Tracking Meaning Over Time in the UMLS® Metathesaurus®

Tammy Powell, MS [1], Suresh Srinivasan, MS [2], Stuart J. Nelson, MD [1], William T. Hole, MD [1],
Laura Roth, MLS [2], Vladimir Olenichev [2]
1. National Library of Medicine, Bethesda, Maryland
2. Management Systems Designers, Vienna, Virginia

Abstract

The Unified Medical Language System® (UMLS) Metathesaurus contains records arranged by concept or meaning. Each concept contains a unique identifier (CUI) that can be used to track the concept over time. Since the January 2001 release, the Metathesaurus has included the file MRCUI that contains mappings for CUIs that disappear. This paper describes the processes that facilitated this effort and the ongoing effort to find suitable mappings for concepts whose meanings no longer exist in the Metathesaurus. This study highlights the need to identify missed synonymy prior to a release. It also shows a need to work more closely with source providers to identify the closest match in the Metathesaurus when they eliminate terms from their vocabularies.

Introduction

The unit of meaning in the UMLS Metathesaurus is the concept. Opaque identifiers, called Concept Unique Identifiers or CUIs, are assigned to each concept and maintain this association forever, allowing CUIs to be used in applications and databases as stand-ins for concepts.

The UMLS Metathesaurus contains concepts and concept names from more than 70 vocabularies and classifications, some in multiple editions. Names of concepts are present in 15 languages. The 2002 edition of the Metathesaurus includes approximately 776,940 concepts and 2.10 million concept names.

What are Concepts, LUIs and CUIs?

Terms from different constituent vocabularies with the same meaning are gathered into a concept. The concept is assigned a concept unique identifier (CUI) - an 8-character identifier beginning with the letter "C" and followed by 7 digits, e.g., C0228498. While the CUI itself has no intrinsic meaning, this identifier remains the same across versions of the Metathesaurus, irrespective of the term designated as the preferred name of the concept. This facilitates file maintenance and management, as well as tracking the meanings assigned to a given term over time. It is "the name (of a concept) that never changes." [1]

For the purposes of Metathesaurus construction, , we chose to define synonymy by the extensional meanings of terms in their source vocabularies, where they can be perceived, and a heuristic. We say two terms are identical in meaning if the vast majority of biomedical professionals would find any distinction in meaning between the two terms is inconsequential, that is, a distinction that was not supportable, a distinction without a significant difference. Of course, the nature of scientific progress is one of continual exploration of what appear to be only minor distinctions. This provides one of the forces that lead to the continual tension between accuracy and currency in vocabulary maintenance [2].

Programs from the Specialist Lexicon (one of the three UMLS Knowledge Sources) are used to generate lexical information about all English terms in the Metathesaurus [3]. The terms that name the concept have a lexical unique identifier or LUI. All terms that have identical normalized forms will share a LUI. The normalization process abstracts away from minor lexical variation such as differences in case, inversion, punctuation, singular-plural inflection, etc. Terms that have an identical string but a different meaning will have identical LUIs but will appear in different concepts [4]. For example, some Metathesaurus source vocabularies use the term "Acetaminophen" to mean a pharmaceutical substance. Other vocabularies use the same string to mean the laboratory procedure to assay for acetaminophen.

The example below serves to illustrate this further. Shown are two concepts that contain the abbreviation "BPD":

Concept #1	**CUI C0006012**
Term	Borderline Personality Disorder
LUI	L0006012
SUI	S0020223
Term	BPD
LUI	L1120288
SUI	S1344541
Concept #2	**C0006287**
Term	Bronchopulmonary Dysplasia
LUI	L0006287
SUI	S0020777
Term	BPD
LUI	L1120288
SUI	S1344541

As shown, these two concepts share a LUI for the term "BPD". The string "BPD" being identical in both concepts, they also share the string unique identifier (SUI).

The UMLS attempts to ensure that the CUI and its associated concept, or more accurately the meaning of the concept, stay together in perpetuity. A meaning that disappears from the Metathesaurus will result in the retirement of its associated CUI. Conversely, new CUIs are assigned only after ensuring that the meaning is not already in the Metathesaurus or has not been in the Metathesaurus in the past. It is this discipline that lets application use the UMLS CUIs as stand-ins for concepts, secure in the knowledge that the association is permanent.

