitive relation is not necessarily transitive. Spatial, temporal, and functional concepts may be organized by a partitive relation.

Informally, a composite concept description can be subsumed to another one for any of the following reasons (Bernauer, 1994):

- Introduction of a specializing criterion to the base concept, or the generic refinement of a concept element;
- Introduction of a partitive criterion to the base concept, or the partitive refinement of a concept element; or
- Introduction of a conjunctive coordination to the base concept, or to a concept element.

For example, the UMLS hierarchy for “Aortic Aneurysm” (the dilatation of the aorta), is organized by “isa” relations. The actual subsumptive principle, however, is not explicit in the UMLS (fig. 2).

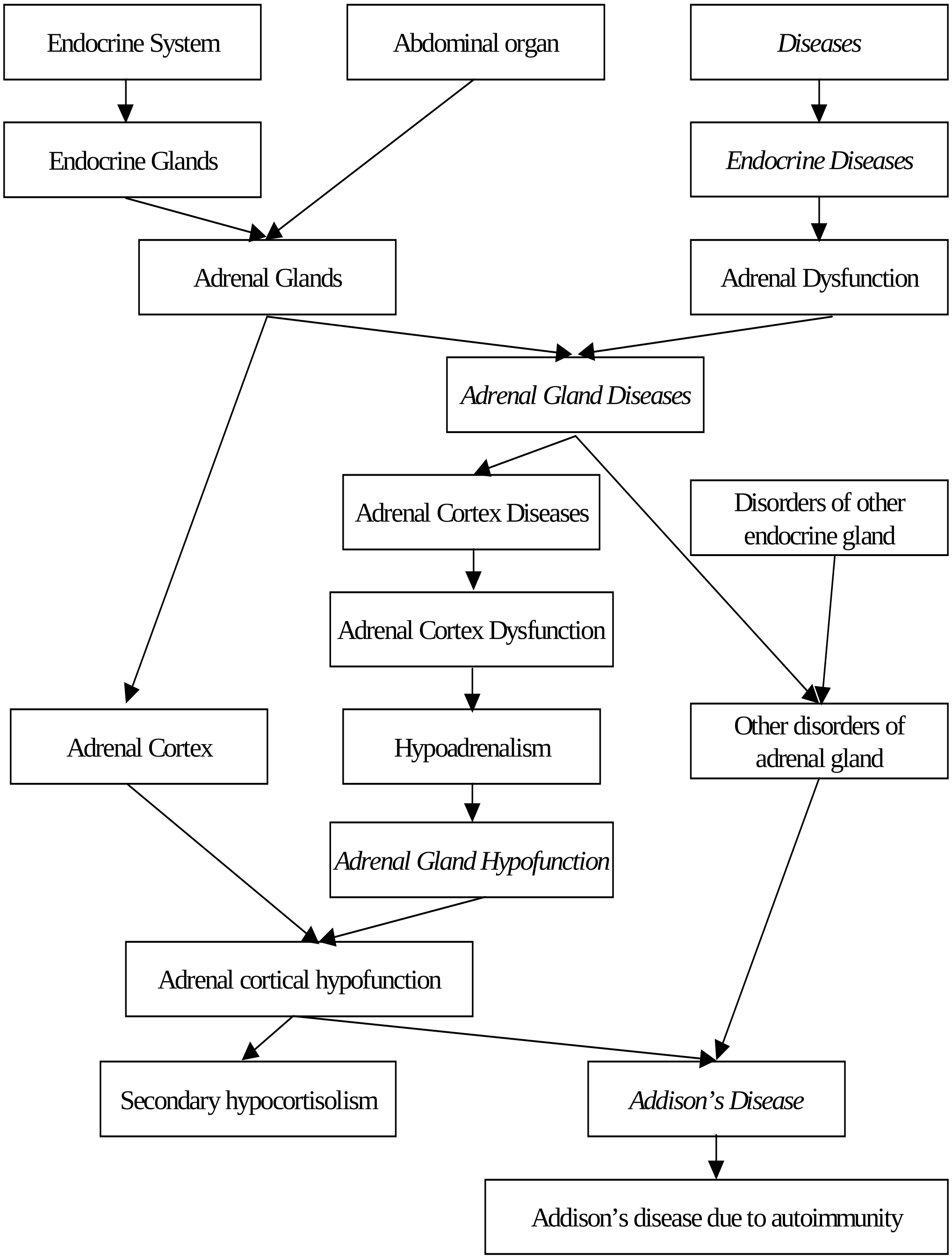
Endocrine System
Abdominal organ
Diseases
Endocrine Glands
Endocrine Diseases
Adrenal Glands
Adrenal Dysfunction
Adrenal Gland Diseases
Adrenal Cortex Diseases
Disorders of other endocrine gland
Adrenal Cortex Dysfunction
Adrenal Cortex
Hypoadrenalism
Other disorders of adrenal gland
Adrenal Gland Hypofunction
Adrenal cortical hypofunction
Secondary hypocortisolism
Addison's Disease
Addison's disease due to autoimmunity

