Approaches to Eliminating Cycles in the UMLS Metathesaurus: Naïve vs. Formal

Fleur Mougin[a], Olivier Bodenreider[b], M.D.,Ph.D.
[a]EA 3888 - IFR 140, Faculté de Médecine, Université de Rennes I, 35033 Rennes, France
[b]National Library of Medicine, Bethesda, Maryland
fleur.mougin@univ-rennes1.fr, olivier@nlm.nih.gov

Applications exploiting the hierarchical relations recorded in the Unified Medical Language System (UMLS) Metathesaurus suffer from the presence of inconsistencies in these relations. A formal approach to identifying and eliminating circular hierarchical relations has been proposed in previous work, leading to the creation of a directed acyclic Metathesaurus graph. However, this approach is at best semi-automatic and its implementation is far from trivial. A simpler, alternative approach consists in avoiding loops while traversing the Metathesaurus graph by preventing nodes from being visited twice. Our objective is to evaluate the benefit of the formal approach to eliminating cycles over a naïve approach to avoiding them. To this end, we compared the size and semantic coherence of sets of descendants obtained by both approaches. 12% of the concepts with descendants exhibit some differences. The formal approach significantly reduces the number of descendants in these cases. The benefits in terms of semantic coherence are more subtle.

INTRODUCTION

Biomedical terminologies are used as a source of knowledge in many applications. More specifically, the hierarchical relations represented in terminologies (*parent*/*child*, *broader*/*narrower than*) provide surrogate subsumption relations (*isa*, *subclass of*). Most terminologies are organized into hierarchies. In terms of data structure, single-inheritance hierarchies are trees and multiple-inheritance hierarchies are directed graphs. Such hierarchical structures can be traversed easily, making it possible for users to find paths among concepts and to compute transitive closures.

By integrating more than 100 vocabularies into a unique semantic space, the Unified Medical Language System® (UMLS®) Metathesaurus® creates a large graph with over one million nodes (concepts) and nearly 5 million hierarchical relations. Because it results from the integration of hierarchical structures representing partial ordering relations, the Metathesaurus graph is expected to be acyclic: no descendant of a concept should be simultaneously an ancestor of this concept. In practice, however, the presence of cycles in the Metathesaurus graph has been identified and recognized as an issue by many researchers [1-3]. There are two major causes for cycles. First, some of the source vocabularies are not acyclic graphs themselves. Once integrated in the Metathesaurus, the cycles existing in these vocabularies become cycles in the Metathesaurus graph. Second, the integration process sometimes creates cycles by allowing conflicting views to coexist in the same structure.

Issues with hierarchical relations in the UMLS Metathesaurus have been studied in [4] from a theoretical perspective and solutions have been proposed for eliminating inappropriate links causing circular hierarchical relations. However, the algorithm proposed is relatively complex, requires some manual intervention and is difficult to implement in UMLS-based applications. A simpler, naïve approach consists in preventing loops in traversing the graph of Metathesaurus concepts by keeping track of the nodes already visited. The disadvantage of this approach, however, is that, while removing *structurally* offending links, it does not discriminate between links *semantically*. In other words, this approach cannot ensure that the links ignored during the graph traversal in order to prevent loops from happening are actually the appropriate links to be removed: which link will be ignored solely depends on the order in which the graph is traversed.

The objective of this study is to evaluate the practical benefit of using the formal approach to eliminating circular hierarchical relations in the UMLS Metathesaurus, compared to the naïve approach. To this end, we compared the size and semantic coherence of sets of descendants obtained by both approaches for all concepts in the Metathesaurus. Our hypothesis is that the formal approach will result in fewer descendants and that the semantic coherence of sets of descendants will be greater, compared to the naïve approach.

BACKGROUND

Resource

The Unified Medical Language System (UMLS) includes two sources of semantic information: the Metathesaurus and the Semantic Network. Several types of relationships among concepts are recorded in the Metathesaurus: *parent*/*child* and *broader*/*narrow-*

wer than essentially correspond to hierarchical relations, while the other relationships are associative. Approximately 5 million hierarchical relations are represented in the Metathesaurus. Compared to its constituent vocabularies taken individually, the Metathesaurus has not only a broader scope, but also a higher level of granularity.

