Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention

Olivier Bodenreider, M.D., Ph.D.
National Library of Medicine, Bethesda, Maryland
olivier@nlm.nih.gov

The Unified Medical Language System (UMLS) is a large repository of some 800,000 concepts for the biomedical domain, organized by several millions of inter-concept relationships, either inherited from the source vocabularies, or specifically generated. This paper focuses on hierarchical relationships in the UMLS Metathesaurus, and especially, on circular hierarchical relationships.
Using the metaphor of a disease, we first analyze the causal mechanisms for circular hierarchical relationships. Then, we discuss methods to identify and remove these relationships. Finally, we briefly discuss the consequences of these relationships for applications based on the UMLS, and we propose some prevention measures.

INTRODUCTION

The Unified Medical Language System® (UMLS®) is intended to help health professionals and researchers use biomedical information from different sources [1]. While the structure of each source vocabulary is preserved, terms which are equivalent in meaning are clustered into a unique concept. Furthermore, interconcept relationships, either inherited from the source vocabularies or specifically generated, give the UMLS Metathesaurus additional semantic structure.

The UMLS building process imposes no restrictions on the source vocabularies prior to integrating their terms and structure into the Metathesaurus. In the source vocabularies, hierarchical relationships are usually not limited to taxonomic relations, but rather reflect the way each vocabulary organizes its terms, according to its purposes. For example, vocabularies that focus on knowledge representation (e.g., the University of Washington Digital Anatomist Symbolic Knowledge Base) use separate hierarchies to represent taxonomic ('is a') and meronymic ('part of') relations, while most vocabularies allow, often implicitly, several types of relationships to be used in hierarchies (relations such as 'manifestation of', for example, are sometimes used in addition to 'is a' and 'part of'). Therefore, hierarchical relationships in the Metathesaurus are not expected to represent homogeneous relations, but rather to reflect several organizational principles inherited from the source vocabularies. Moreover, the precise nature of the relationship is mentioned in only about 25% of the cases; and, because many non-taxonomic relations are used to build hierarchies, it is not possible to assume that a non-labeled hierarchical relationship is probably taxonomic.

Even though they are heterogeneous, the organizational principles used to create hierarchies are expected to share some fundamental characteristics, and, thus, to be compatible. One of these characteristics is *antisymmetry*, one of the properties of the order relation, the mathematical counterpart of hierarchy. Since the hierarchical relation between concepts C_1 and C_2 is antisymmetric, the only possibility for having both C_1 parent of C_2 and C_2 parent of C_1 is that C_1 and C_2 are actually the same concept.

Polyhierarchy refers to the situation in which a concept can have multiple parents. Some vocabularies such as MeSH or Clinical Terms Version 3 (formerly Read Codes) use polyhierarchy as their organizational structure. In the Metathesaurus, polyhierarchical structure results either from such vocabularies or from the combination of multiple single-heritance hierarchies inherited from other source vocabularies. The resulting data structure is called a *directed acyclic graph* (DAG). Concepts are the vertices of the graph, while inter-concept relationships are its edges. Assuming that one direction is arbitrarily selected to represent hierarchy (e.g., 'parent of', not 'child of'), the resulting graph is directed. Assuming further that hierarchical principles used across vocabularies are compatible, no concept C_1 designated as an ancestor (direct or indirect parent) of C_2 in one vocabulary can be a descendant (direct or indirect child) of the same concept C_2 in another vocabulary. In other words, ideally, no circular relationship should result from combining hierarchies, and the resulting graph should be acyclic.

The order relation associated with hierarchies is a *partial* order relation, which means that it is possible for a concept to be hierarchically related to itself (reflexive relation). In a directed acyclic graph, however, no path is allowed to start and end at the same vertex, which means that, when represented in a graph, the reflexive hierarchical relationships create cycles of a particular kind, called loops. In this paper, we will make no distinction among circular

Proceedings of AMIA Annual Symposium 2001;:57-61.

hierarchical relationships on the basis of the number of concepts involved in the cycles, because any cycle has similar detrimental consequences in terms of graph traversal [2].

In fact, circular hierarchical relationships have existed in the UMLS Metathesaurus for quite a long time. Virtually any evaluative study of the Metathesaurus with a focus on relationships mentions them [e.g., 3, 4, 5]. Numerous papers also insist on the necessity for medical vocabularies to include only acyclic relationships [e.g., 6].