Figure 1. UMLS context for "Addison's Disease" (partial).
Concepts in italics belong to the C19 MeSH hierarchy.

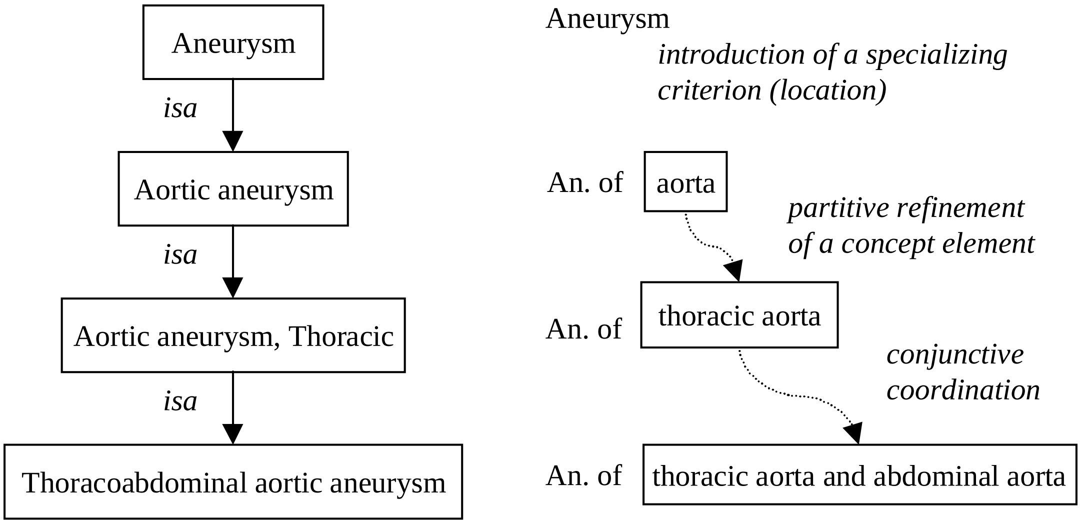
Aneurysm
isa
Aortic aneurysm
isa
Aortic aneurysm, Thoracic
isa
Thoracoabdominal aortic aneurysm
Aneurysm
introduction of a specializing criterion (location)
An. of
aorta
partitive refinement of a concept element
An. of
thoracic aorta
conjunctive coordination
An. of
thoracic aorta and abdominal aorta

Figure 2. Different principles of subsumption (right side) used in the "Aortic Aneurysm" isa hierarchy (left side)

3.3 Integration Issues Related to Hierarchy

In coding systems that were not designed to be used computationally, the criteria of subordination are usually hidden, and the nature of hierarchical relationships is often implicit. Moreover, organizing principles sometimes mix generic and partitive relations, for convenience and simplification. This presents serious limitations for computational usage of such knowledge structures, since it limits the potential for automatic subsumption (Bernauer, 1994).