The UMLS has always included files that contain a list of CUIs that were deleted from the previous release (DELETED.CUI) and a list of CUIs that were merged (MERGED.CUI) with another CUI relative to the previous release. However, these files were not sufficient to provide concept permanence or tracking by themselves. Therefore, in 2000, while editing the 2001

version of the Metathesaurus, we chose to establish a discipline in trying to meet the goal of providing a historical record of all extinct CUIs and of providing a path from them to one or more closely related, extant CUIs. The file MRCUI is a start at making this link and has been an integral part of the UMLS Metathesaurus release since 2001.

The MRCUI file contains each CUI that existed in any prior release but is not present in the current release. When available, mappings to a current CUI along with the relationship between the two concepts are provided.

Column names in MRCUI and description

CUI1 - Retired CUI - was present in some prior release, but is currently missing.

VER - The last release version in which CUI1 was a valid CUI.

CREL - The relationship CUI2 has to CUI1, if present, or DEL if CUI2 is not present. Valid values currently are SY or DEL. As mapping to extant concepts that are not synonymous occurs, further relationships will be allowed.

CUI2 - The current CUI that CUI1 most closely maps to.

Sample Records (CUI1 VER CREL CUI2)

C0435517 1999 SY C0435516
C0361163 1998 DEL
C0785652 2000 BRD C0775088
C0234931 1996 NRW C0152459
C0171313 1995 DEL

The META/CHANGE files, especially MERGED.CUI and DELETED.CUI, contain changes from the last release only without the mappings.

This paper describes the methodology used in preserving and mapping the CUIs that disappear. It also provides data on how many of these deleted CUIs are determined to be broader, narrower, related, or synonymous to existing concepts within the Metathesaurus.

Why CUIs Disappear

There are several reasons why a CUI that was present in a previous release is no longer present in the current release.

1) *Merging of two or more CUIs that were previously released as separate concepts*
Addition of a new or updated source may provide information that two concepts previously distinct are in fact synonymous. In addition, Metathesaurus editors aided by automated processes, look for missed synonymy in the Metathesaurus. Once missed synonymy is detected, the editors merge the records into a single concept. The editing system algorithmically computes

which one of the two CUIs will stay, and which will be retired. Finding missed synonymy is a difficult problem in general and we have an ongoing effort to tackle the problem [5].

2) *Deletion*
The UMLS has seen a steady growth not just in the number of concepts but also in the number of sources contributing to these concepts. When a source is updated, there usually is some amount of reorganization of content within the source itself. This may take the form of: new knowledge resulting in further refinement or reclassification of meaning, dropping of source terms due to lack of use or usefulness, or fixing typographical errors in the source. Such changes may result in the deletion of a concept, or a missed relationship of synonymy with a previously existing concept, which may not be identified prior to a Metathesaurus release. In either case, the consequence for the Metathesaurus is that this results in concepts whose only names are from the old, replaced version of the source, and these concepts become candidates for deletion.

Other Changes that Impact CUIs
The 1992 release of the Metathesaurus introduced a more sophisticated concept structure where there is now a one-to-one correspondence between concepts and CUIs. Prior to 1992, the structure was more term-centric with synonyms being assigned different unique identifiers. This modeling shift resulted in the loss of many numbers.

When the Metathesaurus is subsetted using MetamorphoSys [6] to meet local licensing and other requirements, the resulting subset may not contain all the CUIs in the full release. In this case, applications can determine which CUIs really were deleted and which were eliminated in the subsetting process.

Why do CUIs need to be preserved or mapped?

The Metathesaurus is a Knowledge Source for application developers. Many applications such as patient record systems need to store and manipulate meanings over time. This task is not trivial, for both names and codes change over time as sources are updated. The name used for a meaning may change within source vocabularies as the purpose, science, or style change. Entire vocabularies (and their codes) may come and go or fall from favor. Other meanings for the current name appear, creating ambiguity. The vocabulary required tomorrow may not exist today, and its names for current meanings may well differ.

The consequences of the loss of access to information when these changes occur can be disastrous, for example in patient records. While the concept for a disease may no longer exist in a vocabulary, the information that a patient had a particular disease will still be present in the records [7]. Cimino discusses the need for concept permanence, once a concept has been created it must remain [8]. Because the scope of the Metathesaurus is determined by the scope of all its source vocabularies, once all sources remove a concept from their vocabularies, the concept is no longer part of the Metathesaurus release files. NLM's continuing maintenance and mapping of concepts and their CUIs provides a logical pathway through changes so that developers and researchers can create tools to navigate through them. Note that it is usually prudent to store a specific vocabulary's string and code in addition to the CUI, since in many cases these are also useful.