The relationships discussed in this paper come from the12th edition (2001) of the UMLS [7]. Although all Metathesaurus hierarchical relationships can be found in the MRREL file, two kinds of hierarchical relationships are recorded separately, differentiated by their origin. Relationships inherited from source vocabularies are called "parent / child" and have the types 'PAR' and 'CHD' in the MRREL file, while the other hierarchical relationships, generated during the Metathesaurus building process, are called 'broader than / narrower than' and have the types 'RB' and 'RN'. Since there is no major semantic distinction between these two kinds of relationships, we group them together ('PAR' with 'RB', and 'CHD' with 'RN'), and refer to them simply as hierarchical relationships.

This paper presents the phenomenon of circular hierarchical relationships in the UMLS as if it were a (chronic) disease. After analyzing their causal mechanisms, we discuss methods to identify and remove the circular hierarchical relationships. The last part focuses on the consequences and some prevention measures.

ETIOLOGY

One notion is fundamental to help understand the causal mechanisms for circular hierarchical relationships in the UMLS Metathesaurus: although recorded and used *at the concept level*, many hierarchical relationships were defined *at the term level*. In other words, the clustering of synonymous terms into concepts modifies the original structure of the vocabularies. While this process produces a useful, unified polyhierarchical structure, circular hierarchical relationships can be seen as its side-effect. We have identified the following factors as causes for circular hierarchical relationships.

Granularity. When the level of granularity is higher in a given vocabulary than in the UMLS, two terms represented in a close hierarchical relationship (or micro-relation [8]) in a given vocabulary can be clustered together into a unique concept in the Metathesaurus. For example, the Clinical Terms Version 3 vocabulary considers "Actinomycotic mycetoma" broader in meaning than "Madura foot - actinomycotic" which represents the most common location of this infection. The two terms, however, are clustered into the same UMLS concept, and the relationship between the two terms becomes a reflexive relationship from this concept to itself (Figure 1).

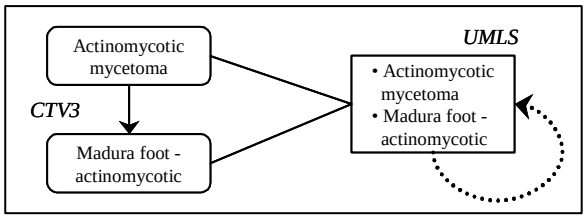

Figure 1 – Reflexive relationship in the UMLS Metathesaurus.

Unspecified terms. 62,000 UMLS terms bear some kind of underspecification marker, the most frequent being "not otherwise specified" or "NOS". In most source vocabularies, "T, NOS" is a child of "T". Although created for terminological purposes, the meaning of this unspecified term is not different from that of the equivalent term without the markers. Thus, "T" and "T, NOS" are clustered into the same UMLS concept, creating a reflexive relationship. Examples of such pairs of terms, found in the same vocabulary, include "Cellulitis" (L03) and "Cellulitis, unspecified" (L03.9), in ICD-10; and "Fracture of humerus" (S22..) and "Fracture of humerus NOS" (S22z.) in CTV3. The term "T" and its unspecified variant can also be found in two distinct vocabularies and related through another term common to the two vocabularies. For example, in MeSH, "Fever" is parent of "Fever of unknown origin", which is itself parent of "Fever, unspecified" in ICD-10. At the concept level, when "Fever" and "Fever, unspecified" are clustered together, this concept appears both parent and child of "Fever of unknown origin" (Figure 2).

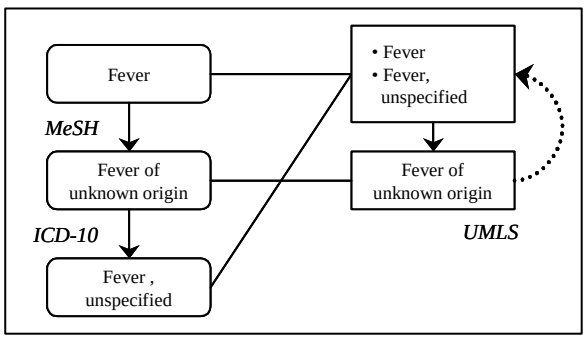

Figure 2 – Direct circular hierarchical relationship in the UMLS Metathesaurus.