3.3.1 Ontological Perspective

As noted by others (e.g., McCray & Nelson, 1995; Pisanelli, Gangemi, & Steve, 1998), and evidenced in the graph of the ancestors of "Addison's disease" (fig. 3), some of the hierarchical relationships in the Metathesaurus express neither hyponymy nor meronymy. For example, "Adrenal Cortex" is the location of "Adrenal cortical hypofunction," and the subsumed concept "Adrenal cortical hypofunction" is neither a specialization nor a part of "Adrenal Cortex." In the UMLS Metathesaurus, explicit "location_of" relationships are usually classified as non-hierarchical relationships. This particular implicit "location_of" relationship, however, is considered hierarchical by one source vocabulary and integrated as such in the UMLS Metathesaurus. Although clearly in the same semantic neighborhood as "Addison's disease," "Adrenal Cortex" cannot be considered an ancestor of "Addison's disease" from an ontological point of view.

In contrast, many hierarchical relationships whose nature is not made explicit are indeed true "isa" relationships and could contribute to defining an ontology for the biomedical domain from the UMLS or from some combination of its source vocabularies. For example, although unqualified in the UMLS, the relationship of “Addison’s disease due to autoimmunity” to “Addison’s Disease” is actually an “isa” relationship (specialization). Making it explicit would allow “Addison’s disease due to autoimmunity” to inherit properties from “Addison’s Disease.”

3.3.2 Granularity, Redundancy, and Simplification

Owing to differences in granularity among medical vocabularies, terms considered siblings in one vocabulary might be hierarchically related in a finer-grained vocabulary. This does not cause problems so long as the hierarchical relationships from different vocabularies are consistent, which is usually the case. For example, “Addison’s Disease” and “Addison’s disease due to autoimmunity” are two direct children of “Adrenal Gland Diseases” in SNOMED International, while other vocabularies provide a more detailed representation (fig. 3).

Using graph theory parlance, several possible paths exist from “Adrenal Gland Diseases” to “Addison’s Disease,” including a direct one provided by SNOMED International and an indirect one coming from MeSH. This redundancy, although useful for certain purposes, also makes it more difficult to process the knowledge, for example to visualize the concepts hierarchically related to a given concept. One solution to simplify the knowledge structure is to remove the relationships that can be inferred from other relationships by transitivity. Performed on graphs, this operation is known as transitive reduction. The graph representing the UMLS context for “Addison’s Disease” (fig. 1 shows 19 vertices connected by 20 edges; there were 46 edges in the graph prior to the transitive reduction).

3.3.3 Implicit Knowledge

As is the case for synonymy, the lack of fully specified terms in certain vocabularies can be a source of erroneous relationships. For example, the term “Infection” found as a child of “Pulmonary disorders” and parent of “Pneumonia” is unambiguously understood as “Lung infection” by a human reader, since it belongs to the “Pulmonary” chapter of COSTAR. For natural language processing tools, however, there is no reason to consider the term “Infection” in COSTAR differently from the same term in any other vocabulary or in another chapter of the same vocabulary.

Inter-concept relationships, particularly those discovered by lexical techniques, also suffer from problems of implicit knowledge. In this example, “Infection” incorrectly becomes a synonym of “Lung infection,” making “Pneumonia” a sibling of “Otitis media” (ear infection). Assuming that the parent/child relationships are “isa” relations, “Pneumonia isa Infection” remains true, whereas “Otitis isa Lung infection” does not.

3.3.4 Circular Hierarchical Relationships

As noted in other studies of inter-concept relationships in the UMLS Metathesaurus (e.g., Pisanelli, Gangemi, & Steve, 1998), the graph of UMLS concepts described by pairs of hierarchically related concepts contains cycles. In other words, some concepts happen to be both ancestors and descendants of themselves (loop), or of another concept (circular hierarchical relationship).

