UMLS® Reference Manual

Last Updated: September 2009

National Library of Medicine (US)
Bethesda (MD)

National Library of Medicine (US), Bethesda (MD)

NLM Citation: UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-.

The UMLS® Reference Manual describes the Unified Medical Language System (UMLS) Knowledge Sources and related tools that are produced and distributed by the National Library of Medicine, a part of the National Institutes of Health in the U.S. Department of Health and Human Services.

Table of Contents

Preface

Purpose of this Documentation

The UMLS® Reference Manual describes the UMLS Knowledge Sources and related tools that are produced and distributed by the National Library of Medicine, a part of the National Institutes of Health in the U.S. Department of Health and Human Services. This documentation explains the following procedures and concepts:

- The purpose, content, and file structure of the three UMLS Knowledge Sources: the Metathesaurus®, the Semantic Network, and the SPECIALIST Lexicon
- The way to use associated UMLS programs:
  - MetamorphoSys, the installation program for the three UMLS Knowledge Sources. MetamorphoSys is also the customization program for the Metathesaurus, providing several output options, including Rich Release and Original Release Formats, and a choice of preferred character sets to produce custom subsets of the Metathesaurus.
  - Lexical Programs, which help to deal with inflectional variation (e.g., treat, treats, treating, treatment) in the English language, convert American English to British English and vice versa, and map text to concepts in the Metathesaurus
- The way to access the UMLS resources using the UMLS Terminology Services, i.e., download, application programming interface (API), or Web browser
- The DVD distribution format for the UMLS Knowledge Sources and associated UMLS programs. The DVD was discontinued as of the 2012AB UMLS release.

Release Schedule

The UMLS is updated biannually in May (AA release) and November (AB release).

Audience

This documentation and the UMLS resources it describes are intended for system developers, informatics researchers, librarians, and other information professionals. The documentation assumes that you are familiar with database concepts and the Internet. If you intend to use the UMLS Knowledge Sources in software applications, it assumes that you have experience with building and using complex databases. If you intend to use any of the UMLS programs, it assumes basic familiarity with Java.

Neither the UMLS resources nor this documentation are intended for end users such as individual health professionals or members of the general public, unless they are also software developers.

A more general overview of the UMLS can be found on the UMLS Basics Web-based Tutorial.

Using this Documentation

Experienced UMLS Users

If you have done substantive work with preceding versions of the UMLS resources, go directly to UMLS Release Notes and Bugs page, which describes any changes in the documentation and in the UMLS resources. This page will direct you to the parts of the documentation that describe any changes to data files, content, or format introduced in the current release.

Novice UMLS Users

If you are new to the UMLS, you should read the rest of the Preface and all of Chapter 1 before moving on to other parts of this documentation. The following brief overview describes what you will find in each chapter of the documentation.

Chapter 1. Introduction to the UMLS

This chapter explains the purpose of the UMLS, the conditions under which you may use the different UMLS components, and how these conditions relate to Open Access/Open Source principles. It also briefly describes each of the UMLS components and the relationships between them, suggests ways to build your understanding of UMLS features and capabilities, and provides a list of additional UMLS reference materials.

Chapters 2-4. Metathesaurus

These chapters describe the content and structure of the Metathesaurus, a large concept-oriented database that incorporates numerous biomedical and health-related vocabularies, classifications, and coding systems. The Metathesaurus categorizes these concepts by assigned basic Semantic Types and makes all information from these terminologies accessible in common, fully-specified file formats. The Metathesaurus includes coding sets and terminologies designated by law and regulation as U.S. standards for electronic exchange of clinical and administrative health data.

Chapter 5. Semantic Network

This chapter describes the content and structure of the Semantic Network, a small database that includes information about the set of basic Semantic Types, or categories, to which Metathesaurus concepts may be assigned. The Semantic Network defines the relationships that may hold between these Semantic Types and between broad groupings of Semantic Types, such as all types that denote disorders (Disease or Syndrome, Acquired Abnormality, Neoplastic Process, etc.).

Chapter 6. SPECIALIST Lexicon and Lexical Tools

This chapter describes the content and structure of the following programs:

- The SPECIALIST Lexicon, a database of syntactic, morphological and orthographic information for commonly occurring English language words and biomedical vocabulary. The SPECIALIST Lexicon is useful for natural language processing applications.
- The Lexical Tools, which detect and abstract away from the inflectional, case, and word order variations encountered in natural language. One of these programs, MetaMap Transfer (MMTx), is specifically designed to map arbitrary terms to concepts in the Metathesaurus, or, equivalently, to discover Metathesaurus concepts within free text.

Chapter 7. UMLS Terminology Services

This chapter describes how to access the UMLS resources from the UMLS Terminology Services via download, application programming interface, and interactive Web browser.

Chapter 8. MetamorphoSys

This chapter describes MetamorphoSys, the installation program for all the UMLS Knowledge Sources and the customization program for the Metathesaurus. You *must* use MetamorphoSys to install the Knowledge Sources. MetamorphoSys allows you to output data in either the 7-bit ASCII (the default) or Unicode UTF-8 character set. MetamorphoSys also provides two file format options (Rich Release Format or Original Release Format) for the Metathesaurus, and provides a number of other customization options.

Chapter 9. UMLS DVD

This chapter gives technical specifications for the UMLS DVD, an alternative method for distributing UMLS content. The DVD was discontinued as of the 2012AB UMLS release.

Chapter 10. Current UMLS Release Information

This chapter briefly describes the UMLS release schedule and provides a link to the current UMLS release information.

1. Introduction to the UMLS

1.1. Purpose of the UMLS

The Unified Medical Language System (UMLS) facilitates the development of computer systems that behave as if they "understand" the language of biomedicine and health. To that end, NLM produces and distributes the UMLS Knowledge Sources (databases) and associated software tools (programs). Developers use the Knowledge Sources and tools to build or enhance systems that create, process, retrieve, and integrate biomedical and health data and information. The Knowledge Sources are multi-purpose and are used in systems that perform diverse functions involving information types such as patient records, scientific literature, guidelines, and public health data. The associated software tools assist developers in customizing or using the UMLS Knowledge Sources for particular purposes. The Lexical Tools work more effectively in combination with the UMLS Knowledge Sources, but can also be used independently.

1.2. Conditions of Use of the UMLS

All UMLS Knowledge Sources and associated software tools are free of charge to U.S. and international users.

The Semantic Network, the SPECIALIST Lexicon, and associated Lexical Tools are accessible on the Internet under open terms, which include appropriate acknowledgment for their use. View the terms and conditions for use of the Semantic Network and of the SPECIALIST Lexicon and Lexical Tools.

To use the Metathesaurus, you must establish a license agreement. This is because the Metathesaurus includes vocabulary content produced by many different copyright holders as well as the substantial content produced by NLM.

Setting up the license agreement is done via the Web. Once the license agreement is in place, much of the content of the Metathesaurus may be used under very open conditions. Your pre-existing licenses for content with use restrictions, e.g., CPT, MedDRA, or NIC, will cover your use of that content as distributed within the Metathesaurus. Some vocabulary producers who require authorization to use their content will generally grant free permission.

The complete text of the License Agreement for Use of the UMLS Metathesaurus appears on the UMLS Terminology Services (UTS), discussed in Chapter 7.

1.3. The UMLS Knowledge Sources and Associated Tools

There are three UMLS Knowledge Sources: the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon. They are distributed with several tools that facilitate their use, including the MetamorphoSys install and customization program.

1.3.1. Metathesaurus

The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts, their various names, and the relationships among them. It is built from the electronic versions of numerous thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing biomedical literature, and/or basic, clinical, and health services research. In this documentation, these are referred to as the "source vocabularies" of the Metathesaurus. In the Metathesaurus, all the source vocabularies are available in a common, fully-specified database format.

A complete list of the source vocabularies present in the current version of the Metathesaurus appears on the UMLS Source Vocabulary Documentation page of the current UMLS release documentation. The list indicates

which coding sets and terminologies are designated by law and regulation as U.S. standards for electronic exchange of clinical and administrative health data.

The Metathesaurus is organized by concept or meaning. In essence, it links alternative names and views of the same concept and identifies useful relationships between different concepts. All concepts in the Metathesaurus are assigned at least one Semantic Type from the Semantic Network (1.3.2) to provide consistent categorization at the relatively general level represented in the Semantic Network. Many of the words and multi-word terms that appear in concept names or strings in the Metathesaurus also appear in the SPECIALIST Lexicon (1.3.3). The Lexical Tools are used to generate the word, normalized word, and normalized string indexes to the Metathesaurus. MetamorphoSys (1.3.5) is used to install the UMLS Knowledge Sources and customize the Metathesaurus.

The Metathesaurus *must* be customized to be used effectively.

A complete description of the Metathesaurus and its file structure begins with Chapter 2 of this documentation.

1.3.2. Semantic Network

The Semantic Network provides a consistent categorization of all concepts represented in the Metathesaurus and provides a set of useful relationships between these concepts. All information about specific concepts is found in the Metathesaurus; the Network provides information about the set of basic Semantic Types, or categories, which may be assigned to these concepts, and it defines the set of relationships that may hold between the Semantic Types. The Semantic Network contains 133 Semantic Types and 54 relationships. The Semantic Network serves as an authority for the Semantic Types that are assigned to concepts in the Metathesaurus. The Network defines these types, both with textual descriptions and by means of the information inherent in its hierarchies.

The Semantic Types are the nodes in the Network, and the Semantic Relations between them are the links. There are major groupings of Semantic Types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. The current scope of the UMLS Semantic Types is quite broad, allowing for the semantic categorization of a wide range of terminology in multiple domains.

A complete description of the Semantic Network and its file structure appears in Chapter 5 of this documentation.

1.3.3. SPECIALIST Lexicon and Lexical Tools

The SPECIALIST Lexicon is intended to be a general English lexicon that includes many biomedical terms. Coverage includes both commonly occurring English words and biomedical vocabulary. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information needed by the SPECIALIST Natural Language Processing System.

The Lexical Tools are designed to address the high degree of variability in natural language words and terms. Words often have several inflected forms which would properly be considered instances of the same word. The verb "treat", for example, has three inflectional variants:

- treats — the third person singular present tense form
- treated — the past and past participle form
- treating — the present participle form

Multi-word terms in the Metathesaurus and other controlled vocabularies may have word order variants in addition to their inflectional and alphabetic case variants. The Lexical Tools allow the user to abstract away from

several types of variation, including British English/American English spelling variation and character set variations.

A complete description of the SPECIALIST Lexicon, its file structure, and the lexical programs appears in Chapter 6 of this documentation.

1.3.4. UMLS Terminology Services

The UMLS Terminology Services (UTS) is a set of Web-based interactive tools and a programmer interface that allows users and developers to access the UMLS Knowledge Sources, including the vocabularies within the Metathesaurus. It also contains the download site for the UMLS data files. The UTS is a useful starting point for gaining an understanding of the content of the UMLS resources. Because it contains the complete Metathesaurus files, access to many UTS components is restricted to registered users who have signed the License Agreement for Use of the UMLS Metathesaurus.

A complete description of the UTS and its capabilities appears in Chapter 7 of this documentation.

1.3.5. MetamorphoSys: The UMLS Installation and Customization Program

MetamorphoSys is a cross-platform Java application that must be used if the UMLS Knowledge Sources (Metathesaurus, Semantic Network, and SPECIALIST Lexicon) are installed locally. MetamorphoSys also supports the creation and refinement of customized subsets of the Metathesaurus. In general, the Metathesaurus must be customized to be used effectively in specific applications.

MetamorphoSys guides you first through the installation of one or more UMLS Knowledge Sources, and then through customization of the Metathesaurus for local use. A variety of options are available, such as the inclusion or exclusion of specific source vocabularies, languages, and term types, specification of output character set (7-bit ASCII or Unicode UTF-8) and output format (Rich Release Format or Original Release Format) for the Metathesaurus files.

A complete description of MetamorphoSys appears in Chapter 8 of this documentation.

1.4. Getting Started

The UMLS resources are powerful - and unusual - tools intended for use by system developers. Here are a few suggestions about how to start building your understanding of UMLS features and capabilities and their potential for enhancing your applications.

Scan the entire UMLS documentation to get a sense of the range of resources available.

If the Metathesaurus interests you, take time to read Chapter 2 of the documentation. The background there will make it easier to understand the actual file descriptions in Chapter 3 and Chapter 4.

Use the UMLS Terminology Services to request a License Agreement for Use of the UMLS Metathesaurus. A license agreement is required because the Metathesaurus contains vocabularies produced by many different copyright holders. You are able to use much of the content of the Metathesaurus with minimal restriction, but you may need to obtain additional licenses from individual vocabulary producers if you wish to use certain vocabularies contained in the Metathesaurus. The various restriction levels are explained in the UMLS license agreement.

Once you have requested a license and activated your UTS account, use the UTS for initial browsing and exploration of the contents of the Metathesaurus, Semantic Network, and SPECIALIST Lexicon and of additional special resources useful to application developers.

If you require local copies of the UMLS files, use the MetamorphoSys install and customization program described in Chapter 8 of this documentation to produce them. You may find it useful to experiment with various options to produce customized subsets. MetamorphoSys is available for download with the UMLS data files from the UTS.

1.5. Sources of Additional Information about the UMLS

In addition to providing links to the UMLS documentation and to the UTS, the UMLS Web site links to fact sheets on the UMLS Knowledge Sources and UTS; FAQs; training materials; and information about NLM applications and research projects that use the UMLS. The UMLS Quick Start Guide provides a brief overview of the UMLS and includes links to more detailed information. Articles on the UMLS project and resources can be retrieved from MEDLINE/PubMed. Click here to obtain a current search. A comprehensive 1986-1996 bibliography on the UMLS project covering additional papers not indexed for MEDLINE/PubMed is also available.

UMLS users are strongly encouraged to subscribe to the UMLS users listserv. NLM uses the listserv to seek advice from users and to distribute news about upcoming UMLS developments; users share experiences or obtain advice about using the UMLS resources.

To subscribe, send an e-mail to listserv@list.nih.gov containing the following message: SUBSCRIBE UMLSUSERS-L <your full name>.

To unsubscribe, send an e-mail to listserv@list.nih.gov containing the following message: SIGNOFF UMLSUSERS-L <your full name>.

To post a message to the list AFTER subscribing, send an e-mail to UMLSUSERS-L@list.nih.gov.

To access subscription information and list archives, go to UMLSUSERS-L Listserv Web page.

2. Metathesaurus

2.1. Overview

The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. Designed for use by system developers, the Metathesaurus is built from the electronic versions of various thesauri, classifications, code sets, and lists of controlled terms used in patient care, health services billing, public health statistics, indexing and cataloging biomedical literature, and/or basic, clinical, and health services research. These are referred to as the "source vocabularies" of the Metathesaurus. The term Metathesaurus draws on Webster's Dictionary third definition for the prefix "meta," i.e., "more comprehensive, transcending." In a sense, the Metathesaurus transcends the specific thesauri, vocabularies, and classifications it encompasses.

The Metathesaurus is organized by concept or meaning. In essence, it links alternative names and views of the same concept and identifies useful relationships between different concepts.

The Metathesaurus is linked to the other UMLS Knowledge Sources – the Semantic Network and the SPECIALIST Lexicon. All concepts in the Metathesaurus are assigned to at least one Semantic Type from the Semantic Network. This provides consistent categorization of all concepts in the Metathesaurus at the relatively general level represented in the Semantic Network. Many of the words and multi-word terms that appear in concept names or strings in the Metathesaurus also appear in the SPECIALIST Lexicon. The Lexical Tools are used to generate the word, normalized word, and normalized string indexes to the Metathesaurus.

MetamorphoSys is the software tool for customizing the Metathesaurus for specific purposes. MetamorphoSys is also the installation program for all of the UMLS resources. UMLS licensees can download the UMLS Knowledge Sources from the UMLS Web site. To ensure proper functionality you should download and extract all UMLS data and zip files to the same directory.

2.1.1. Scope of the Metathesaurus

The scope of the Metathesaurus is determined by the combined scope of its source vocabularies. Many relationships (primarily synonymous), concept attributes, and some concept names are added by the NLM during Metathesaurus creation and maintenance, but essentially all the concepts themselves come from one or more of the source vocabularies. Generally, if a concept does not appear in any of the source vocabularies, it will also not appear in the Metathesaurus.

2.1.2. Preservation of Content and Meaning from Source Vocabularies

The Metathesaurus reflects and preserves the meanings, concept names, and relationships from its source vocabularies. When two different source vocabularies use the same name for differing concepts, the Metathesaurus represents both of the meanings and indicates which meaning is present in which source vocabulary. When the same concept appears in different hierarchical contexts in different source vocabularies, the Metathesaurus includes all the hierarchies. When conflicting relationships between two concepts appear in different source vocabularies, both views are included in the Metathesaurus. Although specific concept names or relationships from some source vocabularies may be idiosyncratic and lack face validity, they are still included in the Metathesaurus.

In other words, the Metathesaurus does not represent a comprehensive NLM-authored ontology of biomedicine or a single consistent view of the world (except at the high level of the semantic types assigned to all its concepts). The Metathesaurus preserves the many views of the world present in its source vocabularies because these different views may be useful for different tasks.

Although it preserves all the meanings and content in its source vocabularies, the Metathesaurus stores this information in a single common format. The native format of each vocabulary is carefully studied and then "inverted" into the common Metathesaurus format. For some vocabularies, this involves representing implied information in a more explicit format. For example, if a source vocabulary stores its preferred concept name as the first occurrence in a list of alternative concept names, that first name is explicitly tagged as the preferred name for that source in the Metathesaurus.

2.1.3. Need to Customize the Metathesaurus

Because it is a multi-purpose resource that includes concepts and terms from many different source vocabularies developed for very different purposes, **the Metathesaurus must be customized for effective use in most specific applications**. Your decisions about what to include in your customized subset(s) of the Metathesaurus will have a significant effect on its utility in your systems. Vocabulary sources that are essential for some purposes, e.g., LOINC for standard exchange of laboratory data, may be detrimental for others, such as Natural Language Processing (NLP). It can also be important to exclude a subset of the concept names found in a vocabulary source that is otherwise useful, e.g., non-standard abbreviations or shortened forms that lack face validity or produce spurious results in NLP.

The Metathesaurus contains source vocabularies produced by many different copyright holders. The majority of the content of the Metathesaurus is available for use under the basic (and quite open) terms described in Sections 1-11 and 13-16 of the Metathesaurus license. However, some vocabulary producers place additional restrictions on the use of their content as distributed within the Metathesaurus. The various levels of additional restrictions are described in Section 12 of the license. The level that applies to individual vocabularies is recorded on the UMLS Source Vocabulary Documentation page of the current UMLS release documentation and in the MetamorphoSys installation and customization program. If you already have a separate license for use of one of the source vocabularies, your existing license also applies to that source as distributed within the Metathesaurus. In some cases, you may have to request permission or negotiate a separate license with a vocabulary producer in order to use that vocabulary in a production system. There may be a charge associated with these separate permissions or license agreements.

The Metathesaurus is designed to facilitate customization. All information in the Metathesaurus is labeled as to its source(s), so it is possible to determine which concept names, attributes, and relationships come from which source vocabularies and which attributes and relationships were added during Metathesaurus construction. The labels allow you to subset the Metathesaurus by excluding information from specific source vocabularies, including those for which you do not have necessary licenses or permissions. It is also easy to exclude all source vocabularies that have particular restriction levels or all information in particular languages. In addition to identifying the source(s), restriction levels, and language of the information it contains, the Metathesaurus includes various more specific concept name flags and relationship labels that can help you to exclude content that is not relevant or helpful for particular applications.

MetamorphoSys, the installation and customization program distributed with the UMLS, makes it easy to generate custom subsets. MetamorphoSys also includes default settings that generate subsets that may be generally useful. MetamorphoSys can also be used to change the default preferred names of concepts; to change the default character set (from 7-bit ASCII to Unicode UTF8); and to include versioned vocabulary source abbreviations in every Metathesaurus file.

2.1.4. Metathesaurus Release Formats

You may select from two relational formats: the Rich Release Format (RRF), introduced in 2004, and the Original Release Format (ORF). Both are available as output options of MetamorphoSys. All Rich Release

Format file names have an extension (.RRF). Original Release Format files have no extension. Both formats are described in Chapter 3 and Chapter 4 (usually abbreviated as RRF and ORF).

The Rich Release Format has a number of advantages and is the preferred format for new users of the Metathesaurus and for most data creation applications.

2.2. Source Vocabularies

The Metathesaurus contains concepts, concept names, and other attributes from more than 100 terminologies, classifications, and thesauri, some in multiple editions. There is a concept in the Metathesaurus for each source vocabulary itself, which is assigned the Semantic Type "Intellectual Product". A special file (MRSAB.RRF and MRSAB in ORF) stores the version of each source vocabulary present in a particular edition of the Metathesaurus. All other Metathesaurus files that reference source vocabularies use "root" or versionless abbreviations, e.g., ICD9CM, not ICD9CM2003, thus avoiding routine wholesale updates to reflect the new versions. If you prefer versioned vocabulary source abbreviations in your custom Metathesaurus subset files, MetamorphoSys offers this option.

A complete list of the Metathesaurus source vocabularies with their root and versioned source abbreviations appears on the UMLS Source Vocabulary Documentation page of the current UMLS release documentation. The list is alphabetized by the abbreviation for that vocabulary source that is used in the Metathesaurus. The UMLS Source Vocabulary Documentation page includes other information: the number of its concept names that are present in the Metathesaurus, the type of hierarchies or contexts it has (if any), and whether it is one of the small number of source vocabularies that is not routinely updated in the Metathesaurus.

The Metathesaurus source vocabularies include terminologies designed for use in patient-record systems; large disease and procedure classifications used for statistical reporting and billing; more narrowly focused vocabularies used to record data related to psychiatry, nursing, medical devices, adverse drug reactions, etc.; disease and finding terminologies from expert diagnostic systems; and some thesauri used in information retrieval. A categorized list of the English-language source vocabularies is available.

2.2.1. Inclusion of U.S. Standard Code Sets and Terminologies

The Metathesaurus includes terminologies and code sets mandated for U.S. use in electronic exchange of clinical and administrative health data.

2.2.2. Inclusion of Languages Other Than English

The Metathesaurus structure can accommodate translations of its source vocabularies into languages other than English. Many translations in many different languages are present in the current edition of the Metathesaurus. The Metathesaurus includes many translations of some source vocabularies, e.g., NLM's Medical Subject Headings (MeSH) and the International Classification of Primary Care; one or a few of others, and, in many cases, only the English version. As previously explained, MetamorphoSys makes it easy to create a subset of the Metathesaurus that excludes the languages that are not relevant in a particular application.

2.3. Concepts, Concept Names, and Their Identifiers

The Metathesaurus is organized by concept. One of its primary purposes is to connect different names for the same concept from many different vocabularies. The Metathesaurus assigns several types of unique, permanent identifiers to the concepts and concept names it contains, in addition to retaining all identifiers that are present in the source vocabularies. The Metathesaurus concept structure includes concept names, their identifiers, and key characteristics of these concept names (e.g., language, vocabulary source, name type). The entire concept

structure appears in a single file in the Rich Release Format (MRCONSO.RRF). An abbreviated version of the concept structure is split between two files in the Original Release Format (MRCON and MRSO).

2.3.1. Concepts and Concept Identifiers

A concept is a meaning. A meaning can have many different names. A key goal of Metathesaurus construction is to understand the intended meaning of each name in each source vocabulary and to link all the names from all of the source vocabularies that mean the same thing (the synonyms). This is not an exact science. The construction of the Metathesaurus is based on the assumption that specially trained subject experts can determine synonymy with a high degree of accuracy. Metathesaurus editors decide what view of synonymy to represent in the Metathesaurus concept structure. Please note that each source vocabulary's view of synonymy is also present in the Metathesaurus, irrespective of whether it agrees or disagrees with the Metathesaurus view.

Each concept or meaning in the Metathesaurus has a unique and permanent concept identifier (CUI). The CUI has no intrinsic meaning. In other words, you cannot infer anything about a concept just by looking at its CUI. In principle, the identifier for a concept never changes, irrespective of changes over time in the names that are attached to it in the Metathesaurus or in the source vocabularies.

A CUI will be removed from the Metathesaurus when it is discovered that two CUIs name the same concept – in other words, when undiscovered synonymy comes to light. In these cases, one of the two CUIs will be retained, all relevant information in the Metathesaurus will be linked to it, and the other CUI will be retired.

Retired CUIs are never re-used. Each edition of the Metathesaurus includes files that detail any such changes from the previous edition. One Metathesaurus file (MRCUI.RRF and MRCUI in ORF) tracks such changes from 1991 to the present, allowing you to check the fate of any CUI that is no longer present in the Metathesaurus.

2.3.2. Concept Names and String Identifiers

Each unique concept name or string in each language in the Metathesaurus has a unique and permanent string identifier (SUI). Any variation in character set, upper-lower case, or punctuation is a separate string, with a separate SUI. The same string in different languages (e.g., English and Spanish) will have a different string identifier for each language. If the same string, e.g., Cold, has more than one meaning, the string identifier will be linked to more than one concept identifier (CUI).

2.3.3. Atoms and Atom Identifiers

The basic building blocks or "atoms" from which the Metathesaurus is constructed are the concept names or strings from each of the source vocabularies. Every occurrence of a string in each source vocabulary is assigned a unique atom identifier (AUI). If exactly the same string appears twice in the same vocabulary, for example, as both the long name and the short name for the same concept or as an alternate name for two different concepts in the same vocabulary source, a unique AUI is assigned for each occurrence. When the same string appears in multiple source vocabularies, it will have AUIs for every time it appears as a concept name in each of those sources. All of these AUIs will be linked to a single string identifier (SUI), since they represent occurrences of the same string. Unlike string identifiers, a single AUI is always linked to a single concept identifier, because each occurrence of a string in a source can only have one meaning.

AUIs appear in the RRF (.RRF files), but not in the ORF.

2.3.4. Terms and Lexical Identifiers

For English language entries in the Metathesaurus only, each string is linked to all of its lexical variants or minor variations by means of a common term identifier (LUI). (In the Metathesaurus, therefore, an English "term" is the group of all strings that are lexical variants of each other.) English lexical variants are detected using the

Lexical Variant Generator (lvg) program, one of the UMLS Lexical Tools. As similar tools become available for other languages, they may be used to create lexical variant groups in other languages. (In the meantime, the LUI for a non-English string is really another string identifier.)