Metadata. In some vocabularies, the same term may appear at different levels in the same hierarchy. In certain cases, one occurrence of the term contains some metadata such as ": GENERAL TERMS" (section header). In other cases, the only difference between the two terms, if any, is case variation. In any event, both terms have the same meaning.
Generally, the higher level term corresponds to some sort of header or section name, while the lower one represents the actual term. For example, the International Classification of Primary Care version 2 (ICPC-2), "HYDROCOELE" (Y86) designates the chapter, while "Hydrocele" (Y86001) is the terminal node that can be used for coding. In SNOMED International, "HEART DISEASES" and "HEART DISEASES: GENERAL TERMS" both correspond to some subdivision of the vocabulary.

Compound terms. As noted by Mendonça et al., terms that contain the conjunctions "and" and "or" do not have a consistent meaning across vocabularies [9]. For example, "nausea and vomiting" may be understood as "nausea with vomiting" (inheriting from both "nausea" and "vomiting") or "nausea or vomiting" (having "nausea" and "vomiting" as its children). In this case, "nausea" may be recorded both as parent and child of "nausea" and "vomiting", leading to a circular hierarchical relationship. A variant of this phenomenon appears with neoclassical compounds. Here, one word, not the term, exhibits the composition. The consequences, however, are similar. For example, a "colorectal neoplasm" is a neoplasm located in either the colon or rectum or both, while an "encephalomyelitis" is an inflammation involving both the brain (encephalitis) and the spinal cord (myelitis).

Classes, instances and implicit knowledge. In most instances, inflectional variation of terms does not modify the meaning (e.g., singular, plural). Therefore, the several inflectional variants of a term are considered synonymous and clustered into the same concept. In some cases, however, the plural form refers to a class, while the singular form refers to an instance, but not necessarily of the same class. For example "purine" is a heterocyclic compound that contribute to produce "purines" (the purine bases). There are two distinct concepts in the UMLS for "purine" and "purines". On the other hand, the terms "Topographic regions" and "body region" are considered synonymous in the UMLS, although "Topographic regions" in SNOMED International actually groups a whole range of physical anatomical entities, including "body regions". Even if the terms "Topographic regions" and "body region" are synonymous, "Topographic regions" has a different meaning in the particular context of SNOMED International (Figure 3). The implicit knowledge associated with a term used in a particular context is difficult to detect and is often not recognized.

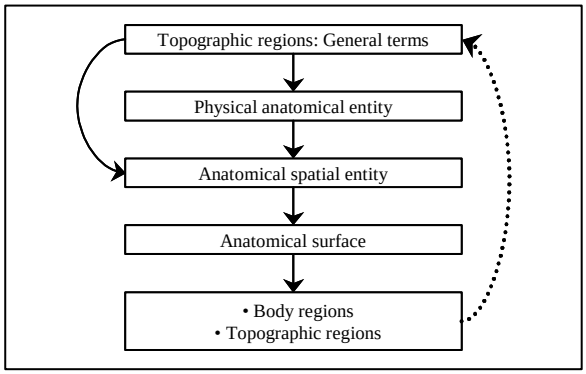

Figure 3 – Indirect circular hierarchical relationship in the UMLS Metathesaurus.

Organizational conventions. In some cases, for convenience, concepts are somewhat arbitrarily organized in hierarchies, although the nature of their relation is not truly hierarchical. Chemical compounds represent one such example. Salts (an acid and a base) and esters (an acid and an alcohol) are sometimes presented as either 'parents' or 'children' of the acid, depending on organizational conventions (e.g., "citric acid" and "citrates"). A similar problem occurs with the relationships between clinical drugs and their active ingredients (e.g., "chloramphenicol product" and "chloramphenicol"). When terms are integrated in the UMLS, conflicting conventions may result in the creation of circular hierarchical relationships.

Idiopathic. Finally, there are many cases for which no obvious cause can be detected. Sometimes, the relationship found in a given vocabulary seems wrong (e.g., "Bladder abnormality" parent of "Urinary tract disorder"). Often, none of the conflicting relationships are really hierarchical (e.g., relationship between "ecology" and "environment"). We will call these cases *idiopathic*, at least until we acquire a better understanding of their origin.