The Semantic Network is a much smaller network of 135 semantic types organized in a tree structure. The semantic types have been aggregated into fifteen coarser semantic groups [5]. Each concept from the Metathesaurus is assigned (at least) one semantic type from the Semantic Network, independently of its hierarchical position in a source vocabulary. Version 2004AA of the UMLS is used in this study.

Set of descendants of a given concept

The set of descendants of a concept consists of the first-generation descendants of this concept (i.e., its children and narrower concepts in Metathesaurus parlance) and their descendants, recursively, all the way to the bottom of the graph. In graph theory, this operation is called the transitive closure of hierarchical relations. It is realized by performing a depth-first traversal of the graph. Because the Metathesaurus graph contains cycles, precautions must be taken for avoiding loops in the graph traversal. The **formal approach** uses a set of heuristics and rules in order to identify and eliminate all cycles from the Metathesaurus graph. For example, redundancy (i.e. the number of sources asserting each relation) is used to select the relation *A parent of B* over *B parent of A*. Confidence criteria associated with vocabularies can be exploited as well. Resolving complex cycles may require manual review by an expert. The interested reader is referred to [4] for more details. In contrast, the **naïve approach** simply consists of marking the nodes visited during the traversal of the graph in order to avoid visiting the same node twice. Although effective in preventing loops, this approach is naïve as nothing but the order in which nodes are visited determines what relation will be ignored in case of cycle.

The concepts *Desire for food* (C0003618), *Appetite Regulation* (C0003622), and *Food Intake Regulation* (C0086311) are used to illustrate this issue. Figure 1 shows that, using the naïve approach and starting from the concept C0003618 to recover its descendants, the relation *C0086311 parent of C0003618* is ignored because it would cause a cycle in the graph (*C0003618 → C0003622 → C0086311 → C0003618*). On the other hand, the same relation is used – not ignored – when the graph is traversed from C0086311 (Figure 2), while the relation *C0003622 broader than C0086311*, used in the previous graph is now ignored because it would cause the cycle (*C0086311 → C0003618 → C0003622 → C0086311*). In contrast, the formal approach consistently identifies the relation *C0086311 parent of C0003618* as inappropriate because the source vocabulary of this relation does not meet the confidence criteria required.

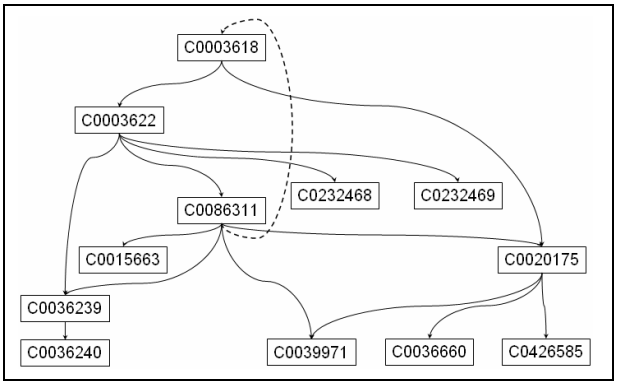

Figure 1 – Descendants of C0003618, ignoring the relation C0086311 parent of C0003618

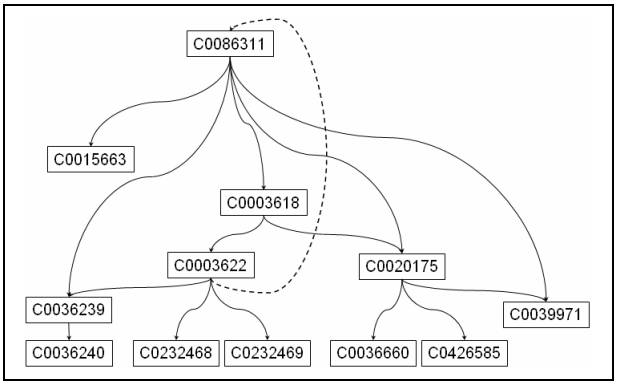

Figure 2 – Descendants of C0086311, ignoring the relation C0003622 broader than C0086311