Like a string identifier, the LUI for an English string may be linked to more than one concept. This occurs when strings that are lexical variants of each other have different meanings. In contrast, each string identifier and each atom identifier can only be linked to a single LUI.

2.3.5. Uses of Concept, String, Atom, and Term Identifiers

In the Metathesaurus, every CUI (concept) is linked to at least one AUI (atom), SUI (string), and LUI (term), but can also be linked to many of each of these. Every AUI (atom) is linked to a single SUI (string), a single LUI (term), and a single CUI (concept). Each SUI (string) can be linked to many AUIs (atoms), to a single LUI (term), and to more than one CUI (concept) – although the typical case is one CUI. Each LUI (term) can be linked to many AUIs (atoms), many SUIs (strings), and more than one CUI (concept) – although the typical case is one CUI.

In the abbreviated example in Table 1, Atrial Fibrillation appears as an atom in more than one source vocabulary and has a distinct AUI for each occurrence. Since each of these atoms has an identical string or concept name, they are linked to a single SUI. Atrial Fibrillations, the plural of Atrial Fibrillation, has a different string identifier. Since the singular and plural are lexical variants of each other, both are linked to the same LUI. There is a different LUI and different SUIs and AUIs for Auricular Fibrillation and its plural Auricular Fibrillations. Since Atrial Fibrillation and Auricular Fibrillation have been judged to have the same meaning, they are linked to the same CUI.

All of these identifiers serve important purposes in building the Metathesaurus, in allowing efficient and accurate customization for specific purposes, and in identifying changes in its concept and concept name coverage over time.

For example, CUIs link all information in the Metathesaurus related to particular concepts. In other words, a CUI can be used to retrieve all the concept names, relationships, and attributes for a particular concept that appear in any Metathesaurus file. CUIs also serve as permanent, publicly available identifiers for biomedical concepts or meanings to which many individual source vocabularies are linked. You are strongly encouraged to incorporate CUIs in your local applications – to support data exchange and linking and to assist migration between the use of individual source vocabularies should that become necessary in the future.

Table 1. Concept, Term, Atom, and String Identifiers.

Concept (CUI)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF Only
C0004238 Atrial Fibrillation (preferred) Atrial Fibrillations Auricular Fibrillation Auricular Fibrillations	**L0004238** Atrial Fibrillation (preferred) Atrial Fibrillations	**S0016668** Atrial Fibrillation (preferred)	**A0027665** Atrial Fibrillation (from MSH) **A0027667** Atrial Fibrillation (from PSY)
		S0016669 (plural variant) Atrial Fibrillations	**A0027668** Atrial Fibrillations (from MSH)
	L0004327 (synonym) Auricular Fibrillation Auricular Fibrillations	**S0016899** Auricular Fibrillation (preferred)	**A0027930** Auricular Fibrillation (from PSY)
		S0016900 (plural variant) Auricular Fibrillations	**A0027932** Auricular Fibrillations (from MSH)

2.3.6. Default Preferred Names for Metathesaurus Concepts

As a convenience for those who build the Metathesaurus, one string from one English term is designated and labeled as the default preferred name of each concept in the Metathesaurus. To avoid laborious selection among alternative terms and strings, selection of the default preferred name for any Metathesaurus concept is based on an order of precedence of all the types of English strings in all the Metathesaurus source vocabularies. Different types of strings, e.g., preferred terms, cross references, and abbreviations from each vocabulary, will have different positions in this order. The factors considered in establishing the default order of precedence include breadth of subject coverage, frequency of update, and the degree to which the source's concept names are used in regular clinical or biomedical discourse. The default order of precedence appears in MRRANK.RRF (MRRANK in ORF), and on the Source and Term Types: Default Order of Precedence and Suppressibility page of the current UMLS release documentation.

The default order of precedence will not be suitable for all applications of the Metathesaurus. The MetamorphoSys can be used to change the selection of preferred names to feature terminology from the source vocabularies most appropriate to particular user populations. For example, concept names from SNOMED CT may be preferred in clinical applications, and terminology from MeSH may be preferred in literature retrieval systems.

2.3.7. Strings with Multiple Meanings

In some cases, the same name (with or without differences in upper-lower case) may apply to different concepts, usually (but not always) in different Metathesaurus source vocabularies. In the abbreviated example that follows, the string "Cold" is a name for the temperature in one vocabulary. In another vocabulary, "Cold" is an alternate name for the "Common cold". In a third vocabulary, "COLD" is an acronym for "chronic obstructive lung disease". As a result, "Cold" or "COLD" appears as a name of more than one concept in the Metathesaurus.

2.3.7.1. Representation of Ambiguity in the Metathesaurus

Separate Metathesaurus files (AMBIGLUI.RRF and AMBIGSUI.RRF (AMBIG.LUI and AMBIG.SUI in ORF)) contain the LUIs and SUIs of all ambiguous terms and strings known to the Metathesaurus. See Table 2.

Table 2. Representation of Ambiguity in the Metathesaurus.

Concepts (CUIs)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF only
C0009264 **Cold Temperature**	**L0215040** cold temperature	**S7669511** Cold Temperature	**A15594156** Cold Temperature (from MTH)
	L0009264 cold	**S0026353** Cold	**A0040709** Cold (from LCH)
			A4711382 Cold (from SNOMEDCT)
C0009443 **Common Cold**	**L0009443** cold common	**S0026747** Common Cold	**A0041261** Common Cold (from MSH)
	L0009264 cold	**S0026353** Cold	**A0040708** Cold (from COSTAR)
			A2880095 Cold (from SNOMEDCT)
C0024117 **Chronic Obstructive Airway Disease**	**L0498186** airway chronic disease obstructive	**S0837575** Chronic Obstructive Airway Disease	**A0896021** Chronic Obstructive Airway Disease (from MSH)
	L0008703 chronic disease lung obstructive	**S0837576** Chronic Obstructive Lung Disease	**A0896023** Chronic Obstructive Lung Disease (from MSH)
	L0009264 cold	**S0474508** COLD	**A10765219** COLD (from NCI)
			A0539536 COLD (from SNMI)

2.3.8. Concept Names Added During Metathesaurus Construction

Although the majority of concept names present in the Metathesaurus come from one or more of its source vocabularies, some concept names are created during Metathesaurus construction. This occurs in the following circumstances:

1. A unique name is created for a string with multiple meanings (the case explained in Section 2.3.7)
2. A more explicit name is created when none of the source vocabulary names for a concept conveys its meaning adequately
3. An American English variant is generated for a British spelling
4. An equivalent basic Latin ASCII character set string is generated for a string in an extended character set, such as Unicode

Like all other concept names in the Metathesaurus, names created during Metathesaurus construction are labeled to indicate their source.

2.4. Relationships and Relationship Identifiers

The Metathesaurus includes many relationships between different concepts (in addition to the synonymous relationships in the Metathesaurus concept structure described in Section 2.3). Most of these relationships come from individual source vocabularies. Some are added by NLM during Metathesaurus construction. Some have been contributed by Metathesaurus users to support certain types of applications.

Relationships are expressed in terms of CUIs (in the RRF and ORF) and AUIs (in the RRF only). Metathesaurus relationship files do not include concept names.

In general, the Metathesaurus indicates the author of each relationship, that is, one of the source vocabularies, the Metathesaurus itself, or another supplier. Some relationships added in the early years of Metathesaurus development (less than 6 percent of the current total and declining) are attributed to the Metathesaurus, but actually came from specific source vocabularies.

2.4.1. Basic Categories of Non-Synonymous Relationships

The Metathesaurus contains non-synonymous relationships between concepts from the same source vocabulary (intra-source vocabulary relationships) and between concepts in different vocabularies (inter-source vocabulary relationships). **The Metathesaurus does not include all possible non-synonymous relationships between the concepts it contains**. It includes all relationships present in its source vocabularies and some additional relationships designed to connect related concepts. In general, the relationships asserted by source vocabularies connect closely related concepts, such as those that share some common property or are related by definition. For example, a member of a class of drugs (e.g., penicillin) will be connected to the name for the class (e.g., antibiotics); a bacterial infection will be connected to the bacterium that causes it.

2.4.1.1. Intra-Source Relationships

The majority of intra-source relationships are asserted or implied by the individual source vocabularies. Such relationships occur in a source vocabulary's explicit or implied hierarchical arrangements or contexts, cross-reference structures, rules for applying qualifiers, or connections between different types of names for the same concept (e.g., abbreviations and full forms). The primary Metathesaurus relationships file, that is, MRREL.RRF and MRREL in the ORF contains the "distance -1" hierarchical relationships, i.e., immediate parents, immediate child, and immediate sibling relationships, as well as other types of intra-source relationships.

A subset of the contextual or hierarchical relationships is also distributed in a special contexts file (MRCXT.RRF and MRCXT in ORF) to facilitate the construction of user displays. A "computable" representation of the complete hierarchies is provided in MRHIER.RRF only. MRHIER.RRF, for example, represents all sibling relationships even when there are thousands of siblings. The UMLS Source Vocabulary Documentation page indicates which source vocabularies have hierarchical contexts, which of these allow concepts to appear in multiple hierarchies, and whether sibling relationships are represented in MRCXT.RRF and MRCXT in ORF or only in MRHIER.RRF.

ORF users may omit MRCXT if they do not want these selected, pre-computed contexts.

2.4.1.2. Inter-Source Relationships

The primary inter-source relationships in the Metathesaurus are the synonymous relationships represented in the Metathesaurus concept structure. The Metathesaurus also includes some relationships between non-synonymous concepts from different source vocabularies. Some of these inter-source relationships are generated

during Metathesaurus construction to connect specific "orphan" concepts (with few or no ancestors, siblings, or children in their own source vocabularies) to the richer contextual information in another source vocabulary. Some are supplied by Metathesaurus users who find "like" or "similar" relationships a useful addition to the Metathesaurus's relatively strict view of synonymy. In both cases, these relationships are distributed in MRREL.RRF and MRREL in ORF.

Many inter-source relationships between non-synonymous concepts are produced through specific efforts to create a mapping between two different source vocabularies. These mappings may be created by an individual source vocabulary producer, by a third party with a particular need for a mapping, or by NLM or under NLM supervision specifically for distribution within the Metathesaurus. The number of NLM-supervised mappings is expected to increase. There are specific Metathesaurus files for mappings in the RRF (MRMAP.RRF and MRSMAP.RRF). A subset of the mappings appears in MRATX in the ORF. Mappings involving SNOMED CT appear in the RRF only.

2.4.2. Relationship Labels

All relationships (outside the basic concept structure) in the Metathesaurus carry a general label (REL), describing their basic nature, such as Broader, Narrower, Child of, Qualifier of, etc., and are identified by their source. Most of these relationships are either directly asserted in a source vocabulary or are implied by the structure of the source vocabulary. A complete list of the general relationship labels appears in MRDOC.RRF and on the Abbreviations Used in Data Elements page of the current UMLS release documentation.

About a quarter of the relationships in the Metathesaurus also carry an additional label (RELA), obtained from a source vocabulary, that explains the nature of the relationship more exactly, such as is_a, branch_of, component_of. The Digital Anatomist vocabulary and RxNorm are examples of source vocabularies that include such relationship labels. A complete list of the additional relationship labels appears in MRDOC.RRF and on the Abbreviations Used in Data Elements page of the current UMLS release documentation.

2.4.3. Relationship Identifiers

Every relationship present in the Metathesaurus has a unique relationship identifier (RUI). The primary purpose of these identifiers is to enable easy detection of changes in relationships across versions of the Metathesaurus. The appearance or disappearance of a relationship identifier indicates a change in the relationships present in the Metathesaurus.

Some source vocabularies have their own relationship identifiers. Where they exist, these identifiers are also present in the Metathesaurus.

2.4.4. Relationship Groups

Relationship groups are source-asserted or implied associations of relationships that may be used to add meaning or clarity when multiple relationships are present. Together, each grouping may express a richer meaning than it would have ungrouped. Relationship groups can be identified in MRREL.RRF by rows that have the same AUI2 and the same numeric value for relationship group (RG). Numeric values for relationship group increase according to the number of relationship groups associated with the same AUI2. A null value indicates no grouping for the relationship is present. Relationships may be suppressible if considered obsolete, which is indicated by a value of O in the SUPPRESS field.

CUI1 | AUI1 | STYPE1 | REL | CUI2 | AUI2 | STYPE2 | RELA | RUI | SRUI | SAB | SL | RG | DIR | SUPPRESS | CVF

C0024109|A3154872|SCUI|RO|C0264408|**A2957612**|SCUI|has_finding_site|R14028961|994883025|SNOMEDCT_US|SNOMEDCT_US|0|Y|O||

C0231335|A2926532|SCUI|RO|C0264408|**A2957612**|SCUI|occurs_in|R123147138|1795540028|SNOMEDCT_US|SNOMEDCT_US|0|Y|N||

C0006255|A3104303|SCUI|RO|C0264408|**A2957612**|SCUI|has_finding_site|R98157815|3465258024|SNOMEDCT_US|SNOMEDCT_US|1|Y|O||

C0028778|A2873893|SCUI|RO|C0264408|**A2957612**|SCUI|has_associated_morphology|R98053314|3419439024|SNOMEDCT_US|SNOMEDCT_US|1|Y|O||

In this example, Relationship Group 0 groups the relationships from A2957612 "Childhood asthma" to A2926532 "Childhood" and A3154872 "Lung structure" to clarify that "Childhood asthma" is found in the lungs of children.

Relationship Group 1 groups the relationships from A2957612 "Childhood asthma" to A3104303 "Bronchial structure" and A2873893 to indicate that "Childhood asthma" is characterized by obstruction of the bronchi.

In SNOMEDCT_US, relationship groups are source-asserted. The SRUI field of MRREL.RRF contains the SNOMEDCT-asserted unique identifier assigned to the relationship. Please refer to the SNOMED CT Technical Implementation Guide for a detailed description of relationship groups in SNOMED CT.

In MeSH and MedlinePlus, relationship groups are source-implied by a Descriptor (SDUI) and Qualifier that appear under the same MeSH Mapped To Heading. Example from MeSH:

```
<HeadingMappedToList>
   <HeadingMappedTo>
    <DescriptorReferredTo>
     <DescriptorUI>D012694</DescriptorUI>
     <DescriptorName>
      <String>Serine</String>
     </DescriptorName>
    </DescriptorReferredTo>
    <QualifierReferredTo>
     <QualifierUI>*Q000031</QualifierUI>
     <QualifierName>
      <String>analogs &amp;amp; derivatives</String>
     </QualifierName>
    </QualifierReferredTo>
   </HeadingMappedTo>
</HeadingMappedToList>
```

This association is represented in UMLS as a grouping of mapped_to and has_mapping_qualifier relationships that connect the AUIs corresponding to the DescriptorName and QualifierName strings to the AUI of the MeSH Mapped To Heading.

CUI1 | AUI1 | STYPE1 | REL | CUI2 | AUI2 | STYPE2 | RELA | RUI | SRUI | SAB | SL | RG | DIR | SUPPRESS | CVF

C0002776|A3879704|SDUI|RO|C0067636|**A0207764**|SDUI|**has_mapping_qualifier**|R148279824||MSH|MSH|1||N||

C0036720|A0115503|SDUI|RN|C0067636|**A0207764**|SDUI|**mapped_to**|R148155946||MSH|MSH|1||N||

The relationship group links the relationships to clarify that A0207764 "N-acetyl-4-nitrophenylserinol" is mapped to A0115503 "Serine" and has mapping qualifier A3879704 "analogs & derivatives".

2.5. Attributes and Attribute Identifiers

In the Metathesaurus, attributes include every discrete piece of information about a concept, an atom, or a relationship that is not (1) part of the basic Metathesaurus concept structure or (2) distributed in one of the relationship files.

2.5.1. Kinds of Attributes

The Metathesaurus includes concept attributes, atom attributes, and relationship attributes.

Concept attributes are added during Metathesaurus construction and apply to all names of a concept. For example, the Semantic Types "Pathologic Function" and "Finding" are attributes of the concept with the preferred name "Atrial Fibrillation" and are applicable to any atom connected to that concept.

Atom attributes come from a particular source vocabulary. Some of them are of general interest; others are relevant only to a particular source vocabulary. For example, the definition "Disorder of cardiac rhythm characterized by rapid, irregular atrial impulses and ineffective atrial contractions" is an attribute of the atom Atrial Fibrillation that comes from the Medical Subject Headings (MeSH). It may be one of several definitions connected to names of this concept, because the Metathesaurus includes all definitions provided by any of its source vocabularies. Although this particular definition comes from MeSH, it might well be useful in Metathesaurus applications that otherwise do not use MeSH. In contrast, the date an occurrence of a string (an atom) was added to a source vocabulary applies only to that specific atom. The utility of specific atom attributes will vary considerably for different applications of the Metathesaurus.

Relationship attributes come from a particular source vocabulary and describe special characteristics of particular relationships in that source, e.g., refinability.

The majority of attributes are distributed in MRSAT.RRF and MRSAT in the ORF. In these files, each row contains the name of the attribute, the source of the attribute, and the value of the attribute, in addition to all appropriate identifiers. There are separate files for selected attributes such as the Semantic Types (MRSTY.RRF and MRSTY in the ORF) and the definitions (MRDEF.RRF and MRDEF in the ORF).

2.5.2. Attribute Identifiers

Each occurrence of each attribute within the Metathesaurus is assigned a unique attribute identifier (ATUI). The appearance or disappearance of ATUIs signals changes in the content of the Metathesaurus, thus ATUIs assist the efficient production of a complete change set for each new version of the Metathesaurus. ATUIs appear only in the RRF, not in the ORF.

2.6. Data About the Metathesaurus

The Metathesaurus contains a number of files that provide useful metadata, i.e., data about the Metathesaurus itself. The metadata files describe (1) characteristics of the current version of the Metathesaurus; (2) changes between the current version and the previous version; and (3) the history of concept identifiers (CUIs) from 1991 to the present.

2.6.1. Characteristics of the Current Metathesaurus

There are discrete Metathesaurus files for:

- The names and sizes of every Metathesaurus file (MRFILES.RRF and MRFILES in ORF)
- The names and size range of every Metathesaurus data element (MRCOLS.RRF and MRCOLS in ORF)

- The possible values for selected data elements that contain a finite set of abbreviated values (MRDOC.RRF only). Note: Eventually this file will include values for every data element that contains a finite set of abbreviated values.
- The source vocabularies in the Metathesaurus (MRSAB.RRF and MRSAB in ORF)
- The LUIs and SUIs for terms and strings that are known to be ambiguous, that is, to have multiple meanings (to be linked to multiple concept identifiers) within the Metathesaurus (AMBIGLUI.RRF and AMBIGSUI.RRF in RRF and AMBIGLUI and AMBIGSUI in ORF)
- The order of precedence of vocabulary source and term types that is used to compute the default preferred concept name for each concept in the Metathesaurus (MRRANK.RRF and MRRANK in ORF). Note: MetamorphoSys can be used to change this order.

MRCOLS, MRDOC, MRSAB, and MRRANK contain data that do not appear in the actual Metathesaurus content files. The others are computable from the Metathesaurus content files. They are pre-computed and provided in separate files as a convenience to users.

2.6.2. Changes Between the Current Metathesaurus and the Previous Version

Each version of the Metathesaurus contains a set of files that summarize changes from the previous version.

CHANGE/MERGEDCUI.RRF in the RRF (CHANGE/MERGED.CUI in the ORF) documents cases in which two discrete concepts in the previous version of the Metathesaurus are now considered to be synonyms.

CHANGE/MERGEDLUI.RRF in the RRF (CHANGE/MERGED.LUI in the ORF) documents cases in which two discrete terms in the previous version of the Metathesaurus are now identified as lexical variants of each other, based on the current version of luinorm (the program used to compute them).

Three files contain the CUIs, LUIs, and SUIs for Metathesaurus concepts, terms, and strings that appeared in the previous version, but are not in the current version (CHANGE/DELETEDCUI.RRF, CHANGE/DELETEDLUI.RRF, CHANGE/DELETEDSUI.RRF in the RRF and CHANGE/DELETED.CUI, CHANGE/DELETED.LUI, CHANGE/DELETED.SUI in the ORF).

Note: Future versions of the Metathesaurus change files will provide for relationships and attributes in the RRF only. The generation of these files is dependent on the relationship and attribute identifiers (RUI and ATUI) introduced in the 2004AA version of the Metathesaurus.

2.6.3. Historical CUIs

The retired CUI file (MRCUI.RRF in RRF and MRCUI in ORF) includes all CUIs present in any previous version of the Metathesaurus, but not in the current version. In general, the file maps the retired CUI to one or more current CUIs.

2.7. Concept Name Indexes

To assist system developers in building applications that retrieve all strings or concept names which include specific words or groups of words, three indexes to the concept names are provided: a Word Index, a Normalized Word Index (for English words only), and a Normalized String Index (for English strings only). The indexes are described in Sections 2.7.1, 2.7.2, and 2.7.3, respectively. To make the distinctions among them clearer, the examples include words or strings that would appear in each index for the following set of Metathesaurus concept names:

Lung Diseases, Obstructive	(C0600260, L0024117, S0058463)

Table continued from previous page.

Obstructive Lung Diseases	(C0600260, L0024117, S0068169)
Lung Disease, Obstructive	(C0600260, L0024117, S0058458)
Obstructive Lung Disease	(C0600260, L0024117, S0068168)

2.7.1. Word Index

2.7.1.1. Description

The word index connects each individual word in any Metathesaurus string to all its related string, term, and concept identifiers. There are separate word index files for each language in the Metathesaurus.

There is one entry for each word found in each unique string in each language. Each entry has five sub-elements.

1. LAT - 3-letter abbreviation for language
2. WD - Word
3. CUI - concept unique identifier
4. LUI - term unique identifier
5. SUI - string unique identifier

2.7.1.2. Definition of a Word

In this index, a word is defined as a token containing only alphanumeric characters with length one or greater; for more information, see the SPECIALIST Lexicon and Lexical Tools.

2.7.1.3. Word Index Example

For the four example concept names listed above, the word index will contain multiple entries for each of the following words: disease, diseases, lung, obstructive. Two of the entries generated for the names Lung Disease, Obstructive and Obstructive Lung Disease are shown below:

ENG|disease|C0600260|L0024117|S0058458|
ENG|disease|C0600260|L0024117|S0068168|

2.7.2. Normalized Word Index

2.7.2.1. Description

The normalized word index connects each individual normalized English word to all its related string, term, and concept identifiers.

There is one entry for each normalized word found in each unique English string. There are no entries for other languages in this index. Each entry has five sub-elements.

1. LAT - (always ENG in this edition of the Metathesaurus)
2. NWD - normalized word
3. CUI - concept unique identifier
4. LUI - term unique identifier
5. SUI - string unique identifier

2.7.2.2. Definition of Normalized Word

The normalization process involves breaking a string into its constituent words, lowercasing each word, and converting it to its uninflected form. Normalized words are generated by uninflecting each word and stripping

out a small number of stop words. The uninflected forms are generated using the SPECIALIST Lexicon if the words appear in the lexicon; otherwise they are generated algorithmically.

2.7.2.3. Normalized Word Example

For the four example concept names listed above, the normalized word index will contain multiple entries for each of the following words: disease, lung, obstructive. Since the normalized word index contains base forms only, it does not contain entries for the plural "diseases". In this index, therefore, all four concept names are linked to the normalized word "disease", as follows:

ENG|disease|C0600260|L0024117|S0058458|
ENG|disease|C0600260|L0024117|S0058463|
ENG|disease|C0600260|L0024117|S0068168|
ENG|disease|C0600260|L0024117|S0068169|

2.7.3. Normalized String Index

2.7.3.1. Description

The normalized string index connects the normalized form of a Metathesaurus string to all its related string, term, and concept identifiers. There is one entry for each unique (non-normalized) English string. There are no entries for other languages in this index. Each entry has five sub-elements.

1. LAT - (always ENG in this edition of the Metathesaurus)
2. NSTR - normalized string
3. CUI - concept unique identifier
4. LUI - term unique identifier
5. SUI - string unique identifier

2.7.3.2. Definition of Normalized String

The normalization process involves breaking a string into its constituent words, lowercasing each word, converting each word to its uninflected form, and sorting the words in alphabetic order. Normalized strings are generated by uninflecting each word, leaving out a small number of stop words. The uninflected forms are generated using the SPECIALIST Lexicon if the words appear in the lexicon; otherwise they are generated algorithmically.

2.7.3.3. Normalized String Example

Since the four example concept names listed above are composed of the same set of normalized words, the Normalized String Index will contain four entries for a single string: disease lung obstructive, in which the component normalized words appear in alphabetical order. The **complete** set of Normalized String Index entries generated by the four concept names is as follows:

ENG|disease lung obstructive|C0600260|L0024117|S0058458|
ENG|disease lung obstructive|C0600260|L0024117|S0058463|
ENG|disease lung obstructive|C0600260|L0024117|S0068168|
ENG|disease lung obstructive|C0600260|L0024117|S0068169|

2.7.4. Word Index Programs

The programs that generate these indexes are written in Java. They may be of use to system developers who are developing their own interfaces to the UMLS data or for other purposes. Chapter 6 includes information about these and other lexical programs provided with the UMLS Knowledge Sources.

2.8. Character Sets

The UMLS Knowledge Sources are distributed in Unicode (specifically, in the UTF-8 encoding of the Unicode 4.0 standard [1]) to avoid complexity and information loss.

Unicode is a single unified and interoperable global standard, which includes the characters needed to write in any language (see www.unicode.org). Unicode also includes diacritical marks, ideographs, and scientific and other symbols. Most modern systems already use Unicode; we strongly encourage you to upgrade to Unicode compliant systems and software.

The 7-bit basic ASCII character set is the 'least common denominator' character set of 96 characters and symbols from the oldest ASCII standard. UTF-8 is identical to the ASCII encoding for characters in the 7-bit ASCII range, so that 7-bit ASCII files are automatically a correct subset of UTF-8. This means that sources originally in 7-bit ASCII are unchanged. In the UMLS, the term 'extended characters' refers to all Unicode characters beyond this 7-bit ASCII subset. All other character sets are converted to, and distributed in, UTF-8.