How are CUIs preserved?

NLM attempts to minimize the change in CUIs between major releases even though concepts may come and go or be merged or split several times while vocabulary is added and edited. This is done by storing the last released CUI while using internal unique identifiers during editing, then computing and assigning the most stable CUI at release time. Additional algorithms look for matches with previously released CUI's strings and meanings. Shown below are the counts of how many CUIs were merged and deleted over the past two major releases.

2002 Release
53,910 concepts were deleted but not mapped to an active concept.
2,017 CUIs were mapped to a broader concept
52 CUIs were mapped to a narrower concept
12,710 CUIs were mapped to a related concept
139,141 CUIs were mapped to a CUI that was synonymous

2001 Release
38,377 - Concepts were deleted.
46,196 - Concepts were mapped to a synonymous concept.

The numbers of deleted CUIs is remarkably large. Rarely are changes in source vocabularies represented in ways which might allow for automatic recognition of the new location of the same meaning [7,9]. The numbers of mapped concepts in 2002 represents those manual efforts.

The 2002 release contained a very large number of synonymous mappings. The change in Medical Subject Headings® (MeSH®) to a concept-oriented maintenance environment, and subsequent editing of the Supplementary Concept Records of MeSH, resulted in approximately the merger of over 100,000 previously unreviewed concepts in the Metathesaurus being merged with other concepts.

Manual CUI Mapping

In 2001 (for the 2002 release) NLM began providing Metathesaurus users with a mapping between a CUI that had become extinct and an existing CUI. In addition to the ten years of deleted CUIs that needed to be mapped, there was the problem of mapping CUIs that were identified as possibly disappearing in the next release.

Concepts that may not be part of the next release
During the editing cycle, queries identify concepts that will become extinct because there are no extant sources asserting names for them. Editors begin by trying to identify possible missed synonymy between these concepts and new or previously released concepts. If missed synonymy is identified then the two concepts are merged together. Programs algorithmically decide which CUI will be released.

In cases where missed synonymy is not identified, editors assign a "bequeathal" non-synonymous relationship to an extant concept closest in meaning which will end up in MRCUI. This work is being performed using the same software used during regular editing. Our goal is to bequeath a relationship for any concept that will disappear prior to a release.

Concepts That Disappeared During the First Ten Years of the Metathesaurus
There is also an effort to work back across time to assign bequeathal relationships to previously deleted CUIs that were not merged with another concept. Since the concepts no longer exist they do not appear in the database of live concepts or in the regular editing database. This required us to devise an alternate approach in order to map these concepts. A tool has been built to use past release data and allow editors to create mappings to concepts found in the current release.

Determining the Best Map
In determining where to map a deleted CUI or a CUI that may go away prior to the next release, the following principles are applied by the editors. Below is a list of these rules listed with the highest priority first.

1) Begin by looking for synonymy to an existing concept in the Metathesaurus. For sources where there is no source unique identifier for every term, this often yields a match. In some cases a source will make minor modifications to a string but still retain the same meaning. In the 2001 release of the Metathesaurus, CUI C036883 contains the preferred name "MOLD CHEESE ANTIBODY. IMMUNOGLOBULIN E". A search of the editing version of the Metathesaurus prior to the 2002 released showed that this concept was going to be deleted. However, there is a concept from the same source with the term "CHEESE MOLD TYPE ANTIBODY. IMMUNOGLOBULIN E".

We believe that it is most useful for the Metathesaurus users if an editor is able to find a concept that is synonymous to the one that has or will be deleted.

2) If not successful, try to find a concept that contains terms from the same source as in the concept that is being eliminated. Begin by looking for the parent in the context of the source. For example, the term "Reagents, Calibration, *Other* was deleted in 2001. This term was a child of "Reagents, Calibration" in the sources hierarchy. The meaning of the term to which "Reagents, Calibration, Other" is mapped is narrower than that of "Reagents, Calibration," to which the deleted CUI is mapped.

3) Try to choose something that is inclusive rather than close in meaning. For example, *"Sister Support"* could be mapped as narrower than *"Sibling Support"* or related to *Sisters*. *"Sister Support"* is closer in meaning to "Sibling Support" than to *Sisters*.

4) Map to a concept that is to be released with the next version of the Metathesaurus, not one that has been or is about to be eliminated.