Most circular hierarchical relationships result from the way terms are integrated in the UMLS, rather than from conflicting organizations of the knowledge among medical vocabularies; that is, conflicts at the concept level come from relationships defined at the term level. Certain medical vocabularies use underspecified terms, containing qualifiers such as "unspecified" or "not otherwise specified," that are clustered by convention together with their specified equivalent in the UMLS. This results in loops if the two terms are in direct hierarchical dependence in one vocabulary, and in circular hierarchical relationships otherwise. Examples of loops include ICD-10 terms "Hodgkin's disease" (C81) and its child "Hodgkin's disease, unspecified" (C81.9), both names for the same UMLS concept C0019829. The following hierarchy extracted from the Clinical Terms Version 3 results in a direct circular hierarchical relationship when integrated in the UMLS. "Ligament reconstruction" is a parent of "Other reconstruction of ligament," which is a parent of "Reconstruction of ligament NOS." Here, the first and the last term in the hierarchy are clustered into the same UMLS concept.

Some cycles involve three concepts or more. Term ambiguity and the use of non-hierarchical relations in hierarchies (e.g., the relations between a disease and its symptoms, or the relations among chemical compounds) are responsible for a large number of these cycles. Other causes include inconsistencies among vocabularies, for example in the semantics of the "and" and "or" conjunctions, as also noted by other authors (Mendonca et al., 1998).

4. EXPLICIT MAPPING RELATIONSHIPS

Some medical vocabularies explicitly include mapping relationships in their structure. The target terms or concepts are either part of the vocabulary itself (internal mapping) or part of another vocabulary (external mapping). Although most mapping relationships come from or are endorsed by the developers of at least one of the vocabularies involved, some are provided by institutions unrelated to either vocabulary.

4.1 External Mapping Relationships

External mapping relationships have been developed for practical reasons: While there is no standard or common structure for medical vocabularies, there is a strong need for terms to be translated from one coding system to another one. Some vocabularies include mapping relationships to other vocabularies, allowing users to produce reports based on a mandatory coding system while using a more clinically oriented terminology instead (Read, Sanderson, & Drennan, 1995).

For example, the International Classification of Primary Care (ICPC), the Clinical Terms Version 3 (CTV3), SNOMED International, and GALEN provide mapping relationships to the International Classification of Diseases (ICD). CTV3 also provides mapping relationships to OPCS-4, the coding system used in the United Kingdom for procedures. Mapping relationships have also been established from one version of ICD to the next or to the previous one. More generally, major coding systems provide cross-references to other coding systems.

Here again, mapping relationships are seldom one-to-one relationships. More often, due to differences in structure and/or granularity between the source and target vocabularies, they are one-to-many or even many-to-many relationships, which makes them difficult to use in an automated coding process. For example, ICPC-2 code P74 ("Anxiety disorder / anxiety state") is mapped to several ICD-10 codes. The ICD terms mapped to (including "Panic disorder" and "Generalized anxiety disorder") are actually narrower than the ICPC-2 term mapped from. As a consequence, the ICPC-2 term can not be translated into an ICD term other than "Anxiety disorder, unspecified" without additional clinical information. For one-to-many mappings, CTV3 provides a list of potential matches in the target coding system and highlight the most likely, to be used as default.

These mapping relationships are often produced manually. GALEN, however, automatically maps terms from different sources, as soon as these terms have been mapped manually to GALEN.

4.2 Mapping Relationships in the UMLS

Among its 626,313 concepts, the UMLS Metathesaurus acknowledges 328,145 mapping relationships (i.e., relationships whose attribute is "mapped_to"). The major source (89%) of mapping relationships is the Medical Subject Headings (MeSH) whose mapping from supplementary concepts to Main Headings is fully preserved in the UMLS. SNOMED International provides an additional 6% of the Metathesaurus mapping relationships.