Note that the UMLS LAT - Language of Term(s) - is the language the source declares. Since the world does not speak or write in 7-bit ASCII, sources often include extended characters for symbols or from other languages, for example in eponyms.

The MetamorphoSys default is to output all records and data in standard UTF-8. Checking the option to "Remove records containing extended UTF-8 characters" will exclude from your subset all terms and other data that contain extended characters. This will create gaps in the hierarchy and may cause loss of vocabulary which matters to your application.

For most English or Spanish sources, i.e., LAT = ENG or SPA, an equivalent 7-bit ASCII string is created for the UMLS to help users of older systems. If you wish to use them, these forms must not be excluded from your subset. These forms are created by the lvg program (see the Lexical Variant Generation section in Section 6.8). This program may be of interest to those who wish to do further conversions; it converts extended characters to an escaped form of the official Unicode character name to ensure that no information is lost. These names may not be "reader friendly" but are useful for some purposes such as indexing.

The initial byte order mark (BOM) character is not present in the UTF-8 encoded Metathesaurus files unless the option "Add UTF-8 BOM characters to output files" is selected on the Output options tab in MetamorphoSys.

Files will be in byte sort order (for example, with data in UTF-8, standard UNIX sort works as expected). Note that the UMLS data are intended to be manipulated with software tools such as database systems, so the sort order of the files should not matter.

2.9. Content Views

A Content View (CV) is any definable subset of the Metathesaurus that is useful for some specific purpose. Content Views are either created by NLM or submitted to the Metathesaurus by external authorities. Membership may be defined in a variety of ways, including:

- A list of Metathesaurus UIs (CUIs, SUIs, AUIs, etc.), maintained over time.
- A list of sources that participate in the view.
- A complex query or algorithm that computes sets of atoms, source concepts, or relationships based on well-defined criteria and may also be configured to include other connected information, such as attributes or relationships.

See the Content Views page for a list of Content Views in the current release.

2.9.1. Representation of Content View Metadata in RRF

Higher level information about each Content View is included as a concept directly in the Rich Release Format (RRF) files as described below.

2.9.1.1. MRCONSO.RRF

Each Content View is represented by a concept in MRCONSO which has an atom with SAB=MTH and TTY=CV:

```
STR = <Content View Name>, e.g. 'MetaMap NLP View'
CODE = NOCODE
TTY = CV
SAB = MTH
SAUI, SCUI, SDUI = null
```

2.9.1.2. MRSTY.RRF

Each Content View concept is assigned "Intellectual Product" as the Semantic Type (STY).

2.9.1.3. MRSAT.RRF

Each Content View has required metadata attributes which appear in MRSAT:

ATN	ATV
CV_ALGORITHM	Content View algorithm
CV_CATEGORY	Content View category
CV_CLASS	Content View class
CV_CODE	Content View code
CV_CONTRIBUTOR_DATE	Date corresponding to the contributor version of this Content View
CV_CONTRIBUTOR_URL	URL corresponding to the contributor version of this Content View
CV_CONTRIBUTOR_VERSION	Version of this Content View submitted by the contributor
CV_CONTRIBUTOR	Content View contributor
CV_DESCRIPTION	Content View description
CV_IS_GENERATED	Content View generated: Y/N
CV_MAINTAINER_DATE	Date corresponding to the maintainer version of this Content View
CV_MAINTAINER_URL	URL corresponding to the maintainer version of this Content View
CV_MAINTAINER_VERSION	Version of this Content View submitted by the maintainer
CV_MAINTAINER	Content View maintainer
CV_PREVIOUS_META	Previous UMLS Metathesaurus version used to generate Content View. A null value means the Content View is generated based on current UMLS Metathesaurus version.
CV_SUBCATEGORY	Content View subcategory

2.9.2. Extracting Content Views

2.9.2.1. Using MetamorphoSys

Content Views are designed to be extracted using MetamorphoSys.

To extract a Content View:

1. Open the File menu on the UMLS MetamorphoSys Configuration screen.
2. Select "Enable/Disable Filter" --> "Content View Filter."
3. Click "OK."
4. Select the desired Content View(s) in the resulting configuration panel.

The Content View Flag (CVF) in a resulting RRF subset is set to an integer representing the sum of the CV_CODE values of the selected view(s) that apply to each data element. See the Content Views page for CV_CODE values. Note: CVFs are not represented in Original Release Format (ORF).

For example, if a subset is created for the "MetaMap NLP View," the CVF is set to 256 (matching the CV_CODE attribute in the "MetaMap NLP View" metadata concept). If a subset is created containing two Content Views, the CVF for atoms participating in both Content Views is the sum of the CV_CODE values that apply in each case. For example, a subset containing atoms that participate in both "MetaMap NLP View" and the "CORE Problem List Subset of SNOMED CT" would have a CVF of 2304 (256 + 2048), the sum of the respective CV_CODE values for those two Content Views. Note: Some atoms in that subset would belong only to "MetaMap NLP View" and would thus still have a CVF value of 256, and others would belong only to the "CORE Problem List Subset of SNOMED CT" and would have a CVF value of 2048.

2.9.2.2. Directly from RRF files

Content View processing outside of MetamorphoSys requires bit field programming. If you have already created a UMLS subset, most RRF files contain a Content View Flag field to denote Content View membership. A CVF consists of an integer representing a bit mask. When interpreted as a binary number, each bit of the integer represents a particular Content View – up to a total of 64 views. These bits are assigned from least significant to most significant digit. Membership in a particular Content View is indicated by the presence of a "1" in the corresponding bit. A "0" indicates that it is not a member. The bit-string is converted into a decimal number for display. Thus if the 9th bit (256) and the 12th bit (2048) were each set to 1, the resulting value would be 2304 (or 100100000000 in binary). The bit used by a corresponding Content View is defined by the CV_CODE attribute in that view's metadata.

Consider the case of trying to find all rows in a MRCONSO table loaded from the MRCONSO.RRF file belonging to the "MetaMap NLP View." Start by identifying all Content Views in the current subset:

```
SELECT * FROM MRCONSO WHERE TTY='CV';
```

Among other results, this query would yield the CUI of the desired Content View: C1700357. Next, query into MRSAT to reveal the Content View metadata, including the CV_CODE value:

```
SELECT ATN, ATV FROM MRSAT WHERE CUI='C1700357';
```

With the CV_CODE known (256) the final step is to identify the entries in MRCONSO participating in this Content View:

```
SELECT * FROM MRCONSO WHERE BITAND(CVF, 256) <> 0;
```

The CV_CODE for this Content View is 256. Any entry in MRCONSO participating in that Content View would yield a non-zero value when a BITAND operation is applied to the CVF using the CV_CODE (256). All entries not participating in the Content View would yield a zero value for this operation.

2.10. Mappings

Inter-source mappings in the Metathesaurus provide links from entities in one terminology (the source terminology) to entities in another terminology (the target terminology). Entities may be terms, codes, concepts, descriptors, or expressions. Mappings may be used for a variety of purposes, including:

- reuse of data for another purpose (e.g. translating clinical information coded with SNOMED CT to ICD-9-CM for reimbursement purposes)
- retaining the value of data when migrating to newer terminology requirements (e.g. updating from ICD-9-CM to ICD-10-CM)

Given the diversity of mapping applications, it is important to understand the purpose, approach, and authority and validation of a mapping when evaluating it for a particular use case.

Inter-source mapping data are represented in MRMAP.RRF and MRSMAP.RRF, with auxiliary data in MRCONSO.RRF, MRSTY.RRF, and MRSAT.RRF. Mapping data may also be redundantly represented as relationships in MRREL.RRF.

2.10.1. Representation of Mappings in the Metathesaurus

Inter-source mapping data is represented using the following specifications (there may be exceptions, e.g. for map sets that have not been updated recently):

2.10.1.1. MRCONSO.RRF

For each map set represented in MRMAP.RRF, there is a single "Cross mapping set" concept in MRCONSO.RRF. Note that the CUI changes when the map set is updated from one version to the next.

Field values are assigned as follows:

- SAB: the source that asserts the mapping information. For example, LCH_NW provides one mapping: LCH_NW_2013 to MSH2015_2014_09_08 Mappings

The SAB for these map set atoms is "LCH_NW."

- TTY: XM for all map set atoms
- STR: The atom name is created as "<VSAB> to <VSAB> Mappings <optional additional information>" For example:
  - SNOMEDCT_2011_07_31 to ICD9CM_2011 Mappings
- CODE: If an appropriate identifier for the map set is available from the source, it will be used as the CODE. SAUI, SCUI and SDUI may also be populated. If no source-asserted identifier is available, a CODE beginning with "MTHU" will be generated during Metathesaurus production.

Example:

C3826804|ENG|P|L11643734|PF|S14441772|Y|A23864609|LCH_NW|XM|MTHU000001|LCH_NW_2013 to MSH2015_2014_09_08|0|N|256||

2.10.1.2. MRSTY.RRF

All map set concepts are assigned an STY of "Intellectual Product".

Example:

C3826807|T170|A2.4|Intellectual Product|AT201718383|256||

2.10.1.3. MRSAT.RRF

Every map set concept has numerous attributes in MRSAT.RRF which provide additional details. The following attributes can be found in MRSAT.RRF. The attributes are attached using STYPE=CODE.

Required Attributes:

ATN	ATV	Valid Values
FROMRSAB	Root source abbreviation for the "from" identifiers of a map set	range=MRSAB.RSAB
FROMVSAB	Versioned source abbreviation for the "from" identifiers of a map set	range=MRSAB.VSAB
MAPSETRSAB	Root source abbreviation for a map set - in general, the same as the value for FROMRSAB	range=MRSAB.RSAB
MAPSETVERSION	Version of the map set	N/A
MAPSETVSAB	Versioned source abbreviation for the provider of a map set	range=MRSAB.VSAB
TORSAB	Root source abbreviation for the "to" identifiers of a map set	range=MRSAB.RSAB
TOVSAB	Versioned source abbreviation for the "to" identifiers of a map set	range=MRSAB.VSAB

Optional Attributes: In general, these attributes are extracted directly from source-provided data and may have a diverse range of values and formats.

ATN	ATV
MAPSETGRAMMAR	Grammar used by expressions in FROMEXPR or TOEXPR fields
MAPSETNAME	Official name of a map set
MAPSETREALMID	Identifier of a "realm" to which a source is mapped, within which this cross mapping table is applicable. Used in cases where Realm specific business rules or guidelines alter the acceptable mappings. Realm is the same as used in SNOMED CT subsets. It includes a four character ISO6523 identifier followed by an optional series of concatenated subdivision codes defined by the registered organization.
MAPSETRULETYPE	Indicates the types of rules used in a map set and cross map targets to which a source is mapped.
MAPSETSCHEMEID	Standard identifier for the scheme to which a map set belongs. This may be an International Coding Scheme Identifier (ISO7826) or an Object Identifier (OID) used as specified by HL7.
MAPSETSCHEMENAME	Full name of the target scheme in a map set.
MAPSETSCHEMEVERSION	Version number of the target scheme (as published by the issuing organization) in a map set.
MAPSETSEPARATORCODE	XML entity code (for example, "|" to represent the vertical-bar character) for the character used as a separator between the individual codes in the target codes field in a map set.
MAPSETSID	Source asserted identifier for a map set. If present, matches the CODE in MRCONSO.RRF.
MAPSETTYPE	Indicates the nature of a map set. Its value is map set specific. It can be used to indicate the inclusion of one to one, one to many, or rule based.
MAPSETXRTARGETID	Map set target identifier used for XR mappings. Only used for map sets that explicitly map source codes to “nothing.”
SOS	Scope statement
TARGETSCHEMEID	Identifier for the target scheme in the map set. This may be an International Coding Scheme Identifier (ISO7826) or an Object Identifier (OID) used as specified by HL7.

Optional MTH Attributes: ATNs for attributes created during Metathesaurus source processing begin with “MTH_”.

ATN	ATV	Valid Values
MTH_MAPFROMCOMPLEXITY	Two-part value indicating the complexity of "from" expressions used in a map set. Valid values can be combined in a comma-separated list	Part 1: SINGLE, LIST, or BOOLEAN_EXPRESSION Part 2: AUI, CODE, CUI, LUI, SAUI, SCUI, SDUI, SUI, or STR

Table continued from previous page.

ATN	ATV	Valid Values
MTH_MAPFROMEXHAUSTIVE	Indicates whether or not the "from" source of a map set is completely mapped	Y/N
MTH_MAPSETCOMPLEXITY	Indicates the overall complexity of a map set. To compute this field: 1. Compute FROMEXPR cardinality (left hand side) based on whether >1 FROMEXPR exists for same TOEXPR OR MTH_MAPTOCOMPLEXITY indicates MULTIPLE. 2. Compute TOEXPR cardinality (right hand side) based on whether >1 TOEXPR exists for same FROMEXPR OR MTH_MAPFROMCOMPLEXITY indicates MULTIPLE. 3. RULE_BASED if >1 non-null distinct MAPSUBSETID	N_TO_N, N_TO_ONE, ONE_TO_N, ONE_TO_ONE, or RULE_BASED
MTH_MAPTOCOMPLEXITY	Two-part value indicating the complexity of "to" expressions used in a map set. Valid values can be combined in a comma-separated list	Part 1: SINGLE, LIST, or BOOLEAN_EXPRESSION Part 2: AUI, CODE, CUI, LUI, SAUI, SCUI, SDUI, SUI, or STR
MTH_MAPTOEXHAUSTIVE	Indicates whether or not the "to" source is completely mapped	Y/N
MTH_UMLSMAPSETSEPARATOR	The character used in the UMLS Metathesaurus as a separator between the individual codes in the target codes field of the cross map targets to which a source is mapped.	AND

Examples:

C3826807|L11643734|S14441772|A23864609|CODE|MTHU000001|AT197916839||MAPSETRSAB|LCH_NW|LCH_NW|N||

C3826807|L11643734|S14441772|A23864609|CODE|MTHU000001|AT197916840||FROMVSAB|LCH_NW|LCH_NW_2013|N||

C3826807|L11643734|S14441772|A23864609|CODE|MTHU000001|AT197916842||TORSAB|LCH_NW|MSH|N||

2.10.1.4. MRMAP.RRF

MRMAP.RRF contains information on entities that are mapped to each other and on the source responsible for the mapping. See Section 3.3.13 for more information on this file.

2.10.1.5. MRSMAP.RRF

This file provides a simpler representation of most of the mappings in MRMAP.RRF to serve applications which do not require the full richness of the MRMAP.RRF data structure. See Section 3.3.14 for more information on this file.

2.10.1.6. MRREL.RRF

A subset of mappings is redundantly represented as relationships in MRREL.RRF, based on the following guidelines:

- FROMEXPR and TOEXPR are simple expressions

- Map set is not rule-based
- REL is not XR
- Partial map sets may be represented in MRREL

There is currently no simple way of identifying cross-source mappings in MRREL.RRF. The RELAs for these relationships currently include “mapped_to/from,” “same_as,” “classified_as/classified_by” and the null RELA. All of these RELAs are also used for within-source RELAs. To identify cross-source mapping relationships, find MRREL.RRF cases where the AUI1 and AUI2 in MRCONSO.RRF have different SAB values and neither STYPE1 nor STYPE2 is CUI.

3. Metathesaurus - Rich Release Format (RRF)

Metathesaurus users may select from two relational formats: the Rich Release Format (RRF), first introduced in 2004, and the Original Release Format (ORF). Both are available as output options of MetamorphoSys, the installation and customization program.

Developers are encouraged to use the RRF, which offers significant advantages in source vocabulary transparency (that is, ability to exactly represent the detailed semantics of each source vocabulary); in the ability to generate complete and accurate change sets between versions of the Metathesaurus; and in more convenient representations of concept name, source, and hierarchical context information.

Neither Metathesaurus format is fully normalized. By design, there is duplication of data among different files and within certain files. In particular, relationships between different Metathesaurus concepts appear twice (e.g., from entry A to entry B and from entry B to entry A). Developers will need to make their own decisions about the extent to which this redundancy should be retained, reduced, or increased for their specfic applications.

All files except MRRANK.RRF are sorted by row.

3.1. Data Files

The data in each Metathesaurus entry may be represented in more than 20 different relations, or files. These files correspond to the four logical groups of data elements described in Sections 2.3 - 2.6 and the indexes described in Section 2.7 as follows:

- Concepts, Concept Names, and their sources (2.3) = MRCONSO.RRF
- Attributes (2.5) = MRSAT.RRF, MRDEF.RRF, MRSTY.RRF, MRHIST.RRF
- Relationships (2.4) = MRREL.RRF, MRCXT.RRF, MRHIER.RRF, MRMAP.RRF, MRSMAP.RRF
- Data about the Metathesaurus (2.6) = MRFILES.RRF, MRCOLS.RRF, MRDOC.RRF, MRRANK.RRF, MRSAB.RRF, AMBIGLUI.RRF, AMBIGSUI.RRF, CHANGE/MERGEDCUI.RRF, CHANGE/MERGEDLUI.RRF, CHANGE/DELETEDCUI.RRF, CHANGE/DELETEDLUI.RRF, CHANGE/DELETEDSUI.RRF, MRCUI.RRF
- Indexes (2.7) = MRXW_BAQ.RRF, MRXW_DAN.RRF, MRXW_DUT.RRF, MRXW_ENG.RRF, MRXW_FIN.RRF, MRXW_FRE.RRF, MRXW_GER.RRF, MRXW_HEB.RRF, MRXW_HUN.RRF, MRXW_ITA.RRF, MRXW_NOR.RRF, MRXW_POR.RRF, MRXW_RUS.RRF, MRXW_SPA.RRF, MRXW_SWE.RRF, MRXNW_ENG.RRF, MRXNS_ENG.RRF

3.2. Columns and Rows

Each file or named table of data values has by definition a fixed number of columns; the number of rows depends on the content of a particular version of the Metathesaurus.

A column is a sequence of all the values in a given data element or logical sub-element. In general, columns for longer variable length data elements will appear to the right of columns for shorter and/or fixed length data elements. The information for all columns in the files is described in MRCOLS.RRF and on the Columns and Data Elements page of the current release documentation.

A row contains the values for one or more data elements or logical sub-elements for one Metathesaurus entry. Depending on the nature of the data elements involved, each Metathesaurus entry may have one or more rows in a given file. The values for the different data elements or logical sub-elements represented in the row are separated by vertical bars (|). If an optional element is blank, the vertical bars are still used to maintain the correct positioning of the subsequent elements. Each row is terminated by a vertical bar and line termination.

3.3. Descriptions of Each File

The descriptions of the files appear in the following order:

1. Key data about the Metathesaurus: Files; Columns or Data Elements; documentation that explains the meaning of abbreviations that appear as values in Metathesaurus data elements and attributes
2. Concept names and their vocabulary sources
3. Attributes
4. Relationships
5. Other data about the Metathesaurus
6. Indexes

Each file description lists the columns or data elements that appear in the file and includes sample rows from the file.

3.3.1. Files (File = MRFILES.RRF)

There is exactly one row in this file for each physical segment of each logical file. Data elements that appear in multiple files, e.g., CUI, AUI, will have multiple rows in this file.

Col.	Description
FIL	Physical FILENAME
DES	Descriptive Name
FMT	Comma separated list of column names (COL), in order
CLS	# of COLUMNS
RWS	# of ROWS
BTS	Size in bytes in this format (ISO/PC or Unix)

Sample Records

MRSTY.RRF|Semantic Types|CUI,TUI,STN,STY,ATUI,CVF|6|2630816|149735178|

3.3.2. Data Elements (File = MRCOLS.RRF)

There is exactly one row in this file for each column or data element in each file. Data elements that appear in multiple files, e.g., CUI, AUI, will have multiple rows in this file.

Col.	Description
COL	Column or data element name
DES	Descriptive Name
REF	Documentation Section Number
MIN	Minimum Length, Characters
AV	Average Length
MAX	Maximum Length, Characters
FIL	Physical FILENAME in which this field occurs
DTY	SQL-92 data type for this column

Sample Records

AUI|Unique identifier for atom||8|8.57|9|MRCONSO.RRF|varchar(9)|

CODE|Unique Identifier or code for string in source||1|7.23|30|MRCONSO.RRF|varchar(50)|

3.3.3. Documentation for Abbreviated Values (File = MRDOC.RRF)

There is exactly one row in this table for each allowed value of selected data elements or attributes that have a finite number of abbreviations as allowed values. Examples of such data elements include TTY, ATN, TS, STT, REL, RELA.

Col.	Description
DOCKEY	Data element or attribute
VALUE	Abbreviation that is one of its values
TYPE	Type of information in EXPL column
EXPL	Explanation of VALUE

Sample Records

ATN|DDF|expanded_form|Drug Doseform (e.g. chewable tablet)|

ATN|FDA_UNII_CODE|expanded_form|FDA UNII Code|

*Note: The MRDOC file produced by MetamorphoSys contains metadata about the release itself. Here is an example of the records:

RELEASE|mmsys.build.date|release_info|2010_10_19_11_52_39|

RELEASE|mmsys.version|release_info|MMSYS-2010AB-20101019|

3.3.4. Concept Names and Sources (File = MRCONSO.RRF)

There is exactly one row in this file for each atom (each occurrence of each unique string or concept name within each source vocabulary) in the Metathesaurus, i.e., there is exactly one row for each unique AUI in the Metathesaurus. Every string or concept name in the Metathesaurus appears in this file, connected to its language, source vocabularies, and its concept identifier. The values of TS, STT, and ISPREF reflect the default order of precedence of vocabulary sources and term types in MRRANK.RRF. (Table 1)

Sample Records

C0001175|ENG|P|L0001175|VO|S0010340|Y|A0019182||M0000245|D000163|MSH|PM|D000163|Acquired Immunodeficiency Syndromes|0|N||

C0001175|ENG|S|L0001842|PF|S0011877|N|A2878223|103840012|62479008||SNOMEDCT_US|PT|62479008|AIDS|9|N|2304|

C0001175|ENG|P|L0001175|VO|S0354232|Y|A2922342|103845019|62479008||SNOMEDCT_US|SY|62479008|Acquired immunodeficiency syndrome|9|N|2304|

C0001175|FRE|S|L0162173|PF|S0226654|Y|A27478989||M0000245|D000163|MSHFRE|ET|D000163|SIDA|3|N||

C0001175|RUS|S|L0904943|PF|S1108760|Y|A13488500||M0000245|D000163|MSHRUS|SY|D000163|SPID|3|N||

Table 1. Concept Names and Sources (File = MRCONSO.RRF)

Col.	Description
CUI	Unique identifier for concept
LAT	Language of term
TS	Term status
LUI	Unique identifier for term
STT	String type
SUI	Unique identifier for string
ISPREF	Atom status - preferred (Y) or not (N) for this string within this concept
AUI	Unique identifier for atom - variable length field, 8 or 9 characters
SAUI	Source asserted atom identifier [optional]
SCUI	Source asserted concept identifier [optional]
SDUI	Source asserted descriptor identifier [optional]
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
TTY	Abbreviation for term type in source vocabulary, for example PN (Metathesaurus Preferred Name) or CD (Clinical Drug). Possible values are listed on the Abbreviations Used in Data Elements page.
CODE	Most useful source asserted identifier (if the source vocabulary has more than one identifier), or a Metathesaurus-generated source entry identifier (if the source vocabulary has none)
STR	String
SRL	Source restriction level
SUPPRESS	Suppressible flag. Values = O, E, Y, or N O: All obsolete content, whether they are obsolesced by the source or by NLM. These will include all atoms having obsolete TTYs, and other atoms becoming obsolete that have not acquired an obsolete TTY (e.g. RxNorm SCDs no longer associated with current drugs, LNC atoms derived from obsolete LNC concepts). E: Non-obsolete content marked suppressible by an editor. These do not have a suppressible SAB/TTY combination. Y: Non-obsolete content deemed suppressible during inversion. These can be determined by a specific SAB/TTY combination explicitly listed in MRRANK. N: None of the above Default suppressibility as determined by NLM (i.e., no changes at the Suppressibility tab in MetamorphoSys) should be used by most users, but may not be suitable in some specialized applications. See the MetamorphoSys Help page for information on how to change the SAB/TTY suppressibility to suit your requirements. NLM strongly recommends that users not alter editor-assigned suppressibility, and MetamorphoSys cannot be used for this purpose.
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.5. Simple Concept and Atom Attributes (File = MRSAT.RRF)

There is exactly one row in this table for each concept, atom, or relationship attribute that does not have a sub-element structure. All Metathesaurus concepts and a minority of Metathesaurus relationships have entries in this file. This file includes all source vocabulary attributes that do not fit into other categories. (Table 2)

Sample Records

C0001175|L0001175|S0010339|A0019180|SDUI|D000163|AT38209082||FX|MSH|D015492|N||

C0001175|||R54775538|RUI||AT173814751||CHARACTERISTIC_TYPE_ID|SNOMEDCT_US| 900000000000011006|O||

C0001175|||R54775538|RUI||AT174785253||MODIFIER_ID|SNOMEDCT_US| 900000000000451002|O||

Table 2. Simple Concept and Atom Attributes (File = MRSAT.RRF)

Col.	Description
CUI	Unique identifier for concept (if METAUI is a relationship identifier, this will be CUI1 for that relationship)
LUI	Unique identifier for term (optional - present for atom attributes, but not for relationship attributes)
SUI	Unique identifier for string (optional - present for atom attributes, but not for relationship attributes)
METAUI	Metathesaurus atom identifier (will have a leading A) or Metathesaurus relationship identifier (will have a leading R) or blank if it is a concept attribute.
STYPE	The name of the column in MRCONSO.RRF or MRREL.RRF that contains the identifier to which the attribute is attached, i.e. AUI, CODE, CUI, RUI, SCUI, SDUI.
CODE	Most useful source asserted identifier (if the source vocabulary contains more than one) or a Metathesaurus-generated source entry identifier (if the source vocabulary has none). Optional - present if METAUI is an AUI.
ATUI	Unique identifier for attribute
SATUI	Source asserted attribute identifier (optional - present if it exists)
ATN	Attribute name. Possible values appear in MRDOC.RRF and are described on the Attribute Names page.
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
ATV	Attribute value described under specific attribute name on the Attributes Names page. A few attribute values exceed 1,000 characters. Many of the abbreviations used in attribute values are explained in MRDOC.RRF and included on the Abbreviations Used in Data Elements page.
SUPPRESS	Suppressible flag. Values = O, E, Y, or N. Reflects the suppressible status of the attribute. See also SUPPRESS in MRCONSO.RRF, MRDEF.RRF, and MRREL.RRF.
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.6. Definitions (File = MRDEF.RRF)

There is exactly one row in this file for each definition in the Metathesaurus. A definition is an attribute of an atom (an occurrence of a string in a source vocabulary). A few approach 3,000 characters in length. (Table 3)

Sample Records

C0001175|A0019180|AT38139119||MSH|An acquired defect of cellular immunity associated with infection by the human immunodeficiency virus (HIV), a CD4-positive T-lymphocyte count under 200 cells/microliter or less than 14% of total lymphocytes, and increased susceptibility to opportunistic infections and malignant

neoplasms. Clinical manifestations also include emaciation (wasting) and dementia. These elements reflect criteria for AIDS as defined by the CDC in 1993.|N||

C0001175|A0021048|AT51221477||CSP|one or more indicator diseases, depending on laboratory evidence of HIV infection (CDC); late phase of HIV infection characterized by marked suppression of immune function resulting in opportunistic infections, neoplasms, and other systemic symptoms (NIAID).|N||

C0001175|A7568512|AT198127773||NCI_NCI-GLOSS|A disease caused by human immunodeficiency virus (HIV). People with acquired immunodeficiency syndrome are at an increased risk for developing certain cancers and for infections that usually occur only in individuals with a weak immune system.|N||

Table 3. Definitions (File = MRDEF.RRF)

Col.	Description
CUI	Unique identifier for concept
AUI	Unique identifier for atom - variable length field, 8 or 9 characters
ATUI	Unique identifier for attribute
SATUI	Source asserted attribute identifier [optional-present if it exists]
SAB	Abbreviated source name (SAB) of the source of the definition Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
DEF	Definition
SUPPRESS	Suppressible flag. Values = O, E, Y, or N. Reflects the suppressible status of the attribute; not yet in use. See also SUPPRESS in MRCONSO.RRF, MRREL.RRF, and MRSAT.RRF.
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.7. Semantic Types (File = MRSTY.RRF)

There is exactly one row in this file for each Semantic Type assigned to each concept. All Metathesaurus concepts have at least one entry in this file. Many have more than one entry. The TUI, STN, and STY are all direct links to the UMLS Semantic Network.