DIAGNOSIS

In the UMLS, information about hierarchical relationships can be found in two files: MRREL and MRCXT. MRREL is the only file in which all relationships are recorded. MRREL can always be used for identifying circular hierarchical relationships, but, in certain cases, other information sources provide better diagnostic solutions. For some vocabularies, the UMLS records interconcept hierarchical relationships within the vocabulary as a context (MRCXT file), in addition to the relationships recorded in MRREL. Such contexts are

often used for display purposes, but can be used for tracking circular hierarchical relationships as well. Since concept identifiers are recorded in the contexts, the presence of the same identifier more than once in a given context reveals a cycle. If the same identifier is found on two consecutive lines, the relationship is reflexive (Figure 1). If both the parent and one child of a given concept have the same identifier, this is a direct circular hierarchical relationship (Figure 2). Otherwise, the circular hierarchical relationship is indirect (Figure 3).

Reflexive relationships. Reflexive hierarchical relationships are easy to diagnose. In the MRREL file, the same concept identifier is both the source and the target of one of the relationship types that represent hierarchies (PAR, RB, CHD, RN). There are some 13,000 reflexive hierarchical relationships in the UMLS.

Differential diagnosis: some reflexive relationships do not involve hierarchical relationships, but associative relationships instead (e.g., mapping relationships between "Hydrocortisone" and "Cortisol", two terms from the same concept). In this case, the relationship type is 'RO'.

Direct relationships. Direct circular hierarchical relationships can also be diagnosed easily. For a given pair of concept (C_1, C_2), there exists a hierarchical relationship from both C_1 to C_2 and C_2 to C_1. In the MRREL file, the same pair of concept identifiers appears twice: once associated with a relationship type representing higher granularity (PAR, RB), and once with a relationship type representing lower granularity (CHD, RN). There are 1800 direct circular hierarchical relationships in the UMLS.

Indirect relationships. Circular hierarchical relationships may involve more than two concepts. In this case, no method allows for identifying the circular relationships by simply parsing MRREL. What needs to be done instead is to represent the hierarchical relationships in a graph data structure, and perform an operation on the graph to detect and locate the cycles. The UMLS is actually too big (800,000 nodes) to be represented entirely as a graph. Moreover, only a small number of cycles are known to persist once the reflexive and direct circular relationships have been removed. A simpler method is to build the graph of the ancestors for each concept successively. A given concept C participates in a cycle if and only if one of its ancestors A_b sends a back-edge to the source concept, which means that A_b is also its direct descendant of C. Analyzing the path (or, possibly, multiple paths) between the C and A_b reveals one or more cycles. For example, as shown in figure 3, the concept "Topographic regions: General terms", ancestor of "Body regions", sends a back-edge to "Body regions". However, another drawing of the same graph could represent "Topographic regions: General terms" as a direct descendant of "Body regions". There are some 120 indirect circular hierarchical relationships in the UMLS.

TREATMENT

As shown in the Etiology section, circular hierarchical relationships may have various causes, and their cure requires using a technique adapted to the causal mechanism. As usual in medicine, several general principles apply: "Primum non nocere" and, at least for HMOs, "Cost control". Therefore, the general strategy consists of treating first the relationships that can be easily identified and whose ablation will have no major consequence on the semantic structure of the Metathesaurus, i.e. the reflexive relationships. Conversely, indirect circular hierarchical relationships will be treated last, because their identification requires building costly graphs of ancestors and their removal can not easily be automated.

Reflexive relationships. Reflexive relationships in the Metathesaurus, and especially the hierarchical ones, are of no use. Therefore, MRREL lines in which the source (CUI1) and the target (CUI2) of the relationship contain the same concept identifier can be ignored or safely removed.

Cycles within contexts. The contexts recorded in MRCXT describe hierarchies of terms in a given vocabulary, along with the concepts to which these terms are associated in the Metathesaurus. If several terms at different levels of a hierarchy are associated with the same UMLS concept, this creates a circular hierarchical relationship. Breaking this cycle consists of detaching from the lowest level of the hierarchy the term whose associated concept appears more than once in the context. For example, let us assume that, in Figure 2, the three terms on the left belong to the same hierarchy. The top and bottom terms are associated with the same UMLS concept. Removing the relationship between the middle term and the bottom term on the left will detach the bottom term from the hierarchy, and, thus, remove the dotted edge between the two UMLS concepts on the right. The cycle is broken.

Other Direct relationships. The treatment of direct circular hierarchical relationships uses several steps, applied in order of increasing aggressiveness. The process stops at the first step that succeeds. We first use a possible redundancy in the number of relationship types for each direction. Then we remove C_1 from the ancestors of C_2, and C_2 from the ancestors of C_1, unless C_1 or C_2 have no other direct

ancestor and become orphan. Finally, we use a possible redundancy in the number of relationship sources for each direction. The MRREL file contains the information needed in these 3 steps.