Although all of them bear the same "mapped_to" relationship attribute, distinctions can be seen among mapping relationships from different source vocabularies. Mapping relationships inherited from MeSH are considered hierarchical relationships. The source concept is considered subsumed by the target concept, which usually holds true since the granularity of the supplementary concepts tends to be finer than that of the main headings. SNOMED International provides the mapping of SNOMED concepts to ICD-9-CM concepts. In this case, the source and target concepts are considered near-synonyms, or at least very close in meaning. In some cases, the terms naming the SNOMED and the ICD concepts are true synonyms and belong to the same UMLS concept. Mapping relationships from other sources are considered "other relationships," meaning that their nature is not necessarily hierarchical and not further specified. Mapping relationships account for 31% of the total number of hierarchical relationships, 15% of the near-synonyms, and 6% of the "other relationships."

In addition to mapping relationships, the UMLS Metathesaurus also provides some

7,000 “associated expressions” (ATXs) to map terms, mostly from ICD-9-CM to MeSH. ATXs are created by human indexers from elementary concepts combined with both logical operators (i.e., AND, OR, NOT) and from relationships between MeSH Main Headings and subheadings. For example, the term “Mumps pancreatitis” has the associated expression “Mumps/complications AND Pancreatitis/etiology” in which the two MeSH main headings “Mumps” and “Pancreatitis” are qualified by a subheading.

Endnotes

1. A list of the medical vocabularies mentioned in this chapter is given in the Appendix.

2. Formerly called “Read Codes.”

3. Announced recently, the merging of SNOMED-RT and Clinical Terms Version 3 should create SNOMED-CT, a comprehensive language of health to support the computerized patient record.

Acknowledgments

We would like to thank Geneviève Botti, Anita Burgun, Stuart Nelson, Jeremy Rogers, and Pierre Zweigenbaum for thoughtful comments on earlier versions of this manuscript.

Appendix

Clinical Terms Version 3 (**CTV3**, formerly called “Read Codes”). England: National Health Service Centre for Coding and Classification, March, 1998. For information: <http://www.nhsccc.exec.nhs.uk> [2000, July 27].

Computer-Stored Ambulatory Records (**COSTAR**). Boston: Massachusetts General Hospital, 1995.

Physicians’ Current Procedural Terminology (**CPT**). 4th ed. Chicago: American Medical Association, 1999. For information: <http://www.ama-assn.org/med-sci/cpt/coding.htm> [2000, July 27].

Generalised Architecture for Languages, Encyclopaedias, and Nomenclatures in medicine (**GALEN**). Manchester, Eng.: *Open*GALEN. For information: <http://www.opengalen.org> [2000, July 27].

International Classification of Diseases: 9th revision, Clinical Modification (**ICD-9-CM**). 6th ed. Washington, DC: Health Care Financing Administration, July, 1998. For information: <http://www.hcfa.gov/stats/pufiles.htm> [2000, July 27].

International Statistical Classification of Diseases and Related Health Problems (**ICD-10**). 10th rev. Geneva World Health Organization, 1998. For information: <http://www.who.int/whosis/icd10/index.html> [2000, July 27].

International Classification of Primary Care (**ICPC**). Denmark: World Organisation of

Family Doctors, 1993. For information: <http://www.wonca.org/wonca_home.htm> [2000, July 27].

Logical Observation Identifiers, Names and Codes (**LOINC**). Version 1.0j. Indianapolis: The Regenstrief Institute, 1997. For information: <http://www.mcis.duke.edu/standards/termcode/loinc.htm> [2000, July 27].

Medical Subject Headings (**MeSH**). Bethesda, MD: National Library of Medicine, 1999. For information: <http://www.nlm.nih.gov/mesh/meshhome.html> [2000, July 27].

Physician Data Query Online System (**PDQ**). Bethesda, MD: National Cancer Institute, August, 1998. For information: <http://cancernet.nci.nih.gov/pdqfull.html> [2000, July 27].

Systematized Nomenclature of Human and Veterinary Medicine: **SNOMED** International. Version 3.5. Northfield, IL: College of American Pathologists; Schaumburg, IL: American Veterinary Medical Association, 1998. For information: <http://www.snomed.org> [2000, July 27].