Col.	Description
CUI	Unique identifier of concept
TUI	Unique identifier of Semantic Type
STN	Semantic Type tree number
STY	Semantic Type. The valid values are defined in the Semantic Network.
ATUI	Unique identifier for attribute
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

Sample Record

C0001175|T047|B2.2.1.2.1|Disease or Syndrome|AT17683839|2304|

3.3.8. History (File = MRHIST.RRF)

This file tracks source-asserted history information. It currently includes SNOMED CT history only. (Table 4)

Sample Records

C0000294|108821000|SNOMEDCT|20001101|0|CONCEPTSTATUS|0|||

C0000294|108821000|SNOMEDCT|20020731|2|CONCEPTSTATUS|0|FULLYSPECIFIEDNAME CHANGE||

C0000294|1185494016|SNOMEDCT|20020731|0|DESCRIPTIONSTATUS|0|||

C0000294|1185494016|SNOMEDCT|20100731|2|DESCRIPTIONSTATUS|0|INITIALCAPITALSTATUS CHANGE||

C0000294|1461100014|SNOMEDCT|20030131|0|DESCRIPTIONSTATUS|0|||

Table 4. History (File = MRHIST.RRF)

Col.	Description
CUI	Unique identifier for concept
SOURCEUI	Source asserted unique identifier
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
SVER	Release date or version number of a source
CHANGETYPE	Source asserted code for type of change
CHANGEKEY	CONCEPTSTATUS (if history relates to a SNOMED CT concept) or DESCRIPTIONSTATUS (if history relates to a SNOMED CT atom)
CHANGEVAL	CONCEPTSTATUS value or DESCRIPTIONSTATUS value after the change took place. Note: The change may have affected something other than the status value.
REASON	Explanation of change if present
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.9. Related Concepts (File = MRREL.RRF)

There is one row in this table (Table 5) for each relationship between concepts or atoms known to the Metathesaurus, with the following exceptions found in other files: pair-wise mapping relationships between two source vocabularies found in MRMAP.RRF and MRSMAP.RRF.

Note that for asymmetrical relationships there is one row for each direction of the relationship. Note also the direction of REL - the relationship which the SECOND concept or atom (with Concept Unique Identifier CUI2 and Atom Unique Identifier AUI2) HAS TO the FIRST concept or atom (with Concept Unique Identifier CUI1 and Atom Unique Identifier AUI1).

Sample Records

C0002372|A0022283|AUI|SY|C0002372|A16796726|AUI||R55153988||RXNORM|RXNORM|||N||

C0002372|A0022283|AUI|RO|C2241537|A14211642|AUI|has_ingredient|R91984327||MMSL|MMSL|||N||

Table 5. Related Concepts (File = MRREL.RRF)

Col.	Description
CUI1	Unique identifier of first concept
AUI1	Unique identifier of first atom
STYPE1	The name of the column in MRCONSO.RRF that contains the identifier used for the first element in the relationship, i.e. AUI, CODE, CUI, SCUI, SDUI.
REL	Relationship of second concept or atom to first concept or atom
CUI2	Unique identifier of second concept
AUI2	Unique identifier of second atom
STYPE2	The name of the column in MRCONSO.RRF that contains the identifier used for the second element in the relationship, i.e. AUI, CODE, CUI, SCUI, SDUI.
RELA	Additional (more specific) relationship label (optional)
RUI	Unique identifier of relationship
SRUI	Source asserted relationship identifier, if present
SAB	Abbreviated source name of the source of relationship. Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
SL	Source of relationship labels
RG	Relationship group. Used to indicate that a set of relationships should be looked at in conjunction.
DIR	Source asserted directionality flag. Y indicates that this is the direction of the relationship in its source; N indicates that it is not; a blank indicates that it is not important or has not yet been determined.
SUPPRESS	Suppressible flag. Reflects the suppressible status of the relationship. See also SUPPRESS in MRCONSO.RRF, MRDEF.RRF, and MRSAT.RRF.
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.10. Co-occurring Concepts (File = MRCOC.RRF - This file is no longer available in the UMLS after the 2013AA release.)

Note: Co-occurrence information is no longer available in the UMLS after the 2013AA release. Updated co-occurrences data are available in text files from the MEDLINE Co-Occurrences (MRCOC) page.

This file includes statistical aggregations of co-occurrences of meanings in external data sources. These exist at the AUI level. There are two rows in this table for each pair of atoms that co-occur in each information source represented: one for each direction of the relationship. (Note that the COA data may be different for each direction of the relationship.) Many Metathesaurus concepts have no entries in this file. Due to the very large number of co-occurrence relationships, they are distributed in a separate file. (Table 6)

Co-occurrences are concepts that occur together in the same entries in some information source. The relationships represented here are obtained from machine-manipulation of the information source. Co-occurrence relationships may exist between similar concepts (e.g., Atrial Fibrillation and Arrhythmia) or

between very different concepts that nevertheless have some important connection in the field of biomedicine (e.g., Atrial Fibrillation and Digoxin), or between a primary concept and a qualifier (e.g., Lithotripsy and instrumentation). A co-occurrence relationship can exist between two concepts that have no other apparent relationship, although the frequency of such co-occurrences will be small.

In the current Metathesaurus, there are three sources of co-occurrence data: MEDLINE, AI/RHEUM, and CCPSS. From MEDLINE, co-occurrence data was computed for concepts that were designated as principal or main points in the same journal article i.e., the co-occurrence counts do not include articles in which either or both of the concepts were present and indexed in MEDLINE but not designated as main points. (A concept is considered to be a main point if the * is attached to the main heading or any of its subheadings.)

Two overall frequencies of MEDLINE co-occurrence are provided: one for recent MEDLINE data (MED) and one for MEDLINE data from a preceding block of years (MBD). Separate counts are provided for the frequencies with which the first concept was qualified by different MeSH qualifiers or by no qualifier at all when it co-occurred with the second concept. There are separate entries for each direction of the co-occurrence relationship. The related subheading occurrence information in each entry belongs to the first concept in the entry and is therefore different for each direction of the relationship.

In addition to the specific qualifier information associated with two co-occurring concepts, this element also includes in entries with LQ and LQB values for type of co-occurrence, totals for the number of times each main concept was qualified by a specific subheading or by no subheading.

The AI/RHEUM co-occurrence data represent the co-occurrence of diseases and findings in the AI/RHEUM knowledge base, i.e., the diseases that co-occur with a particular finding and the findings that co-occur with a particular disease. Each disease/finding pair can co-occur only once in the AI/RHEUM knowledge base.

In CCPSS, the co-occurrence data is extracted from patient records and includes problem-problem co-occurrences within a patient record as well as problem-modifier co-occurrences.

Sample Records

C0000294|A0085139|C0002423|A0022422|MED|L|1|AD=1,TU=1||

C0000294|A0085139|C0003962|A0026887|MBD|L|1|AA=1,BL=1,PK=1||

C0000294|A0085139|C0006434|A0033347|MBD|L|1|AD=1,PD=1||

Table 6. Co-occurring Concepts (File = MRCOC.RRF)

Col.	Description
CUI1	Unique identifier of first concept
AUI1	Unique identifier of first atom
CUI2	Unique identifier of second concept or not present Note: Where CUI2 is not present and COT is LQ (MeSh topical qualifier), the count of citations of CUI1 with no MeSH qualifiers is reported in COF.
AUI2	Unique identifier of second atom
SAB	Abbreviation of the source of co-occurrence information
COT	Type of co-occurrence
COF	Frequency of co-occurrence, if applicable
COA	Attributes of co-occurrence, if applicable
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.11. Computable Hierarchies (File = MRHIER.RRF)

This file contains one row for each hierarchy or context in which each atom appears. If a source vocabulary does not contain hierarchies, its atoms will have no rows in this file. If a source vocabulary is multi-hierarchical (allows the same atom to appear in more than one hierarchy), some of its atoms will have more than one row in this file. MRHIER.RRF (Table 7) provides a complete and compact representation of all hierarchies present in all Metathesaurus source vocabularies. Hierarchical displays can be computed by combining data in this file with data in MRCONSO.RRF. The distance-1 relationships, i.e., immediate parent, immediate child, and sibling relationships, represented in MRHIER.RRF also appear in MRREL.RRF.

Sample Records

C0001175|A2878223|1|A3316611|SNOMEDCT_US|isa|
A3684559.A3886745.A2880798.A24813547.A3082701.A3316611|||

C0001175|A2878223|2|A23017839|SNOMEDCT_US|isa|
A3684559.A3886745.A2880798.A24813547.A3082701.A3398847.A3398762.A2888699,A23017839|||

C0001175|A2878223|3|A3316611|SNOMEDCT_US|isa|
A3684559.A3886745.A2880798.A24813547.A3287869.A3316611|||

To find the specific concept names used in a hierarchy, look up the atom identifiers in the AUI and STR data elements in MRCONSO.RRF.

NLM editors do not assert concept-level (CUI-to-CUI) hierarchical relationships. Hierarchical relationships are asserted by sources at the atom level (AUI-to-AUI).

For most source vocabularies, the value of RELA (if present) applies up the hierarchy to the top or root. In other words, it also applies to the relationship between the atom's parent and the atom's grandparent, etc. The two exceptions in this version of the Metathesaurus are GO (Gene Ontology) and NIC (Nursing Intervention Classification). Except for GO and NIC atoms, the MRHIER rows for an atom's ancestors (parent, grandparent, etc.) contain no added information except the source-asserted hierarchical number or code (HCD). If this is not of interest, there may be no reason to find MRHIER rows for an atom's ancestors.

To find an atom's siblings in a specific context, find all MRHIER.RRF rows that share its SAB, RELA*, and PTR values.

To find an atom's children in a specific context, append a period (.) and the atom's AUI to its PTR and find all MRHIER.RRF rows with its SAB, RELA*, and the expanded PTR.

*The RELA is needed to retrieve correct siblings and children for University of Washington Digital Anatomist (UWDA) hierarchies. Some UWDA atoms appear in multiple hierarchies that are distinguished ONLY by their RELA values.

Table 7. Computable Hierarchies (File = MRHIER.RRF)

Col.	Description
CUI	Unique identifier of concept
AUI	Unique identifier of atom - variable length field, 8 or 9 characters
CXN	Context number (e.g., 1, 2, 3)
PAUI	Unique identifier of atom's immediate parent within this context

Table 7. continued from previous page.

Col.	Description
SAB	Abbreviated source name (SAB) of the source of atom (and therefore of hierarchical context). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
RELA	Relationship of atom to its immediate parent
PTR	Path to the top or root of the hierarchical context from this atom, represented as a list of AUIs, separated by periods (.) The first one in the list is top of the hierarchy; the last one in the list is the immediate parent of the atom, which also appears as the value of PAUI.
HCD	Source asserted hierarchical number or code for this atom in this context; this field is only populated when it is different from the code (unique identifier or code for the string in that source).
CVF	Content View Flag. Bit field used to flag rows included in Content View. This field is a varchar field to maximize the number of bits available for use.

3.3.12. Contexts (File = MRCXT.RRF)

This file is no longer created by default. It has been replaced by MRHIER.RRF which is a correct, complete, and computable representation of hierarchies. Users who require the MRCXT (Table 8) file will need to create that file after creating a subset. To create the MRCXT file use the new MRCXT Builder application, accessible from the MetamorphoSys Welcome screen. Information on the MRCXT Builder can be found at http://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/MRCXT_Builder.html. The information below describes the content of the file when produced by the MRCXT Builder.

This very large file contains pre-computed hierarchical context information (including concept names) intended to facilitate the display of hierarchies present in UMLS source vocabularies. All of the information in this file (plus additional sibling relationships) can be computed by joining the MRHIER.RRF file with MRCONSO.RRF. There can be many rows in this file for each occurrence of an atom in a hierarchy in any of the UMLS source vocabularies - a "context in" this discussion. Many Metathesaurus concepts have many atoms with contexts while others may have none. The number of rows per context differs depending on the number of ancestor, sibling, or child terms an atom has in that context. Because some atoms have multiple contexts in the same source, e.g., MeSH, a context number (CXN - e.g., 1, 2, 3) is used to identify all members of the same context. The CXNs are not global but are created as required for each atom. Each distinct context for a single atom can be retrieved with a CUI-AUI-SAB-CXN key. The "distance-1 relationships" i.e., the immediate parent, immediate child, and sibling relationships represented in MRCXT.RRF are also present in the MRREL.RRF file.

Sample Records

C0001175|S0011877|A0021048|CSP|1560-6271|4|ANC|5|acquired immunodeficiency|C0596032|A1171599|||||

C0001175|S0011877|A0021048|CSP|1560-6271|4|CCP||AIDS|C0001175|A0021048|||||

C0001175|S0011877|A0021048|CSP|1560-6271|4|CHD||AIDS related neoplasm/cancer|C0920774|A1882809|||||

C0001175|S0011877|A0021048|CSP|1560-6271|4|SIB||hairy cell leukemia|C0023443|A0480441|||||

Table 8. Contexts (File = MRCXT.RRF)

Col.	Description
CUI	Unique identifier of concept
SUI	Unique identifier of string used in this context
AUI	Unique identifier of atom that has this context (variable length field, 8 or 9 characters)
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
CODE	Unique identifier or code for string in that source
CXN	The context number (if the atom has multiple contexts)
CXL	Context member label, i.e., ANC for ancestor of this atom, CCP for the atom itself, SIB for sibling of this atom, CHD for child of this atom
RNK	For rows with a CXL value of ANC, the rank of the ancestors (e.g., a value of 1 denotes the most remote ancestor in the hierarchy)
CXS	String or concept name for context member
CUI2	Concept identifier of context member (may be empty if context member is not yet in the Metathesaurus)
AUI2	Atom identifier of context member
HCD	Source hierarchical number or code of context member (if present)
RELA	Additional relationship label providing further categorization of the CXL, if applicable and known. Valid values listed on the Abbreviations Used in Data Elements page.
XC	A plus (+) sign indicates that the CUI2 for this row has children in this context. If this field is empty, the CUI2 does not have children in this context.
CVF	Content View Flag. Bit field used to flag rows included in Content View.

3.3.13. Mappings (File = MRMAP.RRF)

This file contains sets of mappings between vocabularies. Most mappings are between codes/identifiers (or expressions formed by codes/identifiers) from two different vocabularies. At least one of the vocabularies in each set of mappings is present in the Metathesaurus; usually both of them are. The version of a vocabulary that appears in a set of mappings may be different from the version of that vocabulary that appears in the other Metathesaurus release files. The versions of the vocabularies in a map set are specified by the FROMVSAB and TOVSAB attributes of the map set concept (see below). Users should be aware that the mappings are only valid between the versions of the vocabularies specified in these attributes. The version of the map set itself is specified by the MAPSETVERSION attribute of the map set concept.

The MRMAP.RRF (Table 9) file is complex, to allow for more complicated mappings. Where possible, all mappings are also represented in the simpler MRSMAP.RRF file described below.

Each set of mappings is represented by a map set concept in MRCONSO.RRF (with TTY = 'XM') identified by a CUI (MAPSETCUI). Metadata of a map set are found in MRSAT.RRF as attributes of the map set concept. Each map set has three SAB values associated with it: the SAB of the map set itself (MAPSETVSAB), the SAB of the

source being mapped (FROMVSAB) and the SAB of the source being mapped to (TOVSAB). Thus, a single map set asserts mappings from only one source to only one other source.

A subset of the mappings is redundantly represented as mapped_to and mapped_from relationships in MRREL.RRF. These are one-to-one mappings between two vocabularies which are both present in the UMLS. These general relationships are not as precise as the mapping files, since any differences between versions of the vocabularies in the map set and the versions of those vocabularies in the rest of the Metathesaurus files are ignored. Such differences may affect the validity of the relationships in MRREL.RRF in a small number of cases.

There are three map sets that contain mappings from Metathesaurus concepts (represented by CUIs) to expressions formed by one or more concept names. These were formerly called associated expressions, and all have MAPTYPE='ATX'. This data is derived from earlier mapping efforts and is represented in the MRATX file in ORF.

Sample Records
Map set concepts (in MRCONSO.RRF):

C1306694|ENG|P|L14542194|PF|S17644451|Y|A28926527||||MTH|XM|1000|MSH2018_2018_02_05 Associated Expressions|0|N||

Map set metadata (in MRSAT.RRF):

C1306694|L14542194|S17644451|A28926527|CODE|1000|AT232101656||MAPSETVERSION|MTH|2018_2018_02_05|N||

C1306694|L14542194|S17644451|A28926527|CODE|1000|AT232101657||TOVSAB|MTH|MSH2018_2018_02_05|N||

Mappings (in MRMAP.RRF):

C1306694|MTH|||AT28307527||C0011764||C0011764|CUI|||RO||2201||<Developmental Disabilities> AND <Writing>|BOOLEAN_EXPRESSION_STR|||||ATX||||

C1306694|MTH|||AT52620421||C0010700||C0010700|CUI|||RN||1552||<Urinary Bladder>/<surgery>|BOOLEAN_EXPRESSION_STR|||||ATX||||

Table 9. Mappings (File = MRMAP.RRF)

Col.	Description
MAPSETCUI	Unique identifier for the UMLS concept which represents the whole map set.
MAPSETSAB	Source abbreviation (SAB) for the provider of the map set.
MAPSUBSETID	Map subset identifier used to identify a subset of related mappings within a map set. This is used for cases where the FROMEXPR may have more than one potential mapping (optional).
MAPRANK	Order in which mappings in a subset should be applied. Used only where MAPSUBSETID is used. (optional)
MAPID	Unique identifier for this individual mapping. Primary key of this table to identify a particular row.
MAPSID	Source asserted identifier for this mapping (optional).
FROMID	Identifier for the entity being mapped from. This is an internal UMLS identifier used to point to an external entity in a source vocabulary (represented by the FROMEXPR). When the source provides such an identifier, it is reused here. Otherwise, it is generated by NLM. The FROMID is only unique within a map set. It is not a pointer to UMLS entities like atoms or concepts. There is a one-to-one correlation between FROMID and a unique set of values in FROMSID, FROMEXPR, FROMTYPE, FROMRULE, and FROMRES within a map set.
FROMSID	Source asserted identifier for the entity being mapped from (optional).

Table 9. continued from previous page.

Col.	Description
FROMEXPR	Entity being mapped from - can be a single code/identifier /concept name or a complex expression involving multiple codes/identifiers/concept names, Boolean operators and/or punctuation
FROMTYPE	Type of entity being mapped from.
FROMRULE	Machine processable rule applicable to the entity being mapped from (optional)
FROMRES	Restriction applicable to the entity being mapped from (optional).
REL	Relationship of the entity being mapped from to the entity being mapped to.
RELA	Additional relationship label (optional).
TOID	Identifier for the entity being mapped to. This is an internal identifier used to point to an external entity in a source vocabulary (represented by the TOEXPR). When the source provides such an identifier, it is reused here. Otherwise, it is generated by NLM. The TOID is only unique within a map set. It is not a pointer to UMLS entities like atoms or concepts. There is a one-to-one correlation between TOID and a unique set of values in TOSID, TOEXPR, TOTYPE, TORULE, TORES within a map set.
TOSID	Source asserted identifier for the entity being mapped to (optional).
TOEXPR	Entity being mapped to - can be a single code/identifier/concept name or a complex expression involving multiple codes/identifiers/concept names, Boolean operators and/or punctuation.
TOTYPE	Type of entity being mapped to.
TORULE	Machine processable rule applicable to the entity being mapped to (optional).
TORES	Restriction applicable to the entity being mapped to (optional).
MAPRULE	Machine processable rule applicable to this mapping (optional).
MAPRES	Restriction applicable to this mapping (optional).
MAPTYPE	Type of mapping (optional).
MAPATN	The name of the attribute associated with this mapping [not yet in use]
MAPATV	The value of the attribute associated with this mapping [not yet in use]
CVF	The Content View Flag is a bit field used to indicate membership in a content view.

3.3.14. Simple Mappings (File = MRSMAP.RRF)

This file provides a simpler representation of most of the mappings in MRMAP.RRF (Table 10) to serve applications which do not require the full richness of the MRMAP.RRF data structure. Generally, mappings that support rule-based processing need the additional fields of MRMAP.RRF (e.g. MAPRANK, MAPRULE, MAPRES) and will not be represented in MRSMAP.RRF. More specifically, all mappings with non-null values for MAPSUBSETID and MAPRANK are excluded from MRSMAP.RRF.

Sample Records

C1306694|MTH|AT28312030||C0009215|CUI|SY||<Codeine> AND <Drug Hypersensitivity>|BOOLEAN_EXPRESSION_STR||

C1306694|MTH|AT28312033||C0795964|CUI|RU||<Speech Disorders>|BOOLEAN_EXPRESSION_STR||

Table 10. Simple Mappings (File = MRSMAP.RRF)

Col.	Description
MAPSETCUI	Unique identifier for the UMLS concept which represents the whole map set.
MAPSETSAB	Source abbreviation for the map set.

Table 10. continued from previous page.

Col.	Description
MAPID	Unique identifier for this individual mapping. Primary key of this table to identify a particular row.
MAPSID	Source asserted identifier for this mapping (optional).
FROMEXPR	Entity being mapped from - can be a single code/identifier/concept name or a complex expression involving multiple codes/identifiers/concept names, Boolean operators and/or punctuation.
FROMTYPE	Type of entity being mapped from.
REL	Relationship of the entity being mapped from to the entity being mapped to.
RELA	Additional relationship label (optional).
TOEXPR	Entity being mapped to - can be a single code/identifier /concept name or a complex expression involving multiple codes/identifiers/concept names, Boolean operators and/or punctuation.
TOTYPE	Type of entity being mapped to.
CVF	The Content View Flag is a bit field used to indicate membership in a content view.

3.3.15. Source Information (File = MRSAB.RRF)

The Metathesaurus has "versionless" or "root" Source Abbreviations (SABs) in the data files. MRSAB.RRF connects the root SAB to fully specified version information for the current release. For example, the released SAB for MeSH is now simply "MSH". In MRSAB.RRF (Table 11), you will see a current versioned SAB, e.g., MSH2003_2002_10_24. MRSAB.RRF allows all other Metathesaurus files to use versionless source abbreviations, so that all rows with no data change between versions remain unchanged. MetamorphoSys can produce files with either the root or versioned SABs so that either form can be available in custom subsets of the Metathesaurus.

There is one row in this file for every version of every source in the current Metathesaurus; eventually there will also be historical information with a row for each version of each source that has appeared in any Metathesaurus release. Note that the field CURVER has the value Y to identify the version in this Metathesaurus release. Future releases of MRSAB.RRF will also contain historical version information in rows with CURVER value N.

Sources with contexts have "full" contexts, i.e., all levels of terms may have Ancestors, Parents, Children and Siblings. A full context may also be further designated as Multiple, Nosib (No siblings) or both Multiple and Nosib.

Multiple indicates that a single concept in this source may have multiple hierarchical positions.

No siblings (Nosib) indicates that siblings have not been computed for this source.

The UMLS Source Vocabulary Documentation page of the current release documentation lists each source in the Metathesaurus and includes information about the type of context, if any, for each source.