5) If the meaning appears to be aggregated (e.g. > 1 Semantic Type) and all possible concepts to map to are assigned to a single Semantic Type, then map to multiple concepts. Or if the meaning cannot be mapped to any reasonable concept, map to multiple concepts. For example, it would be best to map a concept such as *DNA and DNA Research Techniques* to multiple concepts.

6) Choose simple (and general) as opposed to complex precoordinated expressions or multiple concepts. In trying to map "Non-Narcotic Analgesic Administration" an editor might find "Administration of analgesic" and Check *medical order for drug, dose, and frequency of non-*

narcotic analgesic administered. It would be best to map "Administration of analgesic" as broader.

7) In a very small number of cases it may not be possible to determine an appropriate map. This may because the concept is very general and not appropriate for mapping, the source did not give enough information for an editor to determine the original meaning, or because the concept was added by the source as an error. For example, one source unintentionally released the SGML entity ' ' as a term and this became a concept in the Metathesaurus.

Mapping Results

A review of 500 sample concepts that were deleted prior to the 2000 release and recently assigned mappings yielded the following results:

178 were mapped as narrower
156 were mapped as synonyms
155 were mapped as other related
9 were mapped as broader
2 cases were not mapped.
These results show editors were able to follow the highest priority rules while performing the mapping task. In addition, it highlights the need to aggressively identify missed synonymy prior to a CUI being deleted.

Mappings from Sources

When a concept is deleted or changed in a source vocabulary, the source maintainers have the best understanding of why the change was made and what the closest match is in their source. To the maintainers of the source, these changes are rarely significant [9]. For the Metathesaurus editors, understanding the reasons for a deletion or change can be difficult. This prevents them from easily identifying the best possible mapping. The 2002 release contains mappings from the UWDA vocabulary provided by the University of Washington for all deleted concepts from that source. NLM is looking at ways to use information about deleted concepts in MeSH to ensure they are accurately mapped.

Future work

NLM hopes to complete the mapping of all deleted concepts released during the first ten years over the next two years. There are plans to provide the mappings for every deleted CUI in the release when it first disappears from the Metathesaurus.

NLM has also begun to look for multiple vocabulary sources for concepts that currently appear in one source [10]. NLM is always interested in multiple sources for the meanings in the Metathesaurus, since the concepts are the most vulnerable to being deleted when their meanings appear in a single source.

Conclusion

The Metathesaurus Concept Unique Identifiers (CUIs) were designed to aid applications developers. Providing mappings as these identifiers are removed from the file will assist them even further.

The high number of cases (31%) where synonymy was identified reinforces the idea that this work should be done prior to a release. More sophisticated missed synonymy tools allow us to use information we gain while mapping concepts to create new missed synonymy queries that can be used in the future.

The MRCUI file fills a gap and will now allow users to create tools to logically track all CUI changes over time. In addition, the mappings will provide alternate concepts to use when meanings disappear from the Metathesaurus.

References

1. Nelson SJ, Powell TP, Humphreys BL. The Unified Medical Language System (UMLS) Project. In: Kent, Allen; Hall, Carolyn M., editors. Encyclopedia of Library and Information Science. New York: Marcel Dekker, Inc.; 2002. p.369-378.

2. Nelson SJ. The Unified Medical Language System Applicable Experiences and Observations. Presentation at: The Workshop on Compilation, Maintenance, and Dissemination of Taxonomic Authority Files; 1998 Jun 22; Washington, DC.

3. Specialist Lexicon [Fact Sheet], National Library of Medicine; Bethesda (MD); 2001 Feb 14

4. UMLS, UMLS Knowledge Sources. 12th ed. Bethesda (MD): National Library of Medicine; 2002.

5. Hole WT, Srinivasan S. Discovering Missed Synonymy in a Large Concept-Oriented Metathesaurus; Proc AMIA Symp; 2000.

6. Bodenreider O, Hole WT, Humphreys B, Roth L, Srinivasan S. Customizing the UMLS Metathesaurus for your Applications. Proc AMIA Symp; Nov. 2001.

7. Cimino JJ. Formal descriptions and adaptive mechanisms for changes in controlled medical vocabularies. Methods Inf Med 1996 Sep;35(3):202-10

8. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998 Nov;37(4-5):394-403 9. Tuttle MT, Nelson SJ. A poor precedent. Methods Inf Med 1996 Sep;35(3):211-7

10. Srinivasan S, Rindflesch TC, Hole WT, Aronson AR, Mork JG. Finding UMLS Metathesaurus Concepts in Medline. Proc AMIA Symp 2002, submitted.