Systematized Nomenclature of Human and Veterinary Medicine-Reference Terminology: **SNOMED-RT**. Northfield, IL: College of American Pathologists. For information: <http://www.snomed.org> [2000, July 27].

Unified Medical Language System (**UMLS**). Bethesda (MD): National Library of Medicine, 1999. For information: <http://www.nlm.nih.gov/pubs/factsheets/umls.html> [2000, July 27].

References

Bernauer, J. (1994). Subsumption principles underlying medical concept systems and their formal reconstruction. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, 140-144.

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(1), 76-87.

Bodenreider, O., Nelson, S. J., Hole, W. T., & Chang, H. F. (1998). Beyond synonymy: Exploiting the UMLS semantics in mapping vocabularies. *Proceedings of the 1998 AMIA Annual Fall Symposium*, 815-819.

Campbell, J. R., Carpenter, P., Sneiderman, C., Cohn, S., Chute, C. G., & Warren, J. (1997). Phase II evaluation of clinical coding schemes: Completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. *Journal of the American Medical Informatics Association*, 4(3), 238-251.

Campbell, K. E., Oliver, D. E., Spackman, K. A., & Shortliffe, E. H. (1998). Representing thoughts, words, and things in the UMLS. *Journal of the American Medical Informatics Association*, 5(5), 421-431.

Campbell, K. E., Tuttle, M. S., & Spackman, K. A. (1998). A "lexically-suggested logical closure" metric for medical terminology maturity. *Proceedings of the 1998 AMIA Annual Fall Symposium*, 785-789.

Chute, C. G., Cohn, S. P., & Campbell, J. R. (1998). A framework for comprehensive health terminology systems in the United States: Development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. *Journal of the American Medical Informatics Association*, 5(6), 503-510.

Chute, C. G., Cohn, S. P., Campbell, K. E., Oliver, D. E., & Campbell, J. R. (1996). The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *Journal of the American Medical Informatics Association,* 3(3), 224-233.

Cimino, J. J. (1996). Review paper: Coding systems in health care. *Methods of Information in Medicine,* 35, 273-284.

Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine,* 37, 394-403.

Cimino, J. J., & Clayton, P. D. (1994). Coping with changing controlled vocabularies. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care,* 135-139.

Cimino, J. J., Johnson, S. B., Peng, P., & Aguirre, A. (1993). From ICD9-CM to MeSH using the UMLS: A how-to guide. *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care*, 730-734.

Dessena, S., Rossi Mori, A., & Galeazzi, E. (1999). Development of a cross-thesaurus with Internet-based refinement supported by UMLS. *International Journal of Medical Informatics,* 53, 29-41.

Dolin, R. H., Huff, S. M., Rocha, R. A., Spackman, K. A., & Campbell, K. E. (1998). Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. *Journal of the American Medical Informatics Association,* 5(2), 203-213.

Evans, D. A., Cimino, J. J., Hersh, W. R., Huff, S. M., & Bell, D. S. (1994). Toward a medical-concept representation language. The Canon Group. *Journal of the American Medical Informatics Association,* 1(3), 207-217.

Humphreys, B. L., Lindberg, D. A., Schoolman, H. M., & Barnett, G. O. (1998). The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1), 1-11.

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. (1998). The nature of lexical knowledge. *Methods of Information in Medicine,* 37(4-5), 353-360.

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., 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.

Mendonca, E. A., Cimino, J. J., Campbell, K. E., & Spackman, K. A. (1998). Reproducibility of interpreting "and" and "or" in terminology systems. *Proceedings of the 1998 AMIA Annual Fall Symposium*, 790-794.

Nelson, S. J., Fuller, L. F., Erlbaum, M. S., Tuttle, M. S., Sherertz, D. D., & Olson, N. E. (1992). The semantic structure of the UMLS Metathesaurus. *Proceedings of the 16th*

Annual Symposium on Computer Applications in Medical Care, 649-653.