Sample Records

C4550278|C1140284|RXNORM_17AB_180305F|RXNORM|RxNorm Vocabulary, 17AB_180305F|RXNORM|17AB_180305F|||2018AA||RxNorm Customer Service;;U.S. National Library of Medicine;8600 Rockville Pike;;Bethesda;MD;United States;20894;(888) FIND-NLM;;rxnorminfo@nlm.nih.gov;https://www.nlm.nih.gov/research/umls/rxnorm/|RxNorm Customer Service;;U.S. National Library of Medicine;8600 Rockville Pike;;Bethesda;MD;United States;20894;(888) FIND-NLM;;rxnorminfo@nlm.nih.gov;https://www.nlm.nih.gov/research/umls/rxnorm/|0|319274|208301||
BN,BPCK,DF,DFG,ET,GPCK,IN,MIN,PIN,PSN,SBD,SBDC,SBDF,SBDG,SCD,SCDC,SCDF,SCDG,SY,TMSY|AMBIGUITY_FLAG,NDC,ORIG_CODE,ORIG_SOURCE,RXAUI,RXCUI,RXN_ACTIVATED,RXN_AVAILA

BLE_STRENGTH,RXN_BN_CARDINALITY,RXN_HUMAN_DRUG,RXN_IN_EXPRESSED_FLAG,RXN_OBSOLETED,RXN_QUALITATIVE_DISTINCTION,RXN_QUANTITY,RXN_STRENGTH,RXN_VET_DRUG,RXTERM_FORM|ENG|UTF-8|Y|Y|RXNORM|;;;;RxNorm;;;META2017AB Full Update 2018_03_05;Bethesda, MD;National Library of Medicine;;;;;;;;;|

Table 11. Source Information (File = MRSAB.RRF)

Field	Full Name	Description
VCUI	CUI	CUI of the versioned SRC concept for a source
RCUI	Root CUI	CUI of the root SRC concept for a source
VSAB	Versioned Source Abbreviation	The versioned source abbreviation for a source, e.g., MSH2003_2002_10_24
RSAB	Root Source Abbreviation	The root source abbreviation for a source e.g., MSH
SON	Official Name	The official name for a source
SF	Source Family	The source family for a source
SVER	Version	The source version, e.g., 2001
VSTART	Meta Start Date	The date a source became active, e.g., 2001_04_03
VEND	Meta End Date	The date a source ceased to be active, e.g., 2001_05_10
IMETA	Meta Insert Version	The version of the Metathesaurus in which a source first appeared, e.g., 2001AB
RMETA	Meta Remove Version	The version of the Metathesaurus in which the source last appeared, e.g., 2001AC
SLC	Source License Contact	The source license contact field contains the following semi-colon-separated subfields: Name Title Organization Address 1 Address 2 City State/Prov. Country Zip Telephone Fax Email URL
SCC	Source Content Contact	The source content contact field contains the following semi-colon-separated subfields: Name Title Organization Address 1 Address 2 City State/Prov. Country Zip Telephone Fax Email URL
SRL	Source Restriction Level	0, 1, 2, 3, 4, 9 - explained in the License Agreement

Table 11. continued from previous page.

Field	Full Name	Description
TFR	Term Frequency	The number of terms for this source in MRCONSO.RRF, e.g., 12343
CFR	CUI Frequency	The number of CUIs associated with this source, e.g., 10234
CXTY	Context Type	The type of contexts for this source. Values are FULL, FULL-MULTIPLE, FULL-NOSIB, FULL-NOSIB-MULTIPLE, FULL-MULTIPLE-NOSIB-RELA, null.
TTYL	Term Type List	Term type list from source, e.g., MH, EN, PM, TQ
ATNL	Attribute Name List	The attribute name list (from MRSAT.RRF), e.g., MUI, RN, TH
LAT	Language	The language of the terms in the source
CENC	Character Encoding	All UMLS content is provided in Unicode, encoded in UTF-8. MetamorphoSys will allow exclusion of extended characters with some loss of information. Transliteration to other character encodings is possible but not supported buy NLM; for further information, see http://www.unicode.org
CURVER	Current Version	A Y or N flag indicating whether or not this row corresponds to the current version of the named source
SABIN	Source in Subset	A Y or N flag indicating whether or not this row is represented in the current MetamorphoSys subset. Initially always Y where CURVER is Y, but later is recomputed by MetamorphoSys.
SSN	Source Short Name	The short name of a source as used by the UMLS Terminology Services
SCIT	Source Citation	For sources released in 2014AA and later, the citation field contains the following semi-colon-separated subfields: Author name(s) Personal author address Organization author(s) Editor(s) Title Content Designator Medium Designator Edition Place of Pub. Publisher Date of pub. or copyright Date of revision Location Extent Series Avail. Statement (URL) Language Notes Empty Subfield Empty Subfield The citation field for sources released prior to 2014AA will be updated as resources permit.

3.3.16. Concept Name Ranking (File = MRRANK.RRF)

There is exactly one row for each concept name type from each Metathesaurus source vocabulary (each SAB-TTY combination). The RANK and SUPPRESS values in the distributed file are those used in Metathesaurus production. Users are free to change these values to suit their needs and preferences, then change the naming precedence and suppressibility by using MetamorphoSys to create a customized Metathesaurus. (Table 12)

Sample Records

0624|AIR|SY|N|

0438|PDQ|IS|Y|

0377|LNC|LO|Y|

Table 12. Concept Name Ranking (File = MRRANK.RRF)

Col.	Description
RANK	Numeric order of precedence, higher value wins
SAB	Abbreviated source name (SAB) for source vocabulary. Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
TTY	Abbreviation for term type in source vocabulary, for example PN (Metathesaurus Preferred Name) or CD (Clinical Drug). Possible values are listed in Abbreviations Used in Data Elements page.
SUPPRESS	NLM-recommended Source and Term Type (SAB/TTY) Suppressiblity. Values = Y or N. Indicates the suppressible status of all atoms (names) with this Source and Term Type (SAB/TTY). Note that changes made in MetamorphoSys at the Suppressible tab are recorded in your configuration file. Status E does not occur here, as it is assigned only to individual cases such as the names (atoms) in MRCONSO.RRF. See also SUPPRESS in MRCONSO.RRF, MRDEF.RRF, and MRREL.RRF.

3.3.17. Ambiguous Term Identifiers (File = AMBIGLUI.RRF)

In the instance that a Lexical Unique Identifier (LUI) is linked to multiple Concept Unique Identifiers (CUIs), there is one row in this table for each LUI-CUIs pair. This file identifies those lexical variant classes which have multiple meanings in the Metathesaurus.

In the Metathesaurus, the LUI links all strings within the English language that are identified as lexical variants of each other by the luinorm program found in the UMLS SPECIALIST Lexicon and Tools. LUIs are assigned irrespective of the meaning of each string. This table may be useful to system developers who wish to use the lexical programs in their applications to identify and disambiguate ambiguous terms.

Col.	Description
LUI	Lexical Unique Identifier
CUI	Concept Unique Identifier

Sample Records

L0000003|C0010504|

L0000003|C0917995|

L0000032|C0010206|

L0000032|C0010207|

3.3.18. Ambiguous String Identifiers (File = AMBIGSUI.RRF)

In the instance that a String Unique Identifier (SUI) is linked to multiple Concept Unique Identifiers (CUIs), there is one row in this table for each SUI-CUIs pair.

This file resides in the META directory. In the Metathesaurus, there is only one SUI for each unique string within each language, even if the string has multiple meanings. This table is only of interest to system developers who use the SUI in their applications or in local data files.

Col.	Description
SUI	String Unique Identifier
CUI	Concept Unique Identifier

Sample Records

S0000176|C0042266|

S0000176|C2004487|

S0000217|C0024817|

S0000217|C0555026|

3.3.19. Metathesaurus Change Files

There are six files or relations that identify key differences between entries in the previous and the current edition of the Metathesaurus. Developers can use these special files to determine whether there have been changes that affect their applications.

The usefulness of individual files will depend on how data from the Metathesaurus have been linked or incorporated in a particular application.

Each relation or named table of data has a fixed number of columns and variable number of rows. A column is a sequence of all the values in a given data element. A row contains the values for two or more data elements for one entry. The values for the different data elements in the row are separated by vertical bars (|). Each row ends with a vertical bar and line termination.

3.3.19.1. Deleted Concepts (File = CHANGE/DELETEDCUI.RRF)

Concepts whose meaning is no longer present in the Metathesaurus are reported in this file. There is a row for each concept that existed in the previous release and is not present in the current release. If the meaning exists in the current release, i.e., the missing concept was merged with another current concept, it is reported in the MERGEDCUI.RRF file (Section 3.3.19.2) and not in this file.

Col.	Description
PCUI	Concept Unique Identifier in the previous Metathesaurus
PSTR	Preferred name of this concept in the previous Metathesaurus

3.3.19.2. Merged Concepts (File = CHANGE/MERGEDCUI.RRF)

There is exactly one row in this table for each released concept in the previous Metathesaurus (CUI1) that was merged into another released concept from the previous Metathesaurus (CUI2). When this merge occurs, the first CUI (CUI1) was retired; this table shows the CUI (CUI2) for the merged concept in this Metathesaurus.

Entries in this file represent concepts pairs that were considered to have different meanings in the previous edition, but which are now identified as synonyms.

Col.	Description
PCUI1	Concept Unique Identifier in the previous Metathesaurus

Table continued from previous page.

Col.	Description
CUI	Concept Unique Identifier in this Metathesaurus in format C#######

3.3.19.3. Deleted Terms (File=CHANGE/DELETEDLUI.RRF)

There is exactly one row in this table for each Lexical Unique Identifier (LUI) that appeared in the previous Metathesaurus, but does not appear in this Metathesaurus.

LUIs are assigned by the luinorm program, part of the lvg program in the UMLS SPECIALIST Lexicon and Tools; see Chapter 6.

These entries represent the cases where LUIs identified by the previous release's luinorm program, when used to identify lexical variants in the previous Metathesaurus, are no longer found with this release's luinorm on this release's Metathesaurus. This does not necessarily imply the deletion of a string or a concept from the Metathesaurus.

Col.	Description
PLUI	Lexical Unique Identifier in the previous Metathesaurus
PSTR	Preferred Name of Term in the previous Metathesaurus

3.3.19.4. Merged Terms (File = CHANGE/MERGEDLUI.RRF)

There is exactly one row in this file for each case in which strings had different Lexical Unique Identifiers (LUIs) in the previous Metathesaurus yet share the same LUI in this Metathesaurus; a LUI present in the previous Metathesaurus is therefore absent from this Metathesaurus.

LUIs are assigned by the luinorm program, part of the lvg program in the UMLS SPECIALIST Lexicon and Tools; see Chapter 6.

These entries represent the cases where separate lexical variants as identified by the previous release's luinorm program version are a single lexical variant as identified by this release's luinorm.

Col.	Description
PLUI	Lexical Unique Identifier in the previous Metathesaurus but not present in this Metathesaurus
LUI	Lexical Unique Identifier into which it was merged in this Metathesaurus

3.3.19.5. Deleted Strings (File = CHANGE/DELETEDSUI.RRF)

There is exactly one row in this file for each string in each language that was present in an entry in the previous Metathesaurus and does not appear in this Metathesaurus.

Note that this does not necessarily imply the deletion of a term (LUI) or a concept (CUI) from the Metathesaurus. A string deleted in one language may still appear in the Metathesaurus in another language.

Col.	Description
PSUI	String Unique Identifier in the previous Metathesaurus that is not present in this Metathesaurus
PSTR	Preferred Name of Term in the previous Metathesaurus that is not present in this Metathesaurus

3.3.19.6. Retired CUI Mapping (File = MRCUI.RRF)

There are one or more rows in this file (Table 13) for each Concept Unique Identifier (CUI) that existed in any prior release but is not present in the current release. The file includes mappings to current CUIs as synonymous

or to one or more related current CUI where possible. If a synonymous mapping cannot be found, other relationships between the CUIs can be created. These relationships can be Broader (RB), Narrower (RN), Other Related (RO), Deleted (DEL) or Removed from Subset (SUBX). Rows with the SUBX relationship are added to MRCUI by MetamorphoSys for each CUI that met the exclusion criteria and was consequently removed from the subset. Some CUIs may be mapped to more than one other CUI using these relationships.

CUIs may be retired when (1) two released concepts are found to be synonyms and so are merged, retiring one CUI; (2) the concept no longer appears in any source vocabulary and is not 'rescued' by NLM; or (3) the concept is an acknowledged error in a source vocabulary or determined to be a Metathesaurus production error.

See Sections 3.3.19 1 through 5 for files containing changes from the last release only, without mappings.

Sample Records

C1313903|2004AA|SY|||C0525045|Y|

C1313909|2004AA|RO|||C0476661|Y|

C2732033|2010AA|RO|||C0025942|Y|

Table 13. Retired CUI Mapping (File = MRCUI.RRF)

Col.	Description
CUI1	Unique identifier for first concept - Retired CUI - was present in some prior release, but is currently missing
VER	The last release version in which CUI1 was a valid CUI
REL	Relationship
RELA	Relationship attribute
MAPREASON	Reason for mapping
CUI2	Unique identifier for second concept - the current CUI that CUI1 most closely maps to
MAPIN	Is this map in current subset? Values of Y, N, or null. MetamorphoSys generates the Y or N to indicate whether the CUI2 concept is or is not present in the subset. The null value is for rows where the CUI1 was not present to begin with (i.e., REL=DEL).

3.3.19.7. AUI Movements (File = MRAUI.RRF)

This file records the movement of Atom Unique Identifiers (AUIs) from a concept (CUI1) in one version of the Metathesaurus to a concept (CUI2) in the next version (VER) of the Metathesaurus. The file is historical. (Table 14)

Sample Records

A0000039|C0236824|2004AC|||move|A0000039|C1411876|Y|

A0000077|C1510447|2007AC|||move|A0000077|C0003477|Y|

A9460778|C1696703|2009AB|||move|A9460778|C0023067|Y|

Table 14. AUI Movements (File = MRAUI.RRF)

Col.	Description
AUI1	Atom unique identifier
CUI1	Concept unique identifier
VER	Version in which this change to the AUI first occurred
REL	Relationship

Table 14. continued from previous page.

Col.	Description
RELA	Relationship attribute
MAPREASON	Reason for mapping
AUI2	Unique identifier for second atom
CUI2	Unique identifier for second concept - the current CUI that CUI1 most closely maps to
MAPIN	Mapping in current subset: is AUI2 in current subset? Values of Y, N, or null.

3.3.20. Word Index (File = MRXW_BAQ.RRF, MRXW_DAN.RRF, MRXW_DUT.RRF, MRXW_ENG.RRF, MRXW_FIN.RRF, MRXW_FRE.RRF, MRXW_GER.RRF, MRXW_HEB.RRF, MRXW_HUN.RRF, MRXW_ITA.RRF, MRXW_NOR.RRF, MRXW_POR.RRF, MRXW_RUS.RRF, MRXW_SPA.RRF, MRXW_SWE.RRF)

There is one row in these tables for each word found in each unique Metathesaurus string (ignoring upper-lower case). All Metathesaurus entries have entries in the word index. The entries are sorted in ASCII order.

Col.	Description
LAT	Abbreviation of language of the string in which the word appears
WD	Word in lowercase
CUI	Concept identifier
LUI	Term identifier
SUI	String identifier

Sample Records from MRXW_ENG.RRF

ENG|anaemia|C0002871|L0002871|S0352688|

ENG|anemia|C0002871|L0002871|S0013742|

ENG|disorder|C0002871|L2818006|S3448137|

ENG|unspecified|C0002871|L0503461|S0589617|

Sample Records from MRXW_FRE.RRF

FRE|ANEMIE|C0002871|L0162748|S0227229|

3.3.21. Normalized Word Index (File = MRXNW_ENG.RRF)

There is one row in this table for each normalized word found in each unique English-language Metathesaurus string. All English-language Metathesaurus entries have entries in the normalized word index. There are no normalized string indexes for other languages in the Metathesaurus.

Col.	Description
LAT	Abbreviation of language of the string in which the word appears (always ENG in this edition of the Metathesaurus)
NWD	Normalized word in lowercase (described in Section 2.7.2.1)
CUI	Concept identifier
LUI	Term identifier

Table continued from previous page.

Col.	Description
SUI	String identifier

Sample Records

ENG|anemia|C0002871|L0002871|S0013742|

ENG|anemia|C0002871|L0002871|S0013787|

ENG|disorder|C0002871|L2818006|S3448137|

ENG|unspecified|C0002871|L0503461|S0589617|

3.3.22. Normalized String Index (File = MRXNS_ENG.RRF)

There is one row in this table for each normalized string found in each unique English-language Metathesaurus string (ignoring upper-lower case). All English-language Metathesaurus entries have entries in the normalized string index. There are no normalized word indexes for other languages in this edition of the Metathesaurus.

Col.	Description
LAT	Abbreviation of language of the string (always ENG in this edition of the Metathesaurus)
NSTR	Normalized string in lowercase (described in Section 2.7.3.1)
CUI	Concept identifier
LUI	Term identifier
SUI	String identifier

Sample Records

ENG|anemia disorder|C0002871|L2822821|S3436848|

ENG|anemia unspecified|C0002871|L0503461|S0589617|

ENG|anemia|C0002871|L0002871|S0013742|

4. Metathesaurus - Original Release Format (ORF)

Metathesaurus users may select from two relational formats: the Rich Release Format (RRF), first introduced in 2004, and the Original Release Format (ORF). Both are available as output options of MetamorphoSys, the installation and customization program.

Developers are encouraged to use the RRF, which offers significant advantages in source vocabulary transparency (that is, ability to exactly represent the detailed semantics of each source vocabulary); in the ability to generate complete and accurate change sets between versions of the Metathesaurus; and in more convenient representations of concept name, source, and hierarchical context information.

Neither Metathesaurus format is fully normalized. By design, there is duplication of data among different files and within certain files. In particular, relationships between different Metathesaurus concepts appear twice (e.g., from entry A to entry B and from entry B to entry A). Developers will need to make their own decisions about the extent to which this redundancy should be retained, reduced, or increased for their specfic applications.

Note: The preferred and more complete format is described in Chapter 3, the Metathesaurus Rich Release Format (RRF).

All files except MRRANK are sorted by row.

4.1. Data Files

The data in each Metathesaurus entry may be represented in more than 20 different "relations" or files. These files correspond to the four logical groups of data elements described in Section 2.3 - 2.6 and the indexes described in Section 2.7 as follows:

- Metathesaurus concept names and their sources (2.3) = MRCON, MRSO
- Attributes (2.5) = MRSAT, MRDEF, MRSTY
- Relationships between different concept names (2.4) = MRREL, MRATX, MRCXT
- Data about the Metathesaurus (2.6)=MRSAB, MRRANK, AMBIG.LUI, AMBIG.SUI, DELETED.CUI, MERGED.CUI, DELETED.LUI, MERGED.LUI, DELETED.SUI, MRCUI
- Indexes (2.7) = MRXW.BAQ, MRXW.DAN, MRXW.DUT, MRXW.ENG, MRXW.FIN, MRXW.FRE, MRXW.GER, MRXW.HEB, MRXW.HUN, MRXW.ITA, MRXW.NOR, MRXW.POR, MRXW.RUS,MRXW.SPA, MRXW.SWE, MRXNW.ENG, MRXNS.ENG

The AMBIG* files now provide a convenient way to identify all Metathesaurus terms and strings that have more than one meaning in Metathesaurus source vocabularies.

4.2. Columns and Rows

Each relation or named table of data values has by definition a fixed number of columns; the number of rows depends on the content of a particular version of the Metathesaurus.

A column is a sequence of all the values in a given data element or logical sub-element. In general, columns for longer variable length data elements will appear to the right of columns for shorter and/or fixed length data elements. The information for all columns in the ORF files is described on the Columns and Data Elements page of the current release documentation.

A row contains the values for one or more data elements or logical sub-elements for one Metathesaurus entry. Depending on the nature of the data elements involved, each Metathesaurus entry may have one or more rows in a given file. The values for the different data elements or logical sub-elements represented in the row are

separated by vertical bars (|). If an optional element is blank, the vertical bars are still used to maintain the correct positioning of the subsequent elements. Each row is terminated by a vertical bar and line termination.

4.3. Descriptions of Each File

The descriptions of the files appear in the following order:

- Key data about the Metathesaurus: Files, columns or data elements
- Concept names and their vocabulary sources
- Attributes
- Relationships
- Other data about the Metathesaurus
- Indexes

4.3.1. Files (File = MRFILES)

There is exactly one row in this file for each physical segment of the files in the relational format. The columns or data elements in the file are as follows:

Col.	Description
FIL	Physical FILENAME
DES	Descriptive name
FMT	Comma separated list of COL, in order
CLS	# of COLUMNS
RWS	# of ROWS
BTS	Size in bytes in this format (ISO/PC or Unix)

Sample Records

MRATX|Associated Expressions|CUI,SAB,REL,ATX|4|8451|454611|

MRCOLS|Attribute Relation|COL,DES,REF,MIN,AV,MAX,FIL,DTY|8|220|13546|

4.3.2. Data Elements (File = MRCOLS)

There is exactly one row in this file for each column or data element in each file in the relational format.

Col.	Description
COL	Column or data element name
DES	Descriptive name
REF	Documentation section number
MIN	Minimum length, characters
AV	Average length
MAX	Maximum length, characters
FIL	Physical FILENAME in which this field occurs
DTY	SQL-92 data type for this column

Sample Records

ATN|Attribute name||2|8.03|29|MRSAT|varchar(50)|

ATV|Attribute value||0|7.66|7903|MRSAT|varchar(8000)|

ATX|Associated expression||5|35.79|242|MRATX|varchar(300)|

4.3.3. Concept Names (File = MRCON)

There is exactly one row in this file for each meaning of each unique string in the Metathesaurus, i.e., there is exactly one row for each unique CUI-SUI combination in the Metathesaurus. Any difference in upper-lower case, word order, etc., creates a different unique string.

Col.	Description
CUI	Unique identifier for concept
LAT	Language of term
TS	Term status
LUI	Unique identifier for term
STT	String type
SUI	Unique identifier for string
STR	String
LRL	Least restriction level

Sample Records

C0002871|ENG|P|L0002871|VC|S0352787|ANEMIA|0|

C0002871|ENG|P|L0002871|VC|S0414880|anemia|0|

C0002871|ENG|P|L0002871|VO|S0013787|Anemias|0|

C0002871|ENG|P|L0002871|VO|S0352688|ANAEMIA|0|

C0002871|ENG|P|L0002871|VO|S0470050|Anaemia, NOS|9|

C0002871|ENG|P|L0002871|VO|S0470197|Anemia, NOS|0|

C0002871|ENG|S|L0503461|PF|S0804082|Anemia unspecified|3|

4.3.4. Vocabulary Sources (File = MRSO)

This file contains the vocabulary source(s) for a concept, term, and string.

There is exactly one row in this file for each source of each string in the Metathesaurus. All Metathesaurus concepts have entries in this file.

Col.	Description
CUI	Unique identifier for concept
LUI	Unique identifier for term
SUI	Unique identifier for string

Table continued from previous page.

Col.	Description
SAB	Abbreviated source name (SAB) for source vocabulary. Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
TTY	Abbreviation for term type in source vocabulary, for example PN (Metathesaurus Preferred Name) or CD (Clinical Drug). Possible values are listed on the Abbreviations Used in Data Elements page.
CODE	Unique identifier or code for string in that source
SRL	Source restriction level

Sample Records

C0002871|L0002871|S0013742|SNOMEDCT|OP|154786001|9|

C0002871|L0002871|S0013742|SNOMEDCT|OP|64593003|9|

C0002871|L0002871|S0013742|SNOMEDCT|PT|271737000|9|

C0002871|L0002871|S0013787|MSH|PM|D000740|0|

C0002871|L0002871|S0352688|CST|GT|ANEMIA|0|

C0002871|L0002871|S0352688|WHO|PT|0544|2|

C0002871|L0002871|S0352787|CCPSS|PT|1017210|3|

The information in MRSO can be used in combination with MRCON to determine whether a particular concept, name, or code is present in a particular source, and in what form it appears.

Note: In the RRF, the concept name and vocabulary source information appear in a single file, MRCONSO.RRF.

4.3.5. Simple Concept and String Attributes (File = MRSAT)

There is exactly one row in this table for each concept, term and string attribute that does not have a sub-element structure. All Metathesaurus concepts have entries in this file.

Col.	Description
CUI	Unique identifier for concept
LUI	Unique identifier for term (optional)
SUI	Unique identifier for string (optional)
CODE	Unique identifier or code for entry in the source of the attribute, e.g., for all attributes derived from MeSH, the MeSH unique identifier (optional).
ATN	Attribute name. Possible values are all described in Attribute Names page.

Table continued from previous page.

Col.	Description
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
ATV	Attribute value described under specific attribute name in Attribute Names page. A few attribute values exceed 1,000 characters.

Sample Records

C0002871|L0002871|S0013742|D000740|MMR|MSH|19960610|

C0002871|L0002871|S0013742|D000740|MN|MSH|C15.378.071|

C0002871|L0002871|S0013742|D000740|TERMUI|MSH|T002209|

C0002871|L0002871|S0013742|D000740|TH|MSH|POPLINE (1994)|

C0002871|L0002871|S0470197|DC-10010|SIC|SNMI|285.9|

C0002871|L0002871|S0803242|271737000|LANGUAGECODE|SNOMEDCT|en-GB|

4.3.6. Definitions (File = MRDEF)

There is exactly one row in this file for each definition in the Metathesaurus. A few definitions approach 3,000 characters in length.

Col.	Description
CUI	Unique identifier for concept
SAB	Abbreviated source name (SAB) of the source of the definition. Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
DEF	Definition

Sample Records

C0002871|CSP|subnormal levels or function of erythrocytes, resulting in symptoms of tissue hypoxia.|

C0002871|MSH|A reduction in the number of circulating erythrocytes or in the quantity of hemoglobin.|

C0002871|NCI|(a-NEE-mee-a) A condition in which the number of red blood cells is below normal.|

4.3.7. Semantic Types (File = MRSTY)

There is exactly one row in this file for each semantic type assigned to each concept. All Metathesaurus concepts have at least one entry in this file. Many have more than one entry.

Col.	Description
CUI	Unique identifier of concept
TUI	Unique identifier of Semantic Type
STY	Semantic Type. The valid values are defined in the Semantic Network.

Sample Record

C0002871|T047|Disease or Syndrome|

4.3.8. Locators (File = MRLO)

This file has been deleted from the Metathesaurus effective with the 2004AB release. Some of the information was outdated, some duplicated information contained in other Metathesaurus files, and some was easily obtained from other publicly available sources, e.g., PubMed.

4.3.9. Related Concepts (File = MRREL)

There is one row in this table for each relationship between Metathesaurus concepts known to the Metathesaurus, with the following exceptions found in other files: Associated Expressions found in MRATX.