Semantic coherence of a set of concepts

The semantic characterization of a set of concepts is provided by the distribution of semantic types or semantic groups observed in this set. In other words, the distribution of semantic types or groups for the concepts in the set of descendants represents a measure of the semantic coherence of the set. For example, the concept *Adrenal cortex diseases* (C0001614) is categorized as *Disease or Syndrome*, a semantic type from the group ***Disorders***. All its descendants also belong to the semantic group ***Disorders***. However, while most descendants are categorized as *Disease or Syndrome* or *Neoplastic Process*, the following semantic types which are not descendants of *Disease or Syndrome* in the Semantic Network are also used to categorize some descendants of *Adrenal cortex diseases*: *Anatomical Abnormality*, *Congenital Abnormality*, *Finding*, *Injury or Poisoning*, *Pathologic Function* and *Sign or Symptom*. In this example, the semantic coherence of the descendants

of *Adrenal cortex diseases* is good from the perspective of the semantic groups because only one semantic group (***Disorders***) is represented among the descendants. There is, however, a significant dispersion in terms of semantic types as six semantic types are represented in addition to the descendants of the semantic type of *Adrenal cortex diseases.*

Semantic compatibility among semantic types

Semantic types are often used to assess the semantic compatibility of the descendants of a concept C with respect to C [1]. The semantic types of descendant concepts are expected to be the same as the semantic types of C or one of their descendants in the Semantic Network. For example, the descendants of *Adrenal cortex diseases* include *Adrenal cortex necrosis* (C0151793), also categorized as *Disease or Syndrome* and *Tumors of adrenal cortex* (C0001618), categorized as *Neoplastic Process*. *Adrenal cortex necrosis* and *Adrenal cortex diseases* have the same semantic type *Disease or Syndrome* and are therefore compatible. *Neoplastic Process* is a descendant of *Disease or Syndrome* in the Semantic Network, which makes *Tumors of adrenal cortex* compatible with its ancestor *Adrenal cortex diseases.*

A looser assessment of semantic compatibility is provided by the semantic groups: all descendants of a given concept C are expected to belong to the same semantic group as C. For example, *Accessory adrenal cortex* (C0266277) is another descendant of *Adrenal cortex diseases*, but its semantic type is *Congenital Abnormality*, which is *not* a descendant of *Disease or Syndrome*. The two concepts are nonetheless compatible with respect to semantic groups, because they share the same semantic group ***Disorders*** (i.e., their respective semantic types both belong to the same semantic group ***Disorders***).

METHODS

Establishing sets of descendants

For each Metathesaurus concept, we established the set of its descendants by computing the transitive closure of hierarchical relations, using a depth-first traversal of the Metathesaurus graph[1]. In the **naïve approach**, the relations which would cause loops in the traversal are eliminated on the fly, as required for each set. In order to avoid building large graphs due to inaccurate links, we also limit the maximal depth to 50 levels, knowing that such depth is never reached in the acyclic Metathesaurus. With the **formal approach**, the links causing cycles among hierarchical relations have been identified and removed prior to searching for descendants, transforming Metathesaurus hierarchical relations into a directed acyclic graph.

Comparing sets of descendants

The sets of descendants obtained for a given Metathesaurus concept by naïve and formal approaches respectively were compared as follows. First, a simple intersection of the two sets is performed in order to identify the concepts common to both sets and those specific to each set. We then investigate the semantic coherence of both sets by studying the distribution of semantic types (and groups) in these sets. In addition, we verify the compatibility – defined in the Background section – of each semantic type and group represented in the descendants, with respect to that of the source concept. Numbers of semantic types (and groups) represented in the sets of descendants constitute the quantitative aspects of semantic coherence, while the compatibility of semantic types (and groups) represented in the sets with respect to that of the source concept defines semantic coherence qualitatively.

RESULTS

General

When comparing the sets of descendants obtained for a given Metathesaurus source concept by naïve and formal approaches respectively, we identified four distinct cases, shown in Table 1:

1) The sets of descendants are both empty (the source concept is a leaf concept).
2) The sets of descendants are identical.
3) The traversal of the graph was interrupted after reaching more than 50 levels (naïve approach). The set of descendants recorded is incomplete.
4) The sets of descendants are complete and exhibit differences. Further analysis of the differences focus on this group.

