MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Achieving "Source Transparency" in the UMLS® Metathesaurus®

William T. Hole[a], Brian A. Carlsen [b], Mark S. Tuttle[b], Suresh Srinivasan[c], Stephanie S. Lipow[b], Nels E. Olson[b], David D. Sherertz[b], Betsy L. Humphreys[a]

[a]*National Library of Medicine, Bethesda, MD, USA*
[b]*Apelon Inc., Alameda, CA, USA* [c]*MSD Inc., Vienna VA, USA*

Abstract

The UMLS Metathesaurus is a syntactically uniform, concept-based, semantically enhanced representation of many of the world's authoritative biomedical vocabularies. Released several times a year, the Metathesaurus is becoming a common, longitudinally maintained source of the current versions of these vocabularies. As vocabularies become standards for reimbursement, reporting, interoperation, and use by applications, the vocabulary obtained from the Metathesaurus must be consistent with that obtainable from each vocabulary's authority. Effective with the first 2004 release, the Metathesaurus represents new and updated sources "transparently" – both users and applications are able to "see" each vocabulary in the Metathesaurus without any of the small losses of information introduced by abstractions used in previous versions. Thus, the Metathesaurus can continue to provide its many semantic and lexical value-added features while guaranteeing that original sources will be "visible" in intact form. Vocabulary users and application developers will benefit from the enhancements and economies of scale offered by the Metathesaurus, while preserving distinctions between content provided by external authorities and content added as part of the Metathesaurus development and maintenance process.

Keywords

Unified Medical Language System, Medical Informatics, Controlled Vocabulary, Information Systems, Metathesaurus

Introduction

In the first 2004 release of the U.S. National Library of Medicine (NLM) Unified Medical Language System® [1] (UMLS), the Metathesaurus represents over 1 million concepts with more than 3 million concept names in more than 100 biomedical vocabularies, some in multiple languages [2]. Included with each release is a multi-platform software tool called MetamorphoSys that allows users to create custom subsets of the Metathesaurus for local use. The subsets might exclude vocabularies for which their enterprise does not have a license, include specific vocabularies useful to the user, or further restrict content by some of the value-added features of the Metathesaurus, such as particular Semantic Types.

Selected vocabularies are being recommended and designated as standards within the United States [3], and some of these standards are required for use in the regulated aspects of U.S. healthcare [4]. With these standards come the challenges associated with compliance and certification. Part of the strategy for addressing these challenges will be the creation of computationally authoritative versions of these standard vocabularies. Metathesaurus application developers will want to know that whatever portions of the Metathesaurus they deploy will be certifiably in compliance with the appropriate vocabulary standards.

As emphasized in Section 2.02 of its documentation, the Metathesaurus has always endeavored to preserve the concepts, concept attributes, inter-concept hierarchical connections, and other relationships present in its source vocabularies. Toward this end, the original concept-oriented distribution file format accurately preserves meanings, attributes, and relationships between concepts. However, representing relationships at the concept level can obscure some relationships that are not concept-oriented and, in some cases, can make it difficult to generate completely accurate source hierarchies.

"Source transparency" and the Rich Release Format

In anticipation of the need to address certification and compliance, and in response to the growing richness and refinement in the schemas and content of the Metathesaurus source vocabularies, the NLM and collaborators at Apelon have developed the "Rich Release Format (RRF)" [5]. In contrast to the "Original Release Format (ORF)," the RRF supports complete "source transparency". The ORF will continue to be available as an output of MetamorphoSys for those users desiring to maintain existing Metathesaurus applications.

Source transparency is an important principle applicable to the representation of content from one vocabulary in another. We say the Metathesaurus supports "source transparency" if both users and applications can access its source vocabularies' content without loss of information. The concept-based abstractions in the ORF prevent the perfect extraction of a few sources because of differences between the Metathesaurus concept-based representation and the code-based nature of these sources. This small loss of information has been eliminated in RRF so that, for instance, the distinction between a source's inter-term or inter-code relationships and the information added in creation of the Metathesaurus is made explicit.