Note that for asymmetrical relationships there is one row for each direction of the relationship. Note also the direction of REL - the relationship which the SECOND concept (with Concept Unique Identifier CUI2) HAS TO the FIRST concept (with Concept Unique Identifier CUI1).

Col.	Description
CUI1	Unique identifier of first concept
REL	Relationship of SECOND to first concept
CUI2	Unique identifier of second concept
RELA	Relationship attribute
SAB	Abbreviated source name (SAB) of the source of relationship. Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
SL	Source of relationship labels
MG	Machine-generated and unverified indicator (optional). G indicates 'machine generated'

Sample Records

C0002871|CHD|C0002891||MSH|MSH||

[Anemia, Neonatal (C0002891)

has CHILD REL and isa RELA

to Anemia (C0002871)]

C0002871|RB|C0221016||MTH|MTH||

[Red blood cell disorder, NOS (C0221016)

has broader REL

to Anemia (C0002871)]

C0002871|RL|C0002886|mapped_to|SNMI|SNMI||

[Anemia, Macrocytic (C0002886)

has like relationship

to Anemia (C0002871)]

C0002871|RQ|C0002886|clinically_associated_with|CCPSS|CCPSS||

[Megaloblastic anemia due to folate deficiency, NOS (C0151482)

has clinically_associated_with relationship

to Anemia (C0002871)]

4.3.10. Co-occurring Concepts (File = MRCOC - This file is no longer available in the UMLS after the 2013AA release.)

Note: Co-occurrence information is no longer available in the UMLS after the 2013AA release. Updated co-occurrences data are available in text files from the MEDLINE Co-Occurrences (MRCOC) page.

There are two rows in this table for each pair of concepts that co-occur in each information source represented one for each direction of the relationship. (Note that the COA data may be different for each direction of the relationship). Many Metathesaurus concepts have no entries in this file. Due to the very large number of co-occurrence relationships, they are distributed in a separate file.

Col.	Description
CUI1	Unique identifier of first concept
CUI2	Unique identifier of second concept Note: Where COT is MeSH topical qualifier (LQ) and CUI2 is not present, the count of citations of CUI1 with no MeSH qualifiers is reported.
SOC	Abbreviation of the source of co-occurrence information if applicable
COT	Type of co-occurrence
COF	Frequency of co-occurrence, if applicable
COA	Attributes of co-occurrence, if applicable

Sample Records

C0002871|C0000530|MED|L|1|BL=1,DT=1,ET=1|

C0002871|C0000545|MBD|L|1|BL=1,CI=1,DT=1|

C0002871|C0000589|MBD|L|1|CI=1,PC=1|

C0002871|C0000726|MED|L|1|CO=1|

C0002871|C0000727|MBD|L|1|CO=1,DI=1,TH=1|

Co-occurrences are concepts that occur together in the same entries in some information source. The relationships represented here are obtained from machine-manipulation of the information source. Co-occurrence relationships may exist between similar concepts (e.g., Atrial Fibrillation and Arrhythmia) or between very different concepts that nevertheless have some important connection in the field of biomedicine (e.g., Atrial Fibrillation and Digoxin), or between a primary concept and a qualifier (e.g., Lithotripsy and instrumentation). A co-occurrence relationship can exist between two concepts that have no other apparent relationship, although the frequency of such co-occurrences will be small.

In the current Metathesaurus, there are three sources of co-occurrence data: MEDLINE, AI/RHEUM, and CCPSS. From MEDLINE, co-occurrence data was computed for concepts that were designated as principal or main points in the same journal article i.e., the co-occurrence counts do not include articles in which either or both of the concepts were present and indexed in MEDLINE but not designated as main points. (A concept is considered to be a main point if the * is attached to the main heading or any of its subheadings.)

Two overall frequencies of MEDLINE co-occurrence are provided: one for recent MEDLINE data (MED) and one for MEDLINE data from a preceeding block of years (MBD). Separate counts are provided for the frequencies with which the first concept was qualified by different MeSH qualifiers or by no qualifier at all when it co-occurred with the second concept. There are separate entries for each direction of the co-occurrence relationship. The related subheading occurrence information in each entry belongs to the first concept in the entry and is therefore different for each direction of the relationship.

In addition to the specific qualifier information associated with two co-occurring concepts, in entries with LQ and LQB values for type of co-occurrence, this element also includes totals for the number of times each main concept was qualified by a specific subheading or by no subheading.

The AI/RHEUM co-occurrence data represent the co-occurrence of diseases and findings in the AI/RHEUM knowledge base, i.e., the diseases that co-occur with a particular finding and the findings that co-occur with a particular disease. Each disease/finding pair can co-occur only once in the AI/RHEUM knowledge base.

In CCPSS, the co-occurrence data is extracted from patient records and includes problem-problem co-occurrences within a patient record as well as problem-modifier co-occurrences.

4.3.11. Concept contexts (File = MRCXT)

This file is no longer distributed. To create the MRCXT file (Table 1), use the new MRCXT Builder application, accessible from the MetamorphoSys Welcome screen. Information on the MRCXT Builder can be found at http://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/MRCXT_Builder.html. The information below describes the content of the file when produced by the MRCXT Builder.

There are rows in this file for each occurrence of a concept in a hierarchy in any of the UMLS source vocabularies - a "context" in this discussion. Many Metathesaurus concepts have multiple contexts while others may have none. The number of rows per context differs depending on the number of ancestor, sibling, or child terms the concept has in that context. Because some concepts have multiple contexts in the same source (e.g., MeSH), a context number (CXN - e.g., 1, 2, 3) is used to identify all members of the same context. The CXNs are not global but are created as required for each concept. Since some concepts have multiple contexts in the same vocabulary with the same SUI, each distinct context can be retrieved with a CUI-SUI-SAB-CXN key. The "distance-1 relationships" i.e., the immediate parent, immediate child, and sibling relationships, represented in this file are also present in the MRREL file.

Sample Records

C0002871|S0013742|MSH|D000740|1|ANC|1|MeSH|C0220876||||
C0002871|S0013742|MSH|D000740|1|ANC|2|Diseases (MeSH Category)|C0012674|C|||
C0002871|S0013742|MSH|D000740|1|ANC|3|Hemic and Lymphatic Diseases|C0018981|C15|||
C0002871|S0013742|MSH|D000740|1|ANC|4|Hematologic Diseases|C0018939|C15.378|isa||
C0002871|S0013742|MSH|D000740|1|CCP||Anemia|C0002871|C15.378.71|isa|+|
C0002871|S0013742|MSH|D000740|1|CHD||Anemia, Aplastic|C0002874|C15.378.71.85|isa|+|
C0002871|S0013742|MSH|D000740|1|SIB||Blood Protein Disorders|C0005830|C15.378.147|isa|+|
C0002871|S0013742|MSH|D000740|1|CHD||Anemia, Hemolytic|C0002878|C15.378.71.141|isa|+|

Table 1. Concept contexts (File = MRCXT)

Col.	Description
CUI	Unique identifier of concept
SUI	Unique identifier of string used in this context
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
CODE	Unique identifier or code for string in that source.
CXN	The context number (to distinguish multiple contexts in the same source with the same SUI)
CXL	Context member label, i.e., ANC for ancestor of this concept, CCP for concept, SIB for sibling of this concept, CHD for child of this concept
RNK	For rows with a CXL value of ANC, the rank of the ancestors (e.g., a value of 1 denotes the most remote ancestor in the hierarchy)
CXS	String for context member
CUI2	Unique concept identifier of context member (may be empty if context member is not yet in the Metathesaurus)
HCD	Hierarchical number or code of context member in this source (optional)
RELA	Relationship attribute providing further categorization of the CXL, if applicable and known. Allowed values are listed on the Abbreviations Used in Data Elements page.
XC	A plus (+) sign indicates that the CUI2 for this row has children in this context. If this field is empty, the CUI2 does not have children in this context.

4.3.12. Associated Expressions (File = MRATX)

There is one row in this table for each vocabulary expression (i.e., combination of terms from a specific Metathesaurus source vocabulary) identified as having a relationship to a concept in the Metathesaurus. The majority of Metathesaurus entries have no entries in this table.

Col.	Description
CUI	Unique identifier of concept to which the expression is related

Table continued from previous page.

Col.	Description
SAB	Abbreviated source name (SAB) of source of terms in expression. Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
REL	Relationship of meaning of expression to main concept
ATX	Associated expression

Sample Records

C0001207|MSH|SY|<Acromegaly> AND <Gigantism>|

C0001296|LCH|RU|<Insurance>/<Statistics>|

C0001360|MSH|SY|<Thyroiditis> AND <Acute Disease>|

4.3.13. Source Information (File = MRSAB)

The Metathesaurus has "versionless" or "root" Source Abbreviations (SABs) in the data files. MRSAB (Table 2) connects the root SAB to fully specified version information for the current release. For example, the released SAB for MeSH is now simply "MSH". In MRSAB, you will find the current versioned SAB, e.g., MSH2003_2002_10_24. MetamorphoSys can produce files with either the root or versioned SABs so that either form can be utilized by a user.

There is one row in this file for every version of every source in the current Metathesaurus; when complete, there will also be historical information with a row for each version of each source that has appeared in any Metathesaurus release. Note that the field CURVER has the value Y to identify the version in this Metathesaurus release. Future releases of MRSAB will also contain historical version information in rows with CURVER value N.

MRSAB allows all other Metathesaurus files to use versionless source abbreviations, so that rows with no data change between versions also remain unchanged.

Sources with contexts have "full" contexts, i.e., all levels of terms may have Ancestors, Parents, Children and Siblings. A full context may also be further designated as Multiple, Nosib (No siblings) or both Multiple and Nosib.

Multiple indicates that a single concept in this source may have multiple hierarchical positions.

No siblings (Nosib) indicates that siblings have not been computed for this source.

The UMLS Source Vocabulary Documentation page of the current release documentation lists each source in the Metathesaurus and includes information about the type of context, if any, for each source.

Sample Record

C2930057|C1140284|RXNORM_10AA_100907F|RXNORM|RxNorm Vocabulary, 10AA_100907F|RXNORM|10AA_100907F|||2010AB||Stuart Nelson, M.D. ;Head, MeSH Section;National Library of Medicine;8600 Rockville Pike;Bethesda;Maryland;United States;20894;nelson@nlm.nih.gov|Stuart Nelson, M.D.;Head, MeSH

Section;National Library of Medicine;8600 Rockville Pike;Bethesda;Maryland;United States;
20894;nelson@nlm.nih.gov|0|437305|193737||
BN,BPCK,DF,ET,GPCK,IN,MIN,OCD,PIN,SBD,SBDC,SBDF,SCD,SCDC,SCDF, SY|
AMBIGUITY_FLAG,NDC,ORIG_AMBIGUITY_FLAG,ORIG_CODE,ORIG_SOURCE,ORIG_TTY,ORIG_VSAB,RXAUI,
RXCUI,RXN_ACTIVATED,RXN_BN_CARDINALITY,RXN_HUMAN_DRUG,RXN_IN_EXPRESSED_FLAG,
RXN_OBSOLETED, RXN_QUANTITY,RXN_STRENGTH,RXN_VET_DRUG,UNII_CODE|ENG|UTF-8|Y|Y|

Table 2. Source Information (File = MRSAB)

Field	Full Name	Description
VCUI	CUI	CUI of the versioned SRC concept for a source
RCUI	Root CUI	CUI of the root SRC concept for a source
VSAB	Versioned Source Abbreviation	The versioned source abbreviation for a source e.g., MSH2003_2002_10_24
RSAB	Root Source Abbreviation	The root source abbreviation for a source e.g., MSH
SON	Official Name	The official name for a source
SF	Source Family	The source family for a source
SVER	Version	The source version e.g., 2001
VSTART	Valid Start Date For A Source	The date a source became active, e.g., 2004_04_03
VEND	Valid End Date For A Source	The date a source ceased to be active, e.g., 2003_05_10
IMETA	Meta Insert Version	The version of the Metathesaurus in which a source first appeared, e.g., 2001AB
RMETA	Meta Remove Version	The version of the Metathesaurus in which a source last appeared, e.g., 2001AC
SLC	Source License Contact	The source license contact information
SCC	Source Content Contact	The source content contact information
SRL	Source Restriction Level	0, 1, 2, 3, 4, 9 – explained in the License Agreement
TFR	Term Frequency	The number of terms for this source in MRCON/MRSO, e.g., 12343
CFR	CUI Frequency	The number of CUIs associated with this source, e.g., 10234
CXTY	Context Type	The type of contexts for this source. Values are FULL, FULL-MULTIPLE, FULL-NOSIB, FULL-NOSIB-MULTIPLE, FULL-MULTIPLE-NOSIB-RELA, null.
TTYL	Term Type List	Term type list from source, e.g., MH, EN, PM, TQ
ATNL	Attribute Name List	The attribute name list (from MRSAT), e.g., MUI, RN, TH
LAT	Language	The language of the source
CENC	Character Encoding	All UMLS content is provided in Unicode, encoded in UTF-8. MetamorphoSys will allow exclusion of extended characters with some loss of information. Transliteration to other character encodings is possible but not supported buy NLM; for further information, see http://www.unicode.org.
CURVER	Current Version	A Y or N flag indicating whether or not this row corresponds to the current version of the named source
SABIN	Source in Subset	A Y or N flag indicating whether or not this row is represented in the current MetamorphoSys subset. Initially always Y where CURVER is Y, but later is recomputed by MetamorphoSys.

4.3.14. Concept Name Ranking (File = MRRANK)

There is exactly one row for each concept name type from each Metathesaurus source vocabulary (each SAB-TTY combination). The RANK and SUPPRESS values in the distributed file are those used in Metathesaurus production. Users are free to change these values to suit their needs and preferences, then change the naming precedence and suppressibility (TS in MRCON) by using MetamorphoSys to create a customized Metathesaurus.

Col.	Description
RANK	Numeric order of precedence, higher value wins
SAB	Abbreviated source name (SAB). Maximum field length is 20 alphanumeric characters. Two source abbreviations are assigned: • Root Source Abbreviation (RSAB) — short form, no version information, for example, AI/RHEUM, 1993, has an RSAB of "AIR" • Versioned Source Abbreviation (VSAB) — includes version information, for example, AI/RHEUM, 1993, has an VSAB of "AIR93" Official source names, RSABs, and VSABs are included on the UMLS Source Vocabulary Documentation page.
TTY	Abbreviation for term type in source vocabulary, for example PN (Metathesaurus Preferred Name) or CD (Clinical Drug). Possible values are listed on the Abbreviations Used in Data Elements page.
SUPPRESS	Flag indicating that this SAB and TTY will create a TS=s MRCON entry; see TS

Sample Records

0624|AIR|SY|N|

0623|ULT|PT|N|

0622|CPT|PT|N|

4.3.15. Ambiguous Term Identifiers (File = AMBIG.LUI)

In the instance that a Lexical Unique Identifier (LUI) is linked to multiple Concept Unique Identifiers (CUIs), there is one row in this table for each LUI-CUIs pair. This file identifies those lexical variant classes which have multiple meanings in the Metathesaurus.

In the Metathesaurus, the LUI links all strings within the English language that are identified as lexical variants of each other by the luinorm program found in the UMLS SPECIALIST Lexicon and Lexical Tools. LUIs are assigned irrespective of the meaning of each string. This table may be useful to system developers who wish to make use of the lexical programs in their applications to identify and disambiguate ambiguous terms.

Col.	Description
LUI	Lexical Unique Identifier
CUI	Concept Unique Identifier

Sample Records

L0000003|C0010504|

L0000003|C0917995|

L0000032|C0010206|

4.3.16. Ambiguous String Identifiers (File = AMBIG.SUI)

In the instance that a String Unique Identifier (SUI) is linked to multiple Concept Unique Identifiers (CUIs), there is one row in this table for each SUI-CUIs pair.

This file resides in the META directory. In the Metathesaurus, there is only one SUI for each unique string within each language, even if the string has multiple meanings. This table is only of interest to system developers who make use of the SUI in their applications or in local data files.

Col.	Description
SUI	String Unique Identifier
CUI	Concept Unique Identifier

Sample Records

S0063890|C0026667|

S0063890|C1135584|

S5147722|C1261047|

4.3.17. Metathesaurus Change Files

There are six files or relations that identify key differences between entries in the previous and the current edition of the Metathesaurus. Developers can use these special files to determine whether there have been changes that affect their applications.

The usefulness of individual files will depend on how data from the Metathesaurus have been linked or incorporated in a particular application.

Each relation or named table of data has a fixed number of columns and variable number of rows. A column is a sequence of all the values in a given data element. A row contains the values for two or more data elements for one entry. The values for the different data elements in the row are separated by vertical bars (|). Each row ends with a vertical bar and line termination.

4.3.17.1. Deleted Concepts (File = DELETED.CUI)

Concepts whose meaning is no longer present in the Metathesaurus are reported in this file. There is a row for each concept that existed in the previous release and is not present in the current release. If the meaning exists in the current release, i.e., the missing concept was merged with another current concept, it is reported in the MERGEDCUI file (Section 4.3.17.2) and not in this file.

Col.	Description
CUI	Concept unique identifier in the previous Metathesaurus
STR	Preferred name of this concept in the previous Metathesaurus

4.3.17.2. Merged Concepts (File = MERGED.CUI)

There is exactly one row in this table for each released concept in the previous Metathesaurus (CUI1) that was merged into another released concept from the previous Metathesaurus (CUI2). When this merge occurs, the first CUI (CUI1) was retired; this table shows the CUI (CUI2) for the merged concept in this Metathesaurus.

Entries in this file represent concepts pairs that were considered to have different meanings in the previous edition, but which are now identified as synonyms

Col.	Description
CUI1	Concept unique identifier in the previous Metathesaurus
CUI2	Concept unique identifier in this Metathesaurus in format C#######

4.3.17.3. Deleted Terms (File = DELETED.LUI)

There is exactly one row in this table for each Lexical Unique Identifier (LUI) that appeared in the previous version of the Metathesaurus, but does not appear in this version.

LUIs are assigned by the luinorm program, part of the lvg program in the UMLS SPECIALIST Lexicon and Lexical Tools.

These entries represent the cases where LUIs identified by the previous release's luinorm program, when used to identify lexical variants in the previous Metathesaurus, are no longer found with this release's luinorm on this release's Metathesaurus. This does not necessarily imply the deletion of a string or a concept from the Metathesaurus.

Col.	Description
LUI	Concept unique identifier in the previous Metathesaurus
STR	Preferred name of Term in the previous Metathesaurus

4.3.17.4. Merged Terms (File = MERGED.LUI)

There is exactly one row in this file for each case in which strings had different LUIs in the previous Metathesaurus yet share the same LUI in this Metathesaurus; a LUI present in the previous Metathesaurus is therefore absent from this Metathesaurus.

LUIs are assigned by the luinorm program, part of the lvg program in the UMLS SPECIALIST Lexicon and Lexical Tools.

These entries represent the cases where separate lexical variants as identified by the previous release's luinorm program version are a single lexical variant as identified by this release's luinorm.

Col.	Description
LUI1	Lexical unique identifier in the previous Metathesaurus but not present in this Metathesaurus
LUI2	Lexical unique identifier into which it was merged in this Metathesaurus

4.3.17.5. Deleted Strings (File = DELETED.SUI)

There is exactly one row in this file for each string in each language that was present in an entry in the previous Metathesaurus and does not appear in this Metathesaurus.

Note that this does not necessarily imply the deletion of a term (LUI) or a concept (CUI) from the Metathesaurus. A string deleted in one language may still appear in the Metathesaurus in another language.

Col.	Description
SUI	String unique identifier in the previous Metathesaurus that is not present in this Metathesaurus
LAT	Three-character abbreviation of language of string that has been deleted
STR	Preferred name of term in the previous Metathesaurus that is not present in this Metathesaurus

4.3.17.6. Retired CUI Mapping (File = MRCUI)

There are one or more rows in this file for each Concept Unique Identifier (CUI) that existed in any prior release but is not present in the current release. The file includes mappings to current CUIs as synonymous or to one or more related current CUI where possible. If a synonymous mapping cannot be found, other relationships between the CUIs can be created. These relationships can be Broader (RB), Narrower (RN), Other Related (RO), Deleted (DEL) or Removed from Subset (SUBX). Rows with the SUBX relationship are added to MRCUI by MetamorphoSys for each CUI that met the exclusion criteria and was consequently removed from the subset. Some CUIs may be mapped to more than one other CUI using these relationships.

CUIs may be retired when (1) two released concepts are found to be synonyms and so are merged, retiring one CUI; (2) the concept no longer appears in any source vocabulary and is not 'rescued' by NLM; or (3) the concept is an acknowledged error in a source vocabulary or determined to be a Metathesaurus production error.

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

Col.	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, DEL, RO, RN, RB.
CUI2	The current CUI that CUI1 most closely maps to
MAPIN	Is this map in current subset? Values of Y, N, or null. MetamorphoSys generates the Y or N to indicate whether the CUI2 concept is or is not present in the subset. The null value is for rows where the CUI1 was not present to begin with (i.e., REL=DEL).

Sample Records

C0612278|2001AC|SY|C0612279|Y|

C1146475|2004AA|DEL|||

C2741204|2010AA|RB|C1348543|Y|

C2741243|2010AA|DEL|||

C2741244|2010AA|RO|C1616644|Y|

4.3.18. Word Index (File = MRXW.BAQ, MRXW.DAN, MRXW.DUT, MRXW.ENG, MRXW.FIN, MRXW.FRE, MRXW.GER, MRXW.HEB, MRXW.HUN, MRXW.ITA, MRXW.NOR, MRXW.POR, MRXW.RUS, MRXW.SPA, MRXW.SWE)

There is one row in these tables for each word found in each unique Metathesaurus string (ignoring upper-lower case). All Metathesaurus entries have entries in the word index. The entries are sorted in ASCII order.

Col.	Description
LAT	Abbreviation of language of the string in which the word appears
WD	Word in lowercase
CUI	Concept identifier

Table continued from previous page.

Col.	Description
LUI	Term identifier
SUI	String identifier

Sample Records from MRXW.ENG

ENG|anaemia|C0002871|L0002871|S0352688|

ENG|anemia|C0002871|L0002871|S0013742|

ENG|disorder|C0002871|L2818006|S3448137|

ENG|nos|C0002871|L0002871|S0470050|

ENG|unspecified|C0002871|L0503461|S0589617|

Sample Records from MRXW.FRE

FRE|ANEMIE|C0002871|L0162748|S0227229|

4.3.19. Normalized Word Index (File = MRXNW.ENG)

There is one row in this table for each normalized word found in each unique English-language Metathesaurus string. All English-language Metathesaurus entries have entries in the normalized word index. There are no normalized string indexes for other languages in this edition of the Metathesaurus.

Col.	Description
LAT	Abbreviation of language of the string in which the word appears (always ENG in this edition of the Metathesaurus)
NWD	Normalized word in lowercase (described in Section 2.7.2.1)
CUI	Concept identifier
LUI	Term identifier
SUI	String identifier

Sample Records

ENG|anemia|C0002871|L0002871|S0013742|

ENG|anemia|C0002871|L0002871|S0013787|

ENG|disorder|C0002871|L2818006|S3448137|

ENG|unspecified|C0002871|L0503461|S0589617|

4.3.20. Normalized String Index (File = MRXNS.ENG)

There is one row in this table for each normalized string found in each unique English-language Metathesaurus string (ignoring upper-lower case). All English-language Metathesaurus entries have entries in the normalized string index. There are no normalized word indexes for other languages in this edition of the Metathesaurus.

Col.	Description
LAT	Abbreviation of language of the string (always ENG in this edition of the Metathesaurus)
NSTR	Normalized string in lowercase (described in Section 2.7.3.1)

Table continued from previous page.

Col.	Description
CUI	Concept identifier
LUI	Term identifier
SUI	String identifier

Sample Records

ENG|anemia disorder|C0002871|L2822821|S3436848|

ENG|anemia unspecified|C0002871|L0503461|S0589617|

ENG|anemia|C0002871|L0002871|S0013742|

5. Semantic Network

The Semantic Network consists of (1) a set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and (2) a set of useful and important relationships, or Semantic Relations, that exist between Semantic Types. This section of the documentation provides an overview of the Semantic Network, and describes the files of the Semantic Network. Sample records illustrate structure and content of these files.

The Semantic Network is distributed as one of the UMLS Knowledge Sources and as an open source resource available on the Semantic Network Web site.

5.1. Overview

The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus and to provide a set of useful relationships between these concepts. All information about specific concepts is found in the Metathesaurus. The Network provides information about the set of basic semantic types, or categories, which may be assigned to these concepts, and it defines the set of relationships that may hold between the semantic types. The Semantic Network contains 127 semantic types and 54 relationships. The Semantic Network serves as an authority for the semantic types that are assigned to concepts in the Metathesaurus. The Network defines these types, both with textual descriptions and by means of the information inherent in its hierarchies.

The semantic types are the nodes in the Network, and the relationships between them are the links. There are major groupings of semantic types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. The current scope of the UMLS semantic types is quite broad, allowing for the semantic categorization of a wide range of terminology in multiple domains.

The Metathesaurus consists of terms from its source vocabularies. The meaning of each term is defined by its source, explicitly by definition or annotation; by context (its place in a hierarchy); by synonyms and other stated relationships between terms; and by its usage in description, classification, or indexing. Each Metathesaurus concept is assigned at least one semantic type. In all cases, the most specific semantic type available in the hierarchy is assigned to the concept. For example, the concept "Macaca" receives the semantic type "Mammal" because there is not a more specific type "Primate" available in the Network. The level of granularity varies across the Network. This has important implications for interpreting the meaning (i.e., semantic type) that has been assigned to a Metathesaurus concept. For example, a sub-tree under the node "Physical Object" is "Manufactured Object". It has only two child nodes, "Medical Device" and "Research Device". It is clear that there are manufactured objects other than medical devices and research devices. Rather than proliferate the number of semantic types to encompass multiple additional subcategories for these objects, concepts that are neither medical devices nor research devices are simply assigned the more general semantic type "Manufactured Object".

Figure 1 illustrates a portion of the Network. The semantic type "Biologic Function" has two children, "Physiologic Function" and "Pathologic Function", and each of these in turn has several children and grandchildren. Each child in the hierarchy is linked to its parent by the "isa" link.