Table 1 – Categories of Metathesaurus concepts with respect to differences in their descendants obtained by naïve and formal approaches

Category	**# of source concepts**	
No descendants	765,811	(75.0 %)
Same descendants	221,641	(21.7 %)
Incomplete (interrupted)	6,830	(0.7 %)
Different descendants	26,584	(2.6 %)
Total	*1,020,866*	*(100.0 %)*

[1] In both naïve and formal approaches, we removed from the Metathesaurus graph a limited number of relations causing the root of one vocabulary to appear in the hierarchy of another vocabulary. For example, we delete the link between the MeSH concept *Controlled thesaurus* (C0282502) and *Read thesaurus* (C0338370), which serves as the root for all concepts in this source vocabulary. With this link present, all concepts from Read would be wrongly considered descendants of *Controlled thesaurus.*

Number of descendants

We now consider only the 26,584 source concepts whose sets of descendants are complete and exhibit differences. The statistical characteristics of the numbers of descendants obtained by each approach are summarized in Table 2, as well as those for the difference between the two methods. The number of descendants is always larger with the naïve approach. More precisely, on average, the naïve approach tends to identify nearly 75% more descendants than the formal approach. The concept with the largest number of descendants is *Cyclic compound* (C0596399); the largest difference corresponds to *Chemical bonding* (C0596307).

Table 2 – Numbers of descendants obtained by naïve and formal approaches (minimum, maximum, median and average)

Approach	Min.	Max.	Med.	Avg.
Naïve	1	102,161	58	1112.4
Formal	0	100,333	22	639.0
Diff. (N-F)	1	59,416	13	473.4

Semantic coherence: quantitative aspects

Semantic types. Out of the 26,584 concepts exhibiting differences in the numbers of descendants computed by the two approaches, 14,787 (56%) also exhibit differences in the semantic types represented in the sets of descendants. In other words, in 44% of the cases, the additional descendants computed by the naïve approach have the same semantic types as the descendants computed by the formal approach. The number of additional semantic types ranges from 1 to 68 (median = 2). On average, the naïve approach tends to identify 49% more semantic types in the descendants than the formal approach.

Semantic groups. Only 8,256 (31%) of these 26,584 concepts exhibit differences in the semantic groups represented in the sets of descendants. The number of additional semantic groups ranges from 1 to 11 (median = 1). On average, the naïve approach tends to identify 127% more semantic groups in the descendants than the formal approach.

Semantic coherence: qualitative aspects

Semantic types. For the 14,787 concepts exhibiting additional semantic types represented in the sets of descendants by the naïve approach, the additional semantic types of the descendants are compatible with that of the source concept in only 11% of the cases.

Semantic groups. For the 8,256 concepts exhibiting additional semantic groups represented in the sets of descendants by the naïve approach, the additional semantic groups of the descendants are compatible with that of the source concept in only 27% of the cases.

EXTENDED EXAMPLE

We use the concept *Generally contracted pelvis in pregnancy, labour, and delivery* (C0156969) to illustrate the differences observed in the sets of descendants obtained by the two methods under investigation (Figure 3). The semantic types of this concept are *Acquired Abnormality* and *Disease or Syndrome*. With the formal approach, C0156969 has two descendants: *Generally contracted pelvis, delivered* (C0156971), categorized as *Disease or Syndrome* and *Generally contracted pelvis, antepartum* (C0156972), categorized as *Acquired Abnormality* and *Disease or Syndrome*. The naïve approach identifies four additional descendants for C0156969: *Generally contracted pelvis, unspecified as to episode of care in pregnancy* (C0156970) categorized as *Acquired Abnormality* and *Disease or Syndrome* and its three children: *Small pelvic bone* (C0426852), categorized as *Finding, Midpelvic contraction* (C0405009) and *Pelvic disproportion* (C0558374), both categorized as *Anatomical Abnormality*. The relationship between C0156969 and C0156970 has been eliminated by the formal approach because of the presence of "unspecified" in a leaf term from ICD9CM [4].

The direct descendants of C0156969 are coherent and compatible with the source concept because the two semantic types represented in this group are that of the source concept. Additional semantic types for the second-level descendants include *Anatomical Abnormality* and *Finding*, which are not descendants of the semantic types of the source concept. The semantic coherence of these descendants is weak, because it only exists at the level of the semantic group (the six descendants of C0156969 belong to ***Disorders***).