Oliver, D. E., Shahar, Y., Shortliffe, E. H., & Musen, M. A. (1999). Representation of change in controlled medical terminologies. *Artificial Intelligence in Medicine,* 15, 53-76.

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

Rada, R., Ghaoui, C., Russell, J., & Taylor, M. (1993). Approaches to the construction of a medical informatics glossary and thesaurus. *Medical Informatics (London),* 18, 69-78.

Rassinoux, A. M., Miller, R. A., Baud, R. H., & Scherrer, J. R. (1997). Compositional and enumerative designs for medical language representation. *Proceedings of the 1998 AMIA Annual Fall Symposium,* 620-624.

Read, J. D., Sanderson, H. F., & Drennan, Y. M. (1995). Terming, encoding, and grouping. *Medinfo,* 8 (Pt 1), 56-59.

Rector, A., Rossi Mori, A., Consorti, M. F., & Zanstra, P. (1998). Practical development of re-usable terminologies: GALEN-IN-USE and the GALEN Organisation. *International Journal of Medical Informatics,* 48, 71-84.

Rector, A. L., Bechhofer, S., Goble, C. A., Horrocks, I., Nowlan, W. A., & Solomon, W. D. (1997). The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine,* 9, 139-171.

Rector, A. L., & Nowlan, W. A. (1994). The GALEN project. *Computer Methods and Programs in Biomedicine,* 45, 75-78.

Rector, A. L., Solomon, W. D., Nowlan, W. A., Rush, T. W., Zanstra, P. E., & Claassen, W. M. (1995). A terminology server for medical language and medical information systems. *Methods of Information in Medicine,* 34, 147-157.

Rossi Mori, A., Bernauer, J., Pakarinen, V., Rector, A. L., De Vries-Robbe, P., Ceusters, W., Hurlen, P., Ogonowski, A., & Olesen, H. (1993). Models for representation of terminologies and coding systems in medicine. *Studies in Health Technology and Informatics,* 6, 92-104.

Rossi Mori, A., Consorti, F., & Galeazzi, E. (1998). Standards to support development of terminological systems for healthcare telematics. *Methods of Information in Medicine,* 37, 551-563.

Sherertz, D. D., Olson, N. E., Tuttle, M. S., & Erlbaum, M. S. (1990). Source inversion and matching in the UMLS Metathesaurus. *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care,* 141-145.

Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Spackman, K. A., Campbell, K. E., & Cote, R. A. (1997). SNOMED RT: A reference terminology for health care. *Proceedings of the 1997 AMIA Annual Fall Symposium,* 640-644.

Sperzel, W. D., Tuttle, M. S., Olson, N. E., Erlbaum, M. S., Saurez-Munist, O., Sherertz, D. D., & Fuller, L. F. (1992). The Meta-1.2 engine: A refined strategy for linking biomedical vocabularies. *Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care,* 304-308.

Suarez-Munist, O. N., Tuttle, M. S., Olson, N. E., Erlbaum, M. S., Sherertz, D. D., Lipow, S. S., Cole, W. G., Keck, K. D., & Davis, A. N. (1996). MEME-II supports the cooperative management of terminology. *Proceedings of the 1998 AMIA Annual Fall Symposium*, 84-88.

Tuttle, M. S., Olson, N. E., Campbell, K. E., Sherertz, D. D., Nelson, S. J., & Cole, W. G. (1994). Formal properties of the Metathesaurus. *Proceedings of the 18th Annual Symposium on Compute*r *Applications in Medical Care,* 145-149.

Tuttle, M. S., Suarez-Munist, O. N., Olson, N. E., Sherertz, D. D., Sperzel, W. D., Erlbaum, M. S., Fuller, L. F., Hole, W. T., Nelson, S. J., Cole, W. G., & et al. (1995). Merging terminologies. *Medinfo,* 8 (Pt 1), 162-166.

Volot, F., Joubert, M., & Fieschi, M. (1998). Review of biomedical knowledge and data representation with conceptual graphs. *Methods of Information in Medicine,* 37, 86-96.