The primary link in the Network is the "isa" link. This establishes the hierarchy of types within the Network and is used for deciding on the most specific semantic type available for assignment to a Metathesaurus concept. In addition, a set of non-hierarchical relations between the types has been identified. These are grouped into five major categories, which are themselves relations: "physically related to", "spatially related to", "temporally related to", "functionally related to", and "conceptually related to".

Figure 2 illustrates a portion of the hierarchy for Network relationships. The "affects" relationship, one of several functional relationships, has six children, including "manages", "treats", and "prevents".

The relations are stated between high level semantic types in the Network whenever possible and are generally inherited via the "isa" link by all the children of those types. Thus, for example, the relation "process of" is stated to hold between the semantic types "Biologic Function" and "Organism". Therefore, it also holds between "Organ or Tissue Function" (which is a "Physiologic Function", which is, in turn, a "Biologic Function") and "Animal" (which is an "Organism"). The relations are stated between semantic types and do not necessarily apply to all instances of concepts that have been assigned to those semantic types. That is, the relation may or may not hold between any particular pair of concepts. So, though the relation "evaluation of" holds between the semantic types "Sign" and "Organism Attribute", a particular sign or a particular attribute may not be linked by this relation. Thus, signs such as "overweight" and "fever" are evaluations of the organism attributes "body weight" and "body temperature", respectively. However, "overweight" is not an evaluation of "body temperature", and "fever" is not an evaluation of "body weight".

In some cases there will be a conflict between the placement of types in the Network and the link to be inherited. If so, the inheritance of the link is said to be blocked. For example, by inheritance, the type "Mental Process" would be "process of" "Plant". Since plants are not sentient beings, this link is explicitly blocked. In other cases the nature of the relation is such that it should not be inherited by the children of the types that it links. In that case, the relation is defined for the two semantic types it explicitly links, but blocked for all the children of those types. For example, "conceptual part of" links "Body System" and "Fully Formed Anatomical Structure", but it should not link "Body System" to all the children of "Fully Formed Anatomical Structure", such as "Cell" or "Tissue".

Several portions of the MeSH hierarchy have been labeled with child to parent semantic relationships. All of the anatomy, diseases, and psychiatry and psychology sections have been labeled, as well as a portion of the biological sciences section. The links that are expressed between MeSH terms are, with a few exceptions, reflected in the Semantic Network. That is, if two MeSH terms are linked by a certain relation, then that link is expressed in the Network as a link between the semantic types that have been assigned to those MeSH terms. For example, "Amniotic Fluid", which is a "Body Substance", is a child of "Embryo", which is an "Embryonic Structure". The labeled relationship between "Amniotic Fluid" and its parent "Embryo" is "surrounds". This is allowable, since the relation "Body Substance surrounds Embryonic Structure" is represented in the Network.

Figure 3 shows a portion of the Semantic Network, illustrating the relations, either hierarchical or associative, that exist between semantic types.

The UMLS Semantic Network is provided in two formats: a relational table format and a unit record format.

5.2. Semantic Network ASCII Relational Format

There are two basic tables, two ancillary tables, and two bookkeeping tables included in this format. The two basic tables contain exactly the same information as the unit record file, but the information is presented differently. One table contains definitional information about the semantic types and relations; the other contains information about the structure of the Network. Each semantic type and each relation has been assigned a four character unique identifier (UI). These are of the form "T001", "T002", etc. The ancillary tables are expansions of the table that contains the Network structure. They give the fully inherited set of links represented in the Network. The first table is expressed as triples of UI's. The second is expressed as triples of names. The two bookkeeping tables describe the relational files and their fields. Fields in all tables are separated by a "|". All tables are listed and described below:

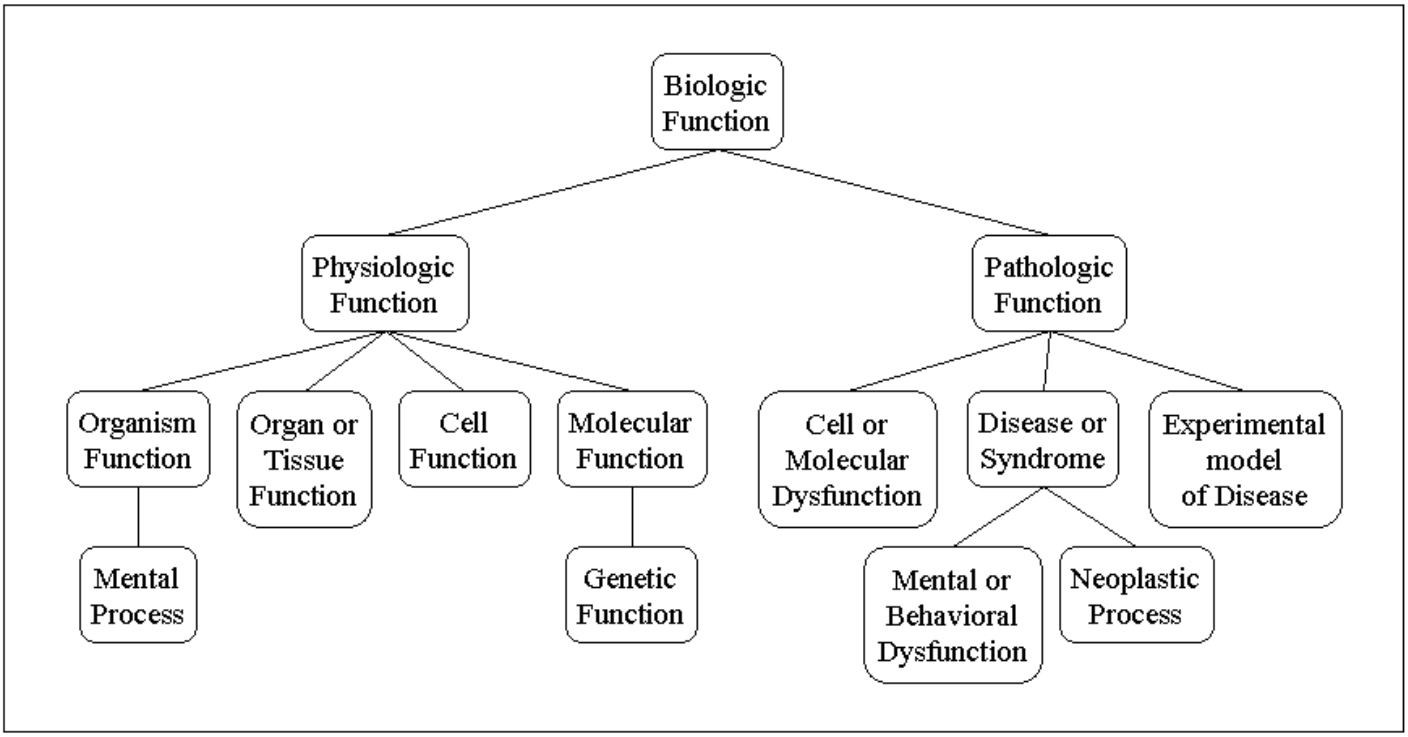

Figure 1. A Portion of the UMLS Semantic Network: "Biologic Function" Hierarchy

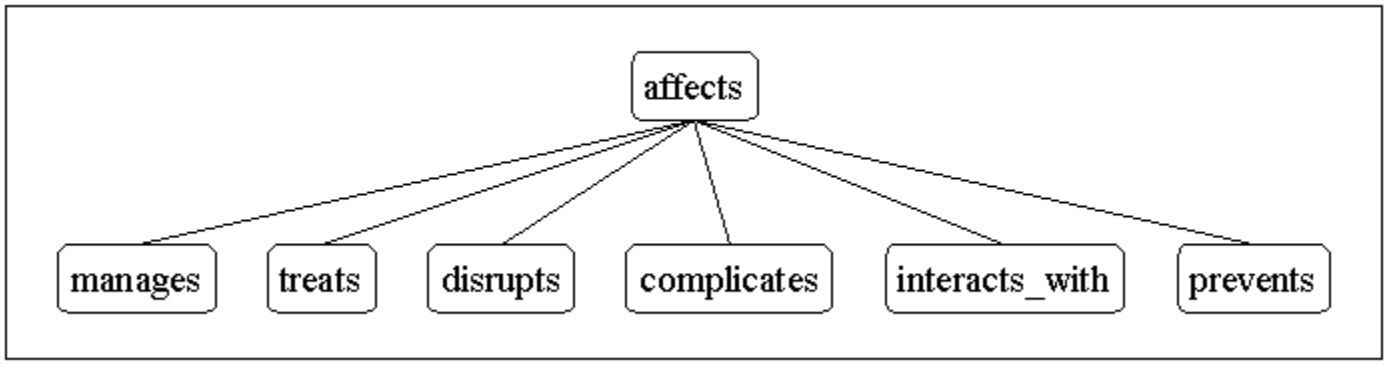

Figure 2. A Portion of the UMLS Semantic Network: "affects" Hierarchy

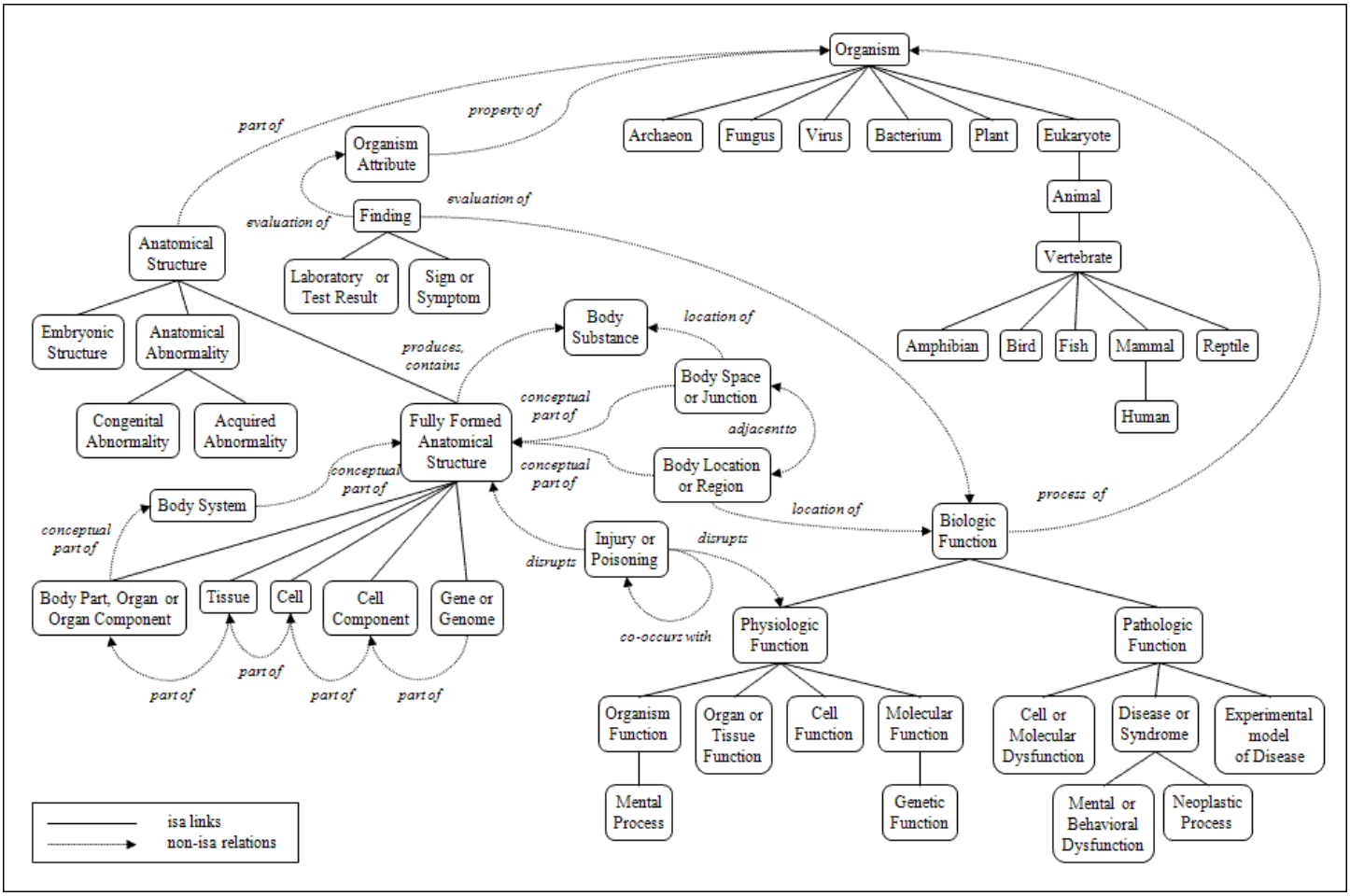

Figure 3. A Portion of the UMLS Semantic Network: Relations

Table	Description
SRDEF	Basic information about the Semantic Types and Relations.
SRSTR	Structure of the Network.
SRSTRE1	Fully inherited set of Relations (UI's).
SRSTRE2	Fully inherited set of Relations (names).
SRFIL	Description of each table.
SRFLD	Description of each field and the table(s) in which it is found.

Specific Descriptions of each Table:

Table: SRDEF

Field	Description
RT:	Record Type (STY = Semantic Type or RL = Relation).
UI:	Unique Identifier of the Semantic Type or Relation.
STY/RL:	Name of the Semantic Type or Relation.
STN/RTN:	Tree Number of the Semantic Type or Relation.
DEF:	Definition of the Semantic Type or Relation.

Table continued from previous page.

Field	Description
EX:	Examples of Metathesaurus concepts with this Semantic Type (STY records only).
UN:	Usage note for Semantic Type assignment (STY records only).
NH:	The Semantic Type and its descendants allow the non-human flag (STY records only).
ABR:	Abbreviation of the Relation Name or Semantic Type.
RIN:	Inverse of the Relation (RL records only).

Table: SRSTR

Field	Description
STY/RL:	Argument 1 (Name of a Semantic Type or Relation).
RL:	Relation ("isa" or the name of a non-hierarchical Relation).
STY/RL:	Argument 2 (Name of a Semantic Type or Relation); if this field is blank this means that the Semantic Type or Relation is one of the top nodes of the Network.
LS:	Link Status (D = Defined for the Arguments and its children; B = Blocked; DNI = Defined but Not Inherited by the children of the Arguments). N.B.: The relations expressed in this table are binary relations and the arguments are ordered pairs. The relations are stated only for the top-most node of the "isa" hierarchy of the Semantic Types to which they may apply.

Table: SRSTRE1 or SRSTRE2

Field	Description
UI/STY:	Argument 1 (UI or name of a Semantic Type).
UI/RL:	Relation (UI or name of a nonhierarchical Relation).
UI/STY:	Argument 2 (UI or name of a Semantic Type). N.B.: The relations expressed in this table are binary relations and the arguments are ordered pairs. All relations have been fully inherited in this table.

Table: SRFIL

Field	Description
FIL:	File Name.
DES:	Description of the file.
FMT:	Format of the file (fields in a comma-separated list).
CLS:	Number of columns in the file.
RWS:	Number of rows in the file.
BTS:	Number of bytes in the file.

Table: SRFLD

Field	Description
COL:	Field name.
DES:	Description of the field.
REF:	Cross-reference to the documentation.
FIL:	File name(s) in which the field is found.

Sample Relational Records

::::::::::::::
SRDEF
::::::::::::::
STY|T020|Acquired Abnormality|A1.2.2.2|An abnormal structure, or one that is abnormal in size or location, found in or deriving from a previously normal structure. Acquired abnormalities are distinguished from diseases even though they may result in pathological functioning (e.g., "hernias incarcerate").||NULL||acab||

STY|T047|Disease or Syndrome|B2.2.1.2.1|A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.||Any specific disease or syndrome that is modified by such modifiers as "acute", "prolonged", etc. will also be assigned to this type. If an anatomic abnormality has a pathologic manifestation, then it will be given this type as well as a type from the 'Anatomical Abnormality' hierarchy, e.g., "Diabetic Cataract" will be double-typed for this reason.||dsyn||

STY|T052|Activity|B1|An operation or series of operations that an organism or machine carries out or participates in.||Few concepts will be assigned to this broad type. Wherever possible, one of the more specific types from this hierarchy will be chosen. For concepts assigned to this type, the focus of interest is on the activity. When the focus of interest is the individual or group that is carrying out the activity, then a type from the 'Behavior' hierarchy will be chosen. In general, concepts will not receive a type from both the 'Activity' and the 'Behavior' hierarchies.||acty||

STY|T059|Laboratory Procedure|B1.3.1.1|A procedure, method, or technique used to determine the composition, quantity, or concentration of a specimen, and which is carried out in a clinical laboratory. Included here are procedures which measure the times and rates of reactions.||NULL||lbpr||

RL|T173|adjacent_to|R2.2|Close to, near or abutting another physical unit with no other structure of the same kind intervening. This includes adjoins, abuts, is contiguous to, is juxtaposed, and is close to.||||AD|adjacent_to|

RL|T151|affects|R3.1|Produces a direct effect on. Implied is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.||||AF|affected_by|

::::::::::::::
SRSTR
::::::::::::::
Acquired Abnormality|co-occurs_with|Injury or Poisoning|
Acquired Abnormality|isa|Anatomical Abnormality|
Acquired Abnormality|result_of|Behavior|
Activity|isa|Event|
Age Group|isa|Group|

::::::::::::::
SRSTRE1
::::::::::::::
T020|T186|T190|
T020|T186|T017|
T020|T186|T072|
T052|T186|T051|

T052|T165|T090|
T052|T165|T091|
T100|T186|T096|
T100|T186|T077|
T100|T186|T071|

::::::::::::::
SRSTRE2
::::::::::::::
Acquired Abnormality|isa|Anatomical Abnormality|
Acquired Abnormality|isa|Anatomical Structure|
Acquired Abnormality|isa|Entity|
Acquired Abnormality|isa|Physical Object|
Acquired Abnormality|affects|Amphibian|
Acquired Abnormality|affects|Animal|
Acquired Abnormality|affects|Archaeon|
Acquired Abnormality|affects|Bacterium|
Acquired Abnormality|affects|Bird|
Acquired Abnormality|affects|Cell Function|
Acquired Abnormality|affects|Eukaryote|
Acquired Abnormality|affects|Fish|
Acquired Abnormality|affects|Fungus|
Acquired Abnormality|affects|Genetic Function|
Acquired Abnormality|affects|Human|
Acquired Abnormality|affects|Mammal|
Acquired Abnormality|affects|Mental Process|
Acquired Abnormality|affects|Molecular Function|
Acquired Abnormality|affects|Organ or Tissue Function|
Acquired Abnormality|affects|Organism Function|
Acquired Abnormality|affects|Organism|
Acquired Abnormality|affects|Physiologic Function|
Acquired Abnormality|affects|Plant|
Acquired Abnormality|affects|Reptile|
Acquired Abnormality|affects|Vertebrate|
Acquired Abnormality|affects|Virus|
Activity|isa|Event|
Age Group|isa|Conceptual Entity|
Age Group|isa|Entity|
Age Group|isa|Group|

5.3. Semantic Network ASCII Unit Record Format

The file "SU" contains individual records for both semantic types and relations.

Each record begins with a unique identifier field (UI) which contains the four character UI. These are of the form "T001", "T002", etc. Each field in a record begins on a new line and may continue over several lines. Some fields are optional.

Semantic Type records contain the following fields:

Field	Description
UI:	Unique Identifier of the Semantic Type.
STY:	Name of the Semantic Type.
STN:	Tree Number of the Semantic Type.
DEF:	Definition of the Semantic Type.
EX:	Examples of Metathesaurus concepts with this Semantic Type (optional field).
UN:	Usage note for Semantic Type assignment (optional field).
NH:	Semantic Type and its descendants allow the non-human flag (optional field).
HL:	Hierarchical links of the Semantic Type to its parent ({isa}) and its children ({inverse_isa}). If there are no hierarchical links, then the value <none> is assigned.

Relation records contain the following fields:

Field	Description
UI:	Unique Identifier of the Relation.
RL:	Name of the Relation.
ABR:	Abbreviation of the Relation.
RIN:	Name of the inverse of the Relation.
RTN:	Tree Number of the Relation.
DEF:	Definition of the Relation.
INH:	"N" if the relation is not inherited (optional field).
HL:	Hierarchical links of the Relation to its parent ({isa}) and its children ({inverse_isa}). If there are no hierarchical links, then the value <none> is assigned.
STL:	Semantic Types linked by this Relation. N.B.: These are binary relations and the arguments are ordered pairs. The relations are stated only for the top-most node of the "isa" hierarchy of the Semantic Types to which they may apply. This field does not appear in the "isa" relation record since its values can be computed from the "HL" field. If there are no semantic types linked by this Relation, then the value <none> is assigned.
STLB:	Semantic Types linked by this Relation are blocked (optional field).

Sample Unit Records

::::::::::::::
SU
::::::::::::::

UI:	T020
STY: ABR:	Acquired Abnormality acab
STN:	A1.2.2.2
DEF:	An abnormal structure, or one that is abnormal in size or location, found in or deriving from a previously normal structure. Acquired abnormalities are distinguished from diseases even though they may result in pathological functioning (e.g., "hernias incarcerate").

Table continued from previous page.

HL:	{isa} Anatomical Abnormality

UI:	T052
STY: ABR:	Activity acty
STN:	B1
DEF:	An operation or series of operations that an organism or machine carries out or participates in.
UN:	Few concepts will be assigned to this broad type. Wherever possible, one of the more specific types from this hierarchy will be chosen. For concepts assigned to this type, the focus of interest is on the activity. When the focus of interest is the individual or group that is carrying out the activity, then a type from the 'Behavior' hierarchy will be chosen. In general, concepts will not receive a type from both the 'Activity' and the 'Behavior' hierarchies.
HL:	{isa} Event; {inverse_isa} Behavior; {inverse_isa} Daily or Recreational Activity; {inverse_isa} Occupational Activity; {inverse_isa} Machine Activity

UI:	T100
STY: ABR:	Age Group aggp
STN:	A2.9.4
DEF:	An individual or individuals classified according to their age. EX: Adult; Infant, Premature; Adolescents; Aged, 80 and over
HL:	{isa} Group

UI:	T173
RL:	adjacent_to
ABR:	AD
RIN:	adjacent_to
RTN:	R2.2
DEF:	Close to, near or abutting another physical unit with no other structure of the same kind intervening. This includes adjoins, abuts, is contiguous to, is juxtaposed, and is close to.
HL:	{isa} spatially_related_to

Table continued from previous page.

STL:	[Body Location or Region|Body Location or Region]; [Body Location or Region|Body Part, Organ, or Organ Component]; [Body Location or Region|Body Space or Junction]; [Body Part, Organ, or Organ Component|Body Part, Organ, or Organ Component]; [Body Part, Organ, or Organ Component|Body Space or Junction]; [Body Part, Organ, or Organ Component|Cell]; [Body Part, Organ, or Organ Component|Tissue]; [Body Space or Junction|Body Space or Junction]; [Cell Component|Body Space or Junction]; [Cell Component|Cell Component]; [Cell|Cell]; [Tissue|Body Space or Junction]; [Tissue|Tissue]

UI:	T151
RL:	affects
ABR:	AF
RIN:	affected_by
RTN:	R3.1
DEF:	Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.
HL:	{isa} functionally_related_to; {inverse_isa} manages; {inverse_isa} treats; {inverse_isa} disrupts; {inverse_isa} complicates; {inverse_isa} interacts_with; {inverse_isa} prevents
STL:	[Natural Phenomenon or Process|Natural Phenomenon or Process]; [Anatomical Abnormality|Physiologic Function]; [Biologic Function|Organism]; [Anatomical Abnormality|Organism]; [Health Care Activity|Biologic Function]; [Diagnostic Procedure|Patient or Disabled Group]; [Therapeutic or Preventive Procedure|Patient or Disabled Group]; [Chemical|Natural Phenomenon or Process]; [Gene or Genome|Physiologic Function]; [Cell Component|Physiologic Function]; [Physiologic Function|Organism Attribute]; [Food|Biologic Function]; [Behavior|Behavior]; [Behavior|Mental Process]; [Mental Process|Behavior]; [Mental or Behavioral Dysfunction|Behavior]; [Research Activity|Mental Process]; [Regulation or Law|Group]; [Regulation or Law|Organization]

6. SPECIALIST Lexicon and Lexical Tools

The SPECIALIST Lexicon has been developed to provide the lexical information needed for the SPECIALIST Natural Language Processing System (NLP). It is intended to be a general English lexicon that includes many biomedical terms. Coverage includes both commonly occurring English words and biomedical vocabulary. The lexicon entry for each word or term records the syntactic, morphological, and orthographic information needed by the SPECIALIST NLP System.

The Lexical Tools are designed to address the high degree of variability in natural language words and terms. Words often have several inflected forms which would properly be considered instances of the same word. The verb "treat", for example, has three inflectional variants: "treats" the third person singular present tense form, "treated" the past and past participle form, and "treating" the present participle form. Multi-word terms in the Metathesaurus and other controlled vocabularies may have word order variants in addition to their inflectional and alphabetic case variants. The Lexical Tools allow the user to abstract away from this sort of variation.

For an overview of the SPECIALIST Lexicon, lexical variant programs, and lexical databases, see Lexical Methods for Managing variation in Biomedical Terminologies, A.T. McCray, S. Srinivasan, A.C. Browne, in the Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care, 1994, 235-239.

The SPECIALIST Lexicon is distributed as one of the UMLS Knowledge Sources and as an open source resource along with the the SPECIALIST NLP tools, subject to these terms and conditions.

6.1. General Description

The Lexicon consists of a set of lexical entries with one entry for each spelling or set of spelling variants in a particular part of speech. Lexical items may be "multi-word" terms made up of other words if the multi-word term is determined to be a lexical item by its presence as a term in general English or medical dictionaries, or in medical thesauri such as MeSH. Expansions of generally used acronyms and abbreviations are also allowed as multi-word terms.

The unit lexical record is a frame structure consisting of slots and fillers. Each lexical record has a base= slot whose filler indicates the base form, and optionally a set of spelling_variants= slots to indicate spelling variants. An "entry=" slot records the unique identifier (EUI) of the record. EUI numbers are seven digit numbers preceded by an "E". Each record has a cat= slot indicating part of speech. The lexical record is delimited by braces ({...}).