Let us consider the following cycle $C_1 \leftrightarrow C_2$.

1. If the $C_1 \rightarrow C_2$ 'parent of' relationship is supported by a $C_1 \rightarrow C_2$ 'broader than' relationship, while there is only one type ('parent of' or 'broader than') for the $C_2 \rightarrow C_1$ relationship, $C_2 \rightarrow C_1$ is removed. If $C_2 \rightarrow C_1$ has two types of hierarchical relationships while $C_1 \rightarrow C_2$ has only one, $C_1 \rightarrow C_2$ is removed. The next step is used only if there is only one type of relationship for each direction.
2. If C_1 has direct ancestors other than C_2, remove $C_2 \rightarrow C_1$, and if C_2 has direct ancestors other than C_1, remove $C_1 \rightarrow C_2$. The next step is used only if both C_1 and C_2 have only one direct ancestor each.
3. Count the number of sources for the relationships 'parent of' and 'broader than' (added together), for $C_1 \rightarrow C_2$, on the one hand (r_{12}), and for $C_2 \rightarrow C_1$ on the other hand (r_{21}). If $r_{12} > r_{21}$, $C_2 \rightarrow C_1$ is removed. If $r_{21} > r_{12}$, $C_1 \rightarrow C_2$ is removed.

Other possible methods for selecting accurate relationships in circular hierarchical relationships include validating the relationships against the Semantic Network, and taking advantage of hyponymic relations detected by lexical techniques. Step 2 is intentionally aggressive, with a significant risk of removing accurate relationships. At a broader level, however, this should not be really detrimental since each concept stays connected to other higher-level concepts in the graphs.

Other Indirect relationships. The treatment of indirect circular hierarchical relationships requires a manual review of all inter-concept relationships involved in the cycle. No useful pattern was identified during our review.

COMPLICATIONS

Certain operations on graphs, such as transitive reduction, cannot be performed unless the graph is acyclic. Transitive reduction consists of removing a direct relationship $C_1 \rightarrow C_3$ when there exists two relationships $C_1 \rightarrow C_2$ and $C_2 \rightarrow C_3$. This operation, used to simplify the graph (e.g., for visualization purposes), would remove the direct relationship between "Topographic regions: General terms" and "Anatomical spatial entity" in Figure 3.

Although the number of cycles is relatively small, the number of concepts having a cycle in the graph of their ancestors is astonishingly large. Not surprisingly, we started this work in the framework of an application in which the graph of the ancestors of a concept is used to discover the MeSH descriptors the most closely associated with this concept.

PREVENTION

Ideally, hierarchies would be restricted to the taxonomic relation, or, at least, hierarchies using other organizing principles (e.g., 'part of') would be kept separate. Since the UMLS does not censor any information provided by the source vocabularies, hierarchical relationships should be tested against the ontological reference provided by the Semantic Network. As suggested by Cimino [10], this could be done by comparing hierarchical inter-concept relationships to the semantic relations defined between the corresponding semantic types in the Semantic Network. The precise nature of hierarchical relationships should also be made explicit for all relationships.

References

1. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281-91.
2. Aho AV, Ullman JD. The graph data model. In: Fundations of computer science. New York: Computer Science Press; 1992.
3. Bodenreider O, Burgun A, Botti G, Fieschi M, Le Beux P, Kohler F. Evaluation of the Unified Medical Language System as a medical knowledge source. J Am Med Inform Assoc 1998;5(1):76-87.
4. Bodenreider O, Bean CA. Relationships among knowledge structures: Vocabulary integration within a subject domain. In: Bean CA, Green R, editors. Relationships in the organization of knowledge: Kluwer; 2001. p. 81-98.
5. Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS Methathesaurus. Proc AMIA Symp 1998:810-4.
6. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.
7. UMLS. UMLS Knowledge Sources. 12th ed. Bethesda (MD): National Library of Medicine; 2001.
8. Cruse DA. Lexical semantics. Cambridge; New York: Cambridge University Press; 1986.
9. Mendonça EA, Cimino JJ, Campbell KE, Spackman KA. Reproducibility of interpreting "and" and "or" in terminology systems. Proc AMIA Symp 1998:790-4.
10. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51.