Table 1: Abbreviated example of MRCONSO.RRF

CUI	LAT	…	AUI	SAUI	SCUI	SDUI	SAB	TTY	CODE	STR
C0000733	ENG		A0017747		M0000007	D000007	MSH	MH	D000007	Abdominal Injuries
C0000733	ENG		A0074197		M0000007	D000007	MSH	PM	D000007	Injury, Abdominal
C0000733	ENG		A0851715				RCD	PT	XA00L	Abdominal injury
C0000733	ENG		A2850179			10060924	MDR	LT	10060924	Abdominal injury
C0000733	ENG		A2850180			10060924	MDR	PT	10060924	Abdominal injury
C0000733	ENG		A3089214	206363015	128069005		SNOMEDCT	SY	128069005	Abdominal injury
C0000733	ENG		A1573318			10060924	MDR	LT	10000076	Abdominal injury NOS
C0000733	ENG		A0074146		M0000007	D000007	MSH	EN	D000007	Injuries, Abdominal
C0000733	ENG		A3515208	732025010	128069005		SNOMEDCT	FN	128069005	Injury of abdomen (disorder)
C0000733	ENG		A3515209	194663011	128069005		SNOMEDCT	PT	128069005	Injury of abdomen
C0000733	GER		A1421546				DMD	MH	D000007	Abdominalverletzungen
C0000733	GER		A1428978				DMD	SY	D000007	Bauchverletzungen
C0000733	GER		A2455670				DMD	SY	D000007	Bauchtrauma

Because of the prospective commitment to source transparency in the Metathesaurus, RRF also accommodates the novel intra-source relationships appearing in SNOMED CT [6] and the NCI Thesaurus [7,8]. *A major purpose of RRF is to make explicit the source of all information - in particular, intra-source relationships in vocabularies and Metathesaurus value-added information - and to carefully maintain the distinction between the two*.

Methods

To implement the RRF, additional fields and identifiers were added to many Metathesaurus files and several new relational files were created. Three ORF files (MRCON, MRSO and MRATX) are deprecated – that is, their continued use is not recommended, although they remain available as an output of MetamorphoSys so that existing applications will continue to work as expected. All of the additions are designed to make it easier for applications developers to subset the UMLS Metathesaurus for particular applications, and to maintain these applications appropriately as the source vocabularies in the Metathesaurus are updated, and Metathesaurus added-value information changes.

The RRF achieves transparency by including identifiers already in the software environment used by the NLM to create, maintain and publish the Metathesaurus on a regular basis. This environment was built on a foundation that makes use of an "atomic" representation of "units of work." In turn these units of work refer to **AUI**s (Atomic Unique Identifiers) that identify each occurrence of each string naming a concept, in each vocabulary. Given this very fine-grained view of every source, the RRF can specify all intra-source relationships in current Metathesaurus vocabularies, whether they are between concepts, concept names, or codes.

Easier Customization

Source transparency makes it possible to extract any source vocabulary from the Metathesaurus and to demonstrate that there is no information loss from the original source input. As described above, the Metathesaurus **AUI** is present in the internal system that NLM uses to maintain the Metathesaurus, but has not previously been distributed in the Metathesaurus release files. The RRF format includes this and related information that enables accurate representation of all known intra-source relationships, including the novel types of relationships, such as the relationship groups present in SNOMED CT [9] and the NCI Thesaurus. There is also a more consistent and explicit approach to labeling source-asserted identifiers and the directionality of source-asserted relationships. We believe that the benefits of source transparency will far outweigh the costs in file size and new complexity from **AUI**s and analogous "atomic" level data – especially since UMLS users may still generate the ORF file formats with MetamorphoSys.