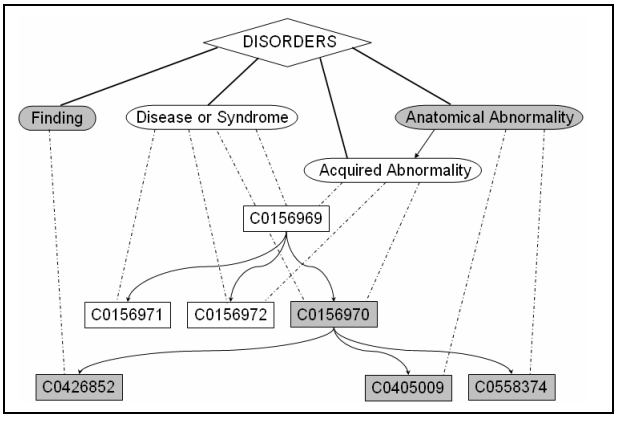

Figure 3 – The descendants, semantic types, and semantic group of C0156969 *(grey components are specific to the naïve approach)*

DISCUSSION

Formal approach vs. naïve approach. This study verified in part our hypothesis. We showed that the formal approach greatly reduces the number of descendants for a large number of concepts and also results in a more modest gain in semantic coherence among the descendants. These findings are contrasted by the fact that they only apply to 12% of the concepts with descendants, i.e., 2.6% of all UMLS concepts. However, we also demonstrated that the formal approach selects consistently the links to be removed, while the order in which the graph is traversed determines which links are ignored in the naïve approach. Finally, we showed that, in practice, the formal approach requires less resources to build the graphs of descendants and, more generally, to traverse the Metathesaurus graph. In contrast, depths over 50 are not uncommon with the unfiltered hierarchical relations in the Metathesaurus, resulting in unnecessarily complex and often larger graphs.

Lessons learned. We have noticed a number of cases where legitimate descendants were absent from the set obtained by the formal approach. For example, while it legitimately removes *Tonsillitis* (C0040425) from the descendants of *Acute tonsillitis* (C0001361), the formal approach also unnecessarily removes *Acute gangrenous tonsillitis* (C0339866), which is indeed also a kind of acute tonsillitis. In this study, some cases where the two approaches yield different numbers of descendants for a given concept may be related to this phenomenon, especially when the descendants obtained by the naïve approach are semantically compatible with the source concept.
We traced this problem back to an error in the formal algorithm used to remove cycles, where *parent/child* relations from SNOMED CT were removed when only the removal of mapping relations recorded as *broader/narrower than* was intended.

Limitations. One limitation of the formal approach is that, while effectively filtering out many illegitimate descendants, the semantic compatibility of the remaining descendants with the source concept is far from complete In fact, 59% of all descendants obtained by the formal approach for the 26,584 concepts investigated in detail in this study are incompatible in terms of semantic type with their respective source concepts. For instance, descendants of *Hostility* (C0020039), categorized as *Mental Process* (semantic group ***Physiology***), include *Reaction belligerent* (C0542181), categorized as *Finding* (semantic group ***Disorders***).
While failure to create semantically compatible and coherent sets of descendants may be in part a limitation of our formal approach, it is essentially indicative of the limited semantic coherence of the Metathesaurus. In other words, our method has been successful in ensuring that hierarchical relations involved in cycles are consistently removed from the Metathesaurus graph. However, only a semantic analysis of all hierarchical relations would be required in order to make the Metathesaurus coherent not only structurally (i.e., acyclic), but also semantically [6]. It is a common misconception that the Metathesaurus is an ontology of biomedicine, while it is in fact the product of terminology integration. Additionally, errors in concept categorization are also partly responsible for the lack of compatibility in the descendants.

Future work. In this study, the issue of eliminating cycles from the Metathesaurus graph is addressed from a theoretical perspective: the creation of sets of descendants. We plan to investigate the difference between formal and naïve approaches in practical applications utilizing hierarchical relations in the UMLS. One such application is mapping among terminologies, for which previous work has only considered the formal approach [7]. One mapping technique consists of exploring the graph of the ancestors of a given concept for identifying a more generic concept in the target vocabulary. We believe this application would provide an interesting case study for comparing the two approaches.

Acknowledgments

This work was funded in part by the Région Bretagne (PRIR).

References

1. Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS Methathesaurus. Proc AMIA Symp 1998:810-4
2. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51
3. Hahn U, Schulz S. Boosting the Medical Knowledge Infrastructure — A Feasibility Study on Very Large Terminological Knowledge Bases Proc Symp on Engineering of Intelligent Systems 2004
4. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001:57-61
5. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 10(Pt 1):216-20
6. McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Boston: Kluwer Academic Publishers; 2002. p. 181-198
7. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998:815-9