While source transparency ensures that there is no information loss when a vocabulary is added to the Metathesaurus, it does *not* mean that the Metathesaurus will reproduce the original file format of each of its source vocabularies. Instead, the Metathesaurus will continue to provide all of its source vocabularies in a common, fully-specified format.

Other Benefits of the Rich Release Format

While the major imperative driving the explicitness in RRF is source transparency, other related priorities have shaped its evolving representation. Among them are desires: a) to make it easier for users to select or exclude designated data elements ("information") from selected sources using MetamorphoSys, b) to enable the computation, representation and use of "change sets" (transactions that represent changes to vocabularies and to the Metathesaurus), c) to support more productive creation and maintenance of inter-vocabulary linking, and d) to improve the degree to which the Metathesaurus is self-documenting. Each of these additional priorities is described in more detail in the sections below.

Selecting and Excluding Designated Information

Specific Metathesaurus vocabularies and groups of vocabularies useful for particular purposes (e.g., clinical applications, natural language processing) can be readily and completely extracted along with their UMLS identifiers. As described in its training resources [10], the UMLS Metathesaurus almost always requires customization for particular applications. A common method of customization is by source.

A new (relational) table, MRCONSO.RRF, combines and expands the concept and vocabulary source information from the original MRCON and MRSO files – thus eliminating the need to

Table 2: Abbreviated example of MRREL.RRF

CUI1	AUI1	STYPE1	REL	CUI2	AUI2	STYPE2	RELA	RUI	SRUI	SAB	SL
C0000733	A0017747	AUI	SIB	C0026771	A0088576	AUI		R07318440		MSH	MSH
C0000733	A0017747	AUI	SIB	C0027531	A1009355	AUI		R07318460		MSH	MSH
C0000733	A0851715	AUI	PAR	C0272429	A1287926	AUI		R08389980		RCD	RCD
C0000733	A2850180	AUI	RQ	C0000733	A1573318	AUI	classified_as	R15410246		MDR	MDR
C0000733	A2850180	AUI	RQ	C0160442	A0286410	AUI	classified_as	R15410234		MDR	MDR
C0000733	A3515208	SCUI	RO	C1272215	A3738566	SCUI	onset_of	R19240361	2306668022	SNOMEDCT	SNOMEDCT
C0000733	A3515209	AUI	CHD	C0160455	A3064559	AUI	isa	R20547235	1940913024	SNOMEDCT	SNOMEDCT
C0000733	A3515209	AUI	CHD	C0160704	A2969893	AUI	isa	R20547222	1910782023	SNOMEDCT	SNOMEDCT
C0000733	A3515209	AUI	PAR	C0272429	A3288331	AUI	inverse_isa	R21046769	163069021	SNOMEDCT	SNOMEDCT
C0000733	A3515209	AUI	PAR	C1291934	A3398961	AUI	inverse_isa	R21046770	163070022	SNOMEDCT	SNOMEDCT
C0000733		CUI	RB	C0860259		CUI		R17010123		MTH	MTH
C0000733		CUI	RN	C0160443		CUI		R17185140		MTH	MTH
C0000733		CUI	RO	C0160420		CUI		R03398638		MTH	MTH

join tables to select concepts and terms from particular sources. This file has rows and identifiers for every occurrence of every string naming a concept in every source. For example, if three different sources contain the exact string "Atrial Fibrillation" there are three rows for that string in MRCONSO.RRF, and, in addition to the Metathesaurus concept (**CUI**), term (**LUI**), and string (**SUI**) identifiers, each row has a unique Metathesaurus **AUI** for each occurrence in each source.

A new "Content View Flag" (**CVF**) has been added to many RRF tables to allow easier extraction of defined sets of content – whether names, attributes, or relationships for particular user purposes, for example for clinical uses or natural language processing. The **CVF** can be used when customization by source alone does not easily eliminate content that is superfluous or detrimental to certain applications, e.g., obsolete terms, terms that lack face validity, or inappropriate hierarchical relationships. The content views will be created and expanded over time based on requests or submissions from the UMLS user community.

Enabling of "Change Sets"

The ORF includes files that track the disappearance of concepts and strings from the previous version of the Metathesaurus and, in the case of concept identifiers, over most of the history of the Metathesaurus. The ORF does not allow easy detection of other types of changes in the Metathesaurus, such as the addition or disappearance of specific relationships and attributes.

In addition to the **AUI**s and other source specific identifiers described above, the RRF includes persistent identifiers for all relationships (**RUI**s) and all attributes (**ATUI**s) released in the Metathesaurus. The continued existence of these identifiers will indicate content that is unchanged across versions of the Metathesaurus. The appearance or disappearance of these identifiers will signal change. This enables the generation of complete Metathesaurus change sets. The sets provide a straight-forward way to update applications incrementally as new releases of the Metathesaurus are issued. Again the objective is to support explicit representation of change in both Metathesaurus source vocabularies and in Metathesaurus value-added information.

Improved Support for Inter-Terminology Linking

The RRF format makes it easier to create and distribute robust, purpose-specific mappings between different Metathesaurus vocabularies at various levels. Specifically, mappings are supported between atoms and codes as well as concepts. Although the original Metathesaurus release format can represent one-to-one, one-to-many, and one-to-Boolean expression mappings, the more complex mappings are cumbersome to maintain and to use and the original format does not accommodate rule-based mappings.

In the RRF, the Associated Expressions file (MRATX.RRF) is deprecated in favor of two new mappings files (MRMAP.RRF and MRSMAP.RRF), which have a more robust structure for representing simple, complex, and rule-based mappings using Metathesaurus or source-asserted unique identifiers.

Self-Documentation

A new file (MRDOC.RRF) lists all possible values for fields containing a finite set of such values, e.g., **TTY**, **ATN**, **TS**, **STT**, **REL**, and **RELA**. By joining this file with MRCOLS.RRF, a user may identify which files contain these fields (columns).

Results

A simplified illustration of the MRCONSO.RRF table implementation is presented in A new "Content View Flag" (CVF) has been added to many RRF tables to allow easier extraction of defined sets of content – whether names, attributes, or relationships for particular user purposes, for example for clinical uses or natural language processing. The CVF can be used when customization by source alone does not easily eliminate content that is superfluous or detrimental to certain applications, e.g., obsolete terms, terms that lack face validity, or inappropriate hierarchical relationships. The content views will be created and expanded over time based on requests or submissions from the UMLS user community.. Interested readers should explore the full RRF file structure[11] , documented in a way so those familiar with the ORF file structure [12] should find easy to understand. Sample files are also available [13]

Table 1 shows selected information for a single Concept, UMLS CUI C0000733. Several languages of origin are shown as **LAT**). Term Status, String Type, and the Lexical and string unique identifier columns are omitted for simplicity. The Atomic Unique Identifiers (**AUI**s) are followed by Source Atomic Unique Identifiers (**SAUI**s). Source Concept Unique Identifiers

(**SCUI**s) are present for MeSH and SNOMED CT. MedDRA and MeSH have a descriptor structure which groups related concepts, and so show a Source Descriptor Unique Identifier (**SDUI**). The Source abbreviation (**SAB**) identifies the source vocabulary. Note that versioned **SAB**s are not present in the example, though users may elect to include versioned **SAB**s in all Metathesaurus files as an option in MetamorphoSys; version information also present in MRSAB.RRF. The Term Type (**TTY**) classifies the role this term plays within its source vocabulary, and the **CODE** is a vocabulary's main identifier for this term. Two additional fields are not shown; one is the "Suppressible flag" which is used to tag content unlikely to be of use to many users, such as face-invalid terms with implied context or redundant or idiosyncratic abbreviations. The Content view flag is also not shown; it will identify set membership in various Metathesaurus views as described above.

The single **CUI** represents the Metathesaurus assertion that all these terms are synonymous. All **AUI**s are distinct, since the "atom" is distinct for a string appearing with a term type and code in a Source, and are constant across versions of that Source. The variety of roles of terms is shown by a selection of various Term Types (**TTY**s) including Main Heading (**MH**s), Synonyms (**SY**s), and Preferred Term (**PT**).

A simplified example of the MRREL.RRF table is shown in **Table 2**. **CUI1** is the unique identifier for the first concept and **AUI1** is the unique identifier for first atom, present when the relationship is asserted between atoms. **STYPE1** is type of identifier *originally* or *natively* used by the source to express the first anchor of the relationship. **REL** labels the kind of relationship. Similarly for the second anchor **CUI2** is the unique identifier of the second concept; **AUI2** is the unique identifier for second atom; and **STYPE2** is the identifier type used by the source**. RELA** is the additional relationship label, a more detailed description of the relationship. **RUI** is the unique identifier for relationship, and **SRUI** is the Source attributed relationship identifier**. SAB** is the Source abbreviation, while **SL** identifies the Source of relationship labels which may have been supplied by the Metathesaurus rather than the source. Not shown in this example are **RG** which identifies Relationship Groups and **DIR** which labels the Source asserted directionality, since all relationships are represented in both directions in the Metathesaurus**.** As in MRCONSO.RRF, there is a Suppressible flag **SUPPRESS** and a Content View Flag **CVF**.

Along with the atomic representation in MRCONSO.RRF, the RRF relationships table may now represent relationships between any of the identifiers as needed. For example, where a source asserts relationships between its atoms, this relationship is transparently shown between **AUI**s and attributed in the **SAB** and **SL**. Alternatively, atoms in one source used for clinical documentation may be mapped to atoms in another vocabulary required for reporting, to other **CUI**s to identify a view of synonymy that differs from the Metathesaurus view, or even to Source Descriptors used in a target Source's descriptor aggregation.

Another new file, MRMAP.RRF represents simple and complex mappings between Metathesaurus UIs or the identifiers and concept names in one source to Metathesaurus UIs or identifiers and concept names in another source. In addition, MRMAP.RRF contains source asserted historical mappings (i.e. mappings between obsolete terms or concepts and current ones. More complex mappings are accommodated with set unique identifiers, machine processible rules for when mapping is to be applied and the rule to apply. Most data in MRMAP does not require its full complexity and is also represented in a simpler table, MRSMAP**.**

A standard model has been defined for the candidate "objects" in the Metathesaurus such as concepts, attributes, and relationships; this model is the foundation of proposed Java APIs that "surface" RRF content which are further documented in draft Javadoc documents [14]. MetamorphoSys uses this model internally and it will be able to consume or produce representations of these objects in RRF or other formats. The UMLS Knowledge Source Server (KSS) plans to deploy this model in its Metathesaurus API.

Discussion

The RRF adds more detailed information to both the original concept-based representation of vocabularies and added-value information in the Metathesaurus. The new format allows identification of the source of all information, making it possible to select or exclude information to create useful subsets or to demonstrate the complete and correct representation of a source vocabulary's information. In future versions of the Metathesaurus, change sets will be computed to implement updates using this detailed identification.

It is important to note that the initial version of the RRF was fully populated only for each new or updated vocabulary. In this version, vocabularies previously released in the Metathesaurus and not yet updated lack some new data, such as representation of relationships at the atom level. Full RRF data for these vocabularies will be available when they are updated in the Metathesaurus.

One measure of the utility of RRF will be the degree to which users and developers can be certifiably compliant with terminology standards and, at the same time, take advantage of the Metathesaurus value-added features. Early users of the new Metathesaurus functionality will probably be the vocabulary maintainers themselves, especially those creating and maintaining inter-vocabulary links. For the first time these maintainers will be able to view both their own content and others' content at both atomic and concept levels, and be able to encode a richer repertoire of inter-vocabulary mappings as desired.

As the constituent vocabularies themselves become more explicit, more formal, and represent change more fully, the RRF representation will seem more and more natural; eventually, the RRF may not be "seen" at all as users become accustomed to the evolving model.

The cost to users of this increased source and longitudinal explicitness is mainly the increased size and complexity of the Metathesaurus release files. We expect that this growth will be easily accommodated by the ongoing decline in cost of secondary storage and processor power. It is a hypothesis that the increase in scale will be more than offset by increases in the simplicity of the required processing.

Conclusion

The introduction of the RRF allows the Metathesaurus to serve as a more formal and explicit vocabulary release management system for source vocabularies that are increasingly being specified as standards for creation and use of electronic health data in the U.S. and internationally. Importantly, the RRF enables the production of complete, specific change sets for individual source vocabularies and for the Metathesaurus itself. Such change sets are an essential part of the solution to the "update problem", that is, keeping local data creation applications in synchrony with changes in standard vocabularies as both the local systems and the standard vocabularies evolve.

The maintenance of the Metathesaurus involves an extreme example of the general vocabulary update and maintenance problem. The Metathesaurus must accommodate changes in scores of source vocabularies as well as its own value-added features. It is therefore not surprising that important features of the RRF were implemented first within the internal system that is used to maintain the Metathesaurus. As our understanding of how to address the update problem evolved over time, these features became mature enough to include. Although the slight concept-based abstraction in the ORF may be preferable for some natural language processing and cross-database retrieval applications, the RRF's added precision is likely to be essential for efficient updating of future data creation and retrieval systems.

Acknowledgements

The details of RRF have benefited greatly from UMLS users around the world who continue to share their experiences, successes and problems.

References

[1] Lindberg, DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inf Med.* 1993;32: 281-91.

[2] Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc* 1993 Apr;81(2):217-22

[3] Federal Government Announces First Federal eGov Health Information Exchange Standards. Available from: URL: http://www.hhs.gov/news/press/2003pres/20030321a.html

[4] BPHC - HIPAA transactions. Available from: URL: http://bphc.hrsa.gov/hipaa/hipaa_transactions.htm

[5] White Paper: UMLS Metathesaurus Rich Release (RRF) Format. Available from: URL: http://www.nlm.nih.gov/research/umls/white_paper.html.

[6] SNOMED International The Systematized Nomenclature of Medicine. Available from: URL: www.snomed.org

[7] Cancer.gov - Terminology Resources: NCI Thesaurus and Enterprise Vocabulary Services (EVS). Available from: URL: http://cancer.gov/cancerinfo/terminologyresources

[8] de Coronado, S, Haber, MW, Sioutos, N, Tuttle, MS, Wright, LW. Using Science-Based Terminology to Help Integrate Cancer Research Results, *MedInfo2004*.

[9] Spackman KA, Dionne R, Mays E, Weis J. Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED. *Proc AMIA Fall Symp*. 2002:712-6.

[10] UMLS Information: Customizing the UMLS Metathesaurus for Your Applications. http://umlsinfo.nlm.nih.gov/Customizing_the_UMLS.html

[11]Metathesaurus Production - Rich Release Format (Draft). Available from: URL: http://umlsinfo.nlm.nih.gov/rich_release_format.html

[12]Metathesaurus documentation. Available from: URL: http://www.nlm.nih.gov/research/umls/UMLSDOC.HTML

[13]RRF sample files. Available from: URL: http://umlsinfo.nlm.nih.gov/RRFsample.zip

[14]UMLS Objects API. Available from: URL: http://umlsinfo.nlm.nih.gov/Objects/

Address for correspondence

William T. Hole, MD,

National Library of Medicine, 8600 Rockville Pike,

Bethesda, MD 20894.

E-mail: wth@nlm.nih.gov.