The unit lexical records for "anaesthetic" given below illustrate some of the features of the SPECIALIST lexical record:

{base=anesthetic spelling_variant=anaesthetic entry=E0354094 cat=noun variants=reg variants=uncount}
{base=anesthetic spelling_variant=anaesthetic entry=E0330019 cat=adj variants=inv position=attrib(3) position=pred stative}

The base form "anesthetic" and its spelling variant "anaesthetic" appear in two lexical records, one an adjective entry, the other a noun entry. The variants= slot contains a code indicating the inflectional morphology of the entry; the filler reg in the noun entry indicates that the noun "anesthetic" is a count noun which undergoes regular English plural formation ("anaesthetics"); inv in the variants= slot of the adjective entry indicates that the adjective "anesthetic" does not form a comparative or superlative. The position= slot indicates that the adjective "anaesthetic" is attributive and appears after color adjectives in the normal adjective order. "pred" in the position slot of the adjective entry indicates that this adjective can appear in predicate position.

Lexical entries are not divided into senses. Therefore, an entry represents a spelling-category pairing regardless of semantics. The noun "act" has two senses both which show a capitalized and lower case spelling; an act of a

play and an act of law. Since both senses share the same spellings and syntactic category, they are represented by a single lexical entry in the current lexicon. The unit record for "Act" is shown below.

{base=Act spelling_variant=act entry=E0000154 cat=noun variants=reg }

When different senses have different syntactic behavior, codes for each behavior are recorded in a single entry. For example, "beer" has two senses: the alcoholic beverage and the amount of a standard container of that beverage.

A Patients who drank beer recovered more slowly than patients who drank wine.
B. Fifty-six patients reported drinking more than five beers a day.

The first sense illustrated in A. above is a mass (uncount) noun. The second sense illustrated in B. is a regular (count) noun. In cases like this the appropriate codes for both senses are included in the entry.

{base=beer entry=E0012226 cat=noun variants=uncount variants=reg }

Two codes will also appear in cases where the lexical item is both count and uncount without a sense distinction. "Abdominal delivery" denotes the same procedure whether it appears as an uncount noun as in C. or a count noun as in D.

C. Abdominal delivery is the procedure of choice in this situation.
D. Abdominal deliveries are more common these days.

The unit lexical record for "abdominal delivery" includes both codes.

{base=abdominal delivery entry=E0006453 cat=noun variants=uncount variants=reg }

Other syntactic codes such as complement codes for verbs, adjectives and nouns are similarly grouped without regard to sense.

6.2. The Scope of the Lexicon

Words are selected for lexical coding from a variety of sources. Approximately 20,000 words from the UMLS Test Collection of MEDLINE abstracts together with words which appear both in the UMLS Metathesaurus and Dorland's Illustrated Medical Dictionary form the core of the words entered. In addition, an effort has been made to include words from the general English vocabulary. The 10,000 most frequent words listed in The American Heritage Word Frequency Book and the list of 2,000 words used in definitions in Longman's Dictionary of Contemporary English have also been coded. Since the majority of the words selected for coding are nouns, an effort has been made to include verbs and adjectives by identifying verbs in current MEDLINE citation records, by using the Computer Usable Oxford Advanced Learner's Dictionary, and by identifying potential adjectives from Dorland's Illustrated Medical Dictionary using heuristics developed by McCray and Srinivasan (1990).

A variety of reference sources are used in coding lexical records. Coding is based on actual usage in the UMLS Test Collection and MEDLINE, dictionaries of general English, primarily learner's dictionaries which record the kind of syntactic information needed for NLP, and medical dictionaries. Longman's Dictionary of Contemporary English, Dorland's Illustrated Medical Dictionary, Collins COBUILD Dictionary, The Oxford Advanced Learner's Dictionary, and Webster's Medical Desk Dictionary were used.

The SPECIALIST Lexicon also exists in relational format generated from the unit records. The full SPECIALIST Lexicon technical report entitled "The SPECIALIST Lexicon," found in the file techrpt.pdf, fully describes the unit record format. The remainder of the present chapter describes the relational form of the Lexicon. Section 6.3 describes the data elements that make up the relational tables and Section 6.4 describes the tables.

6.3. Lexicon Data Elements

Each of the elements below is represented as fields (columns) in the relational format.

6.3.1. String Properties

These data elements refer to properties of the strings generated by the entries.

6.3.1.1. STR - String

A Lexical entry generates a variety of forms (strings) including all the inflectional forms (the citation form, as well) of each spelling variant. Case, punctuation and spaces are considered significant.

6.3.1.2. AGR - Agreement/Inflection Code

This element encodes agreement and inflection information.

Agreement between nouns and verbs and between determiners and nouns involves person and number. Person and Number are indicated by the following codes.

Code	Person	Number
second	Second	Singular & Plural
third	Third	Singular & Plural
fst_sing	First	Singular
fst_plur	First	Plural
thr_sing	Third	Singular
thr_plur	Third	Plural

For Nouns, the agreement/inflection code indicates countability, person and number. Person and number are indicated by the person/number codes given above which are parenthesized after the countability code. Nouns can be either count or uncount.

For Pronouns, the agreement/inflection indicates person and number using the codes given above.

For verbs, including auxiliaries and modals, the agreement/inflection code indicates tense, person and number. Persons and numbers are indicated by the same person/number codes given above. These codes are parenthesized after the tense. No person number codes are given for non-finite tenses. "pres(thr_sing)" indicates third person singular present tense and "pres(fst_sing,fst_plur,thr_plur,second)" indicates present tense for all persons and numbers other than third singular. Negative forms of auxiliaries (didn't) and modals (can't) have "negative" after a colon at the end of the agreement/inflection code.

Code	Tense
past	Past Tense
pres	Present Tense
past_part	Past Participle
pres_part	Present Participle
infinitive	Infinitive

Determiners agree with nouns in terms of countability and number. The agreement/inflection codes for determiners are "free", "plur", "sing" and "uncount". "free" indicates that the determiner places no restrictions on

its noun. Determiners marked "plur" allow plural nouns, those marked "sing" allow singular nouns and those marked "uncount" allow uncountnouns.

6.3.1.3. CAS - Case

See Section 4.3.1 of "The SPECIALIST Lexicon" technical report.

Pronouns in English may be in one of two cases, subjective (nominative) or objective (accusative). This field contains "subj", "obj" or both separated by a comma to indicate the case of the pronoun.

6.3.1.4. GND - Gender

This field indicates the gender of pronouns.

Pronouns may be marked pers or neut to indicate whether they refer to people or non-people respectively. Pronouns marked pers may be masculine (masc) or feminine (fem) referring to male or female people respectively. See Section 14.2 of "The SPECIALIST Lexicon" technical report. There are four codes possible in this field:

Code	Gender
pers	person
neut	neuter
pers(masc)	person masculine
pers(fem)	person feminine

Notice that pers as used here does not correspond to the traditional term "personal pronoun". For example "it" and "they" are traditionally called personal pronouns since they both participate in the person/number paradigm. A pronoun like "none" is not traditionally called a personal pronoun.

6.3.2. Entry Properties

6.3.2.1. EUI - Unique Identifier Number for Lexical Entries

The EUI identifies a lexical entry. Information about a set of spelling variants in a particular part of speech is represented as an entry in the unit record. A particular string may be assigned several EUI numbers as it may occur in several parts of speech.

6.3.2.2. CIT - Citation Form

This field records the citation form of strings in the agreement/inflection table (Section 6.4.3.1 - lragr). The citation form is the singular for nouns, infinitive for verb and positive for adjectives and adverbs. The base form and the spelling variants if any are the citation forms of each of their respective inflections. This form is sometimes referred to as the un-inflected form.

6.3.2.3. BAS - Base Form

This field records the base form of a lexical entry. The base form is the citation form of one of a set of spelling variants chosen to represent the whole set. It might be thought of as the name of a lexical entry. The base form is the filler of the base= slot.

6.3.2.4. SCA - Syntactic Category

The syntactic category (part of speech) of the lexical entry. This field may be filled by one of the following. See Section 3 of "The SPECIALIST Lexicon" technical report.

Code	Category
noun	nouns
adj	adjectives
adv	adverbs
pron	pronouns
verb	verbs
det	determiners
prep	prepositions
conj	conjunctions
aux	auxiliaries
modal	modals
compl	complementizers

6.3.2.5. PER - Periphrastic

The code "periph" in this field indicates that an adjective or adverb is periphrastic. An adjective is periphrastic if it can form its comparative with "more" and its superlative with "most". See Section 4.3.5 of "The SPECIALIST Lexicon" technical report for discussion.

6.3.2.6. COM - Complements

These are complement codes. See Sections 5.1, 5.2, 5.4 and 5.5 in "The SPECIALIST Lexicon" technical report for a description of SPECIALIST complement codes.

6.3.2.7. TYP - Inflectional Type

The inflectional type(s) of an entry indicate the ways in which its forms may be inflected, or in the case of determiners the inflection of the heads they may determine. These codes are used to generate the variant strings (STR) found in other tables.

For nouns the following types may appear:

Code	Pluralization Pattern	See "The SPECIALIST Lexicon" Section
reg	regular	4.5.2
glreg	Greco-Latin regular	4.5.3
metareg	metalinguistic regular	4.5.4
irreg( )	irregular	4.5.5
sing	fixed singular	4.5.6
plur	fixed plural	4.5.7
inv	invariant	4.5.8
group(irreg( ))	group irregular	4.5.9
group(reg)	group regular	4.5.9
uncount	uncountable	4.5.10
groupuncount	group uncount	4.5.11

For verbs the following types may appear:

Code	Inflection Type	See "The SPECIALIST Lexicon" Section
reg	regular	4.1.1
regd	regular doubling	4.1.2
irreg( )	irregular	4.1.3

For pronouns the following types may appear:

Code	Inflection Type
fst_plur	first person plural
fst_sing	first person singular
sec_plur	second person plural
sec_sing	second person singular
second	second person
third	third person
thr_plur	third person plural
thr_sing	third person singular

See Section 14.1 of "The SPECIALIST Lexicon" technical report.

For adjectives and adverbs the following types can appear:

Code	Inflectional Type	See "The SPECIALIST Lexicon" Section
reg	regular	4.3.1 and 4.4.1
regd	regular doubling	4.3.2
inv	invariant	4.3.4 and 4.4.3
inv;periph	periphrastic	4.3.5 and 4.4.4
irreg( )	irregular	4.3.3 and 4.4.2

For determiners the inflection type indicates the inflection of the noun heads they may determine. The following types may appear:

Code	Inflectional Type	See "The SPECIALIST Lexicon" Section
sing	singular	4.7.1
plur	plural	4.7.2
uncount	uncount	4.7.3
singuncount	singular uncount	4.7.4
pluruncount	plural uncount	4.7.5
free	free	4.7.6

6.3.2.8. POS - Possession

English pronouns may be possessive or possessive nominal. The codes poss, possnom or both (comma separated) may appear in this field.

See Section 14.3.2 of "The SPECIALIST Lexicon" technical report.

6.3.2.9. QNT - Quantification

This field indicates the quantification properties inherent in certain pronouns. The four codes possible in this field are:

Code	Properties
univ	universal quantification
indef(nonassert)	non-assertive indefinite
indef(neg)	negative indefinite
indef(assert)	assertive indefinite

See Section 14.3.4 in "The SPECIALIST Lexicon" technical report for discussion of quantification in pronouns.

6.3.2.10. FEA - Features

This field represents various features of terms in various categories. The possible features are:

Feature	See "The SPECIALIST Lexicon" Section
reflexive	14.3.3
negative	14.3.4
demonstrative	14.3.5
interrogative	12.1
proper	8.
negative	13.1
broad_negative	13.2
stative	10.

PSN - Position for Adjectives

Adjectives are marked in the SPECIALIST Lexicon with position codes showing whether they are attributive postmodifying or predicative. If attributive, the code indicates where they appear in the pre-nominal sequence of adjectives. An additional attributive code, attribc, is used to indicate adjectives which can take complements in attributive position. One or more of the following codes can appear:

Code	Position	See "The SPECIALIST Lexicon" Section
attrib(1)	attributive (1st position)	9.1.1.1
attrib(2)	attributive (2nd position)	9.1.1.2
attrib(3)	attributive (3rd position)	9.1.1.3
attribc	attributive with complement	9.1.2
post	post modifying	9.2

Table continued from previous page.

Code	Position	See "The SPECIALIST Lexicon" Section
pred	predicative	9.3

6.3.2.12. MOD - Modification Type for Adverbs

Adverbs are marked in the SPECIALIST Lexicon to indicate their modification type. The possible values of this field are:

Code	See "The SPECIALIST Lexicon" Section
intensifier	11.2
particle	11.1
sentence_modifier; TYPE	11.3
verb_modifier; TYPE	11.4

TYPE is one of locative, temporal or manner. See Section 11.5 in "The SPECIALIST Lexicon" technical report.

6.3.2.13. GEN - Generic Name for a Trademark

The GEN field represents a generic or public name for the thing referred to by the trademark. The trademark "Alphalin" has the generic term "vitamin A".

6.3.3. Entry Relations

6.3.3.1. ABR - Acronym or Abbreviation

This field indicates whether a term listed in the acronym-abbreviation table (lrabr) is an acronym or abbreviation. It contains either:

"abbreviation_of" or "acronym_of".

6.3.3.2. SPV - Spelling Variant

A base form in the SPECIALIST Lexicon may have one or more spelling variants, subject to the same inflectional pattern. This field contains the citation form of a particular spelling variant. See Section 2 of "The SPECIALIST Lexicon" technical report.

6.3.4. Data Description

The data elements describe the relational table files or provide index entries into the Lexicon.

6.3.4.1. WRD - Word

Each string is broken into "words" and indexed in lrwd. Words are strings of alpha-numeric characters more than one character long, separated by space or punctuation.

6.3.4.2. DES - Description

A short definition of a file or field. This is free text.

6.3.4.3. FMT - Format

An ordered comma separated list of field names appearing in a file.

6.3.4.4. RWS - Number of Rows

The number of rows (lines or records) in a file.

6.3.4.5. FIL - File Name(s)

One or more file names denoting the files containing relational tables.

6.3.4.6. BTS - Size in Bytes

The size of a file in bytes (characters).

6.3.4.7. CLS - Number of Columns

The number of columns (fields) in a record (or row) of a table. The same number as the number of lines in the file.

6.3.4.8. COL - Three Letter Field Name

A three letter identifier for a field.

6.3.4.9. REF - Cross Reference to Document

A cross reference to a section of this document.

6.4. Lexicon Relational Tables

6.4.1. Introduction

In this format the data in each lexical entry is represented in ten different "relations" or "tables" each in a file.

The lexicon relational format is not fully normalized. By design, there is duplication of data among different relations and within certain relations. Developers will need to make their own decisions about the extent to which this redundancy should be retained, reduced, or increased for their specific applications.

6.4.2. General Description of the Relational Format

As in the Metathesaurus ASCII relational format, each relation or table of data values has by definition a fixed number of columns; the number of rows depends on the content of a particular version of the Lexicon. A column is a sequence of all the values in a given data element or logical sub-element. In general, columns for longer variable length data elements will appear to the right of columns for shorter and/or fixed length data elements. A row contains the values for one or more data elements or logical sub-elements for one Lexicon entry or string. Depending on the nature of the data elements involved, each Lexicon entry or string may have one or more rows in a given file. The values for the different data elements or logical sub-elements represented in the row are separated by vertical bars (|). If an optional element is blank, the vertical bars are still used to maintain the correct positioning of the subsequent elements. Each row is terminated by a vertical bar and a carriage return followed by a line feed. (|<CR><LF>).

6.4.3. Summary of the Contents of Each of the Relational Files

In the following descriptions, the numbers in parentheses beside each element refer to the section of this document that describes the element's contents.

6.4.3.1. - Agreement and Inflection (File = lragr)

Rows of the agreement table have six fields. There is a row in lragr for each inflected form of each spelling variant. This table links those forms to their citation forms and base forms. It provides information about agreement between subjects (nouns and pronouns) and verbs and between determiners and nouns.

EUI	The Entry Unique ID Number (6.3.2.1)
STR	String (6.3.1.1)
SCA	Syntactic Category (6.3.2.4)
AGR	Agreement/Inflection Code (6.3.1.2)
BAS	Base Form (6.3.2.3)
CIT	Citation Form (6.3.2.2)

6.4.3.2. - Inflection Type (File = lrtyp)

The lrtyp table has one or more rows for each lexical entry, indicating the inflectional pattern(s) to which it belongs.

EUI	The Entry Unique ID Number (6.3.2.1)
CIT	Citation Form (6.3.2.2)
SCA	Syntactic Category (6.3.2.4)
TYP	Inflectional Type (6.3.2.7)

6.4.3.3. - Complementation (File = lrcmp)

In lrcmp there is one line for each complement code for each entry.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
SCA	Syntactic Category (6.3.2.4)
COM	Complement Code. (6.3.2.6)

6.4.3.4. - Pronouns (File = lrprn)

lrprn has one or more rows for each pronoun entry in the Lexicon. Each row has nine columns.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
AGR	Agreement/Inflection Code (6.3.1.2)

See Section 14.1 in "The SPECIALIST Lexicon" technical report.

The agreement/inflection field in lrprn indicates person and number for anaphoric reference, AGR in lragr indicates person for agreement. These differ in the case of possessive nominal pronouns. The possessive nominal "mine" is "third" for purposes of subject verb agreement and "fst_sing" in its anaphoric reference.

GND	Gender (6.3.1.4)
CAS	Case (6.3.1.3)
POS	Possession (6.3.2.8)

Table continued from previous page.

QNT	Quantification (6.3.2.9)
FEA	Other Features (for pronouns) (6.3.2.10)

6.4.3.5. Modifiers (file = lrmod)

The modifier table includes position information for adjectives and modification type information for adverbs, and a variety of features.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
SCA	Syntactic Category (6.3.2.4)

All the entries represented in this table have the category "adj" or "adv" indicating adjectives or adverbs respectively. The fourth field of lrmod may be PSN or MOD depending on whether the term is an adjective or adverb.

PSN/MOD	Position (6.3.2.11) - for adjectives / Modification Types (6.3.2.12) - for adverbs
FEA	Features (6.3.2.10)

6.4.3.6. - Properties (file = lrprp)

lrprp indicates properties of terms in various categories.

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	The Base Form (6.3.2.3)
STR	String (6.3.1.1)

STR is only indicated in lrprp when a feature applies to a single string out of those generated by the entry, as in the negative contractions.

SCA	Syntactic Category (6.3.2.4)
FEA	Features (6.3.2.10)

6.4.3.7. - Abbreviations and Acronyms (file = lrabr)

This file links acronyms and abbreviations to their expansions.

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of the acronym or abbreviation.

BAS	The Base Form (6.3.2.3)

This field contains the Base form of the acronym or abbreviation.

ABR	Acronym or Abbreviation (6.3.3.1)
BAS	The Base Form (6.3.2.3)

This field contains the Base form of the expansion of the acronym or abbreviation.

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of the expansion of the abbreviation or acronym.

6.4.3.8. - Spelling Variants (file = lrspl)

EUI	The Entry Unique ID Number (6.3.2.1)
SPV	Spelling Variant (6.3.3.2)
BAS	The Base Form (6.3.2.3)

6.4.3.9. - Nominalizations (file = lrnom)

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of the nominalization.

BAS	The Base Form (6.3.2.3)

This field contains the Base form of the nominalization.

SCA	Syntactic Category (6.3.2.4)

This field contains the category of the nominalization (noun).

EUI	The Entry Unique ID Number (6.3.2.1)

This field contains the EUI of a verb or adjective of which the noun is a nominalization.

BAS	The Base Form (6.3.2.3)

This field contains the base form of the verb or adjective of which the noun is a nominalization.

SCA	Syntactic Category (6.3.2.4)

This field contains the syntactic category (adj or verb) of the adjective or verb.

6.4.3.10. - Trademarks (file = lrtrm)

EUI	The Entry Unique ID Number (6.3.2.1)
BAS	Base (6.3.2.3)
GEN	Generic Term (6.3.2.13)

The appearance of a form in the lrtrm table indicates that it is a trademark. It may or may not have a generic term associated with it.

6.4.3.11. - Files (file = lrfil)

The lrfil table describes each file in the ASCII relational form of the Lexicon.

FIL	File Name(s) (6.3.4.5)
DES	Description (6.3.4.2)
FMT	Format (6.3.4.3)
CLS	Number of Columns (6.3.4.7)
RWS	Number of Rows (6.3.4.4)
BTS	Size in Bytes (6.3.4.6)

6.4.3.12. - Word Index. (file = lrwrd)

WRD	Word (6.3.4.1)
EUI	The Entry Unique ID Number (6.3.2.1)

6.4.3.13. - Fields (file = lrfld)

COL	Three Letter Field Name (6.3.4.8)
DES	Description (6.3.4.2)
REF	Cross Reference to Document (6.3.4.9)
FIL	File Name(s) (6.3.4.5)

6.5. The SPECIALIST Lexicon Unit Record

The unit lexical record is a frame structure consisting of slots and fillers. Each lexical record has a base= slot whose filler indicates the base form, and optionally a set of spelling_variants= slots to indicate spelling variants. Lexical entries are delimited by entry= slots filled by the EUI number of the entry. EUI numbers are seven digit numbers preceded by an "E". Each entry has a cat= slot indicating part of speech. The lexical record is delimited by braces ({...}).

The unit lexical records for "anaesthetic" given below illustrate some of the features of a SPECIALIST unit lexical record:

{base=anesthetic spelling_variant=anaesthetic entry=E0354094 cat=noun variants=reg variants=uncount}
{base=anesthetic spelling_variant=anaesthetic entry=E0330019 cat=adj variants=inv position=attrib(3) position=pred stative}

The base form "anesthetic" and its spelling variant "anaesthetic" appear in two lexical records containing a noun and a verb entry. The variants= slot contains a code indicating the inflectional morphology of the entry; the filler reg in the noun entry indicates that the noun "anaesthetic" is a count noun which undergoes regular English plural formation ("anaesthetics"); inv in the variants= slot of the adjective entry indicates that the adjective "anesthetic" does not form a comparative or superlative. The position= slot indicates that the adjective "anaesthetic" is attributive and appears after color adjectives in the normal adjective order.

The SPECIALIST technical report "The SPECIALIST Lexicon" gives a full description of the Lexicon in unit format.

6.6. Lexical Databases Introduction

The lexical databases contain lexical information that we have found to be useful for Natural Language Processing. They are not finished products but are under continuous development.

6.6.1. Semantically Related Terms SM.DB

This database (SM.DB) contains pairs of semantically related terms. Each row of the database has the following form.

Index String|BAS1|SCA1|BAS2|SCA2|SRC (source of synonym)

Such a row indicates that BAS1 in syntactic category SCA1 is semantically related to BAS2 in syntactic category SCA2. Both terms are given in base form.

Examples:

able|able|1|ability|128|E0006510

alar|alar|1|wing|128|NLP_LVG

ocular|ocular|128|eye|128|C0015392

auditory area|auditory area|128|auditory cortex|128|C0004302

vomitive|vomitive|128|emetic|128|NLP_LVG

vomitive|vomitive|1|emetic|1|NLP_LVG

iridescent virus|iridescent virus|128|iridovirus|128|NLP_LVG

typhloteritis|typhloteritis|128|cecitis|128|NLP_LVG

6.6.2. Derivationally Related Terms: DM.DB

This database (DM.DB) contains pairs of terms related by derivational morphology. Both terms are given in base form.

BAS1|SCA1|EUI1|BAS2|SCA2|EUI2|negation|Type|Prefix|

Examples:

abase|verb|E0006432|abasement|noun|E0006433|O|S|None

abdominal|adj|E0006444|abdominal|noun|E0554771|O|Z|None
acrosome|noun|E0007035|acrosomeless|adj|E0237024|N|S|None
adenohypophyseal|adj|E0007295|adenohypophysis|noun|E0007296|O|S|None
arithmetician|noun|E0359753|arithmetic|noun|E0010398|O|S|None

bone|noun|E0013675|boneless|adj|E0359802|N|S|None
immobilize|verb|E0033519|immobilizer|noun|E0408339|O|S|None

pretest|noun|E0312255|test|noun|E0060348|O|P|pre

unburden|verb|E0062940|burden|verb|E0014409|N|P|un

DM.DB is derived from the morphological fact files (derivation.data) used in lvg (See Lexical Variant Generation section in Section 6.8).

6.6.3. Neo-classical Combining Forms NC.DB

This database (NC.DB) contains morphemes that are used to form neo-classical compounds. Each row of the database has the following form.

MORPHEME|MEANING|TYPE

Morphemes may have optional connecting vowels indicated in parentheses. The types are: prefix, root, and terminal.

Examples:

abdomin(o)|abdomen|root
ab|away from|prefix
acou(o)|hearing|root
cardi(o)|heart|root

cele|swelling|terminal
desis|binding|terminal
de|negate|prefix

Our analysis of combining forms divides them into roots and terminals, which are distinguished from prefixes and suffixes. A neo-classical compound can consist of any number of roots ending in a terminal or suffix. Prefixes normally must precede roots and cannot attach directly to terminals. Users interested in suffixation rules and facts should consult the dm.rul and dm.fct files included with lvg.

For further discussion see McCray et. al., 1988, "The Semantic Structure of Neo-Classical Compounds", In the Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care, Washington DC.

6.7. Sample Records

::::::::::::::
lragr.sam
::::::::::::::
E0007127|acute|adj|positive;periph|acute|acute|
E0014875|cans|noun|count(thr_plur)|can|can|
E0014875|can|noun|count(thr_sing)|can|can|
E0014876|canned|verb|past_part|can|can|
E0014876|canned|verb|past|can|can|
E0014876|canning|verb|pres_part|can|can|
E0014876|cans|verb|pres(thr_sing)|can|can|
E0014876|can|verb|infinitive|can|can|
E0014876|can|verb|pres(fst_sing,fst_plur,thr_plur,second)|can|can|
E0014877|can't|modal|pres:negative|can|can|
E0014877|cannot|modal|pres:negative|can|can|
E0014877|can|modal|pres|can|can|
E0014877|couldn't|modal|past:negative|can|can|
E0014877|could|modal|past|can|can|
E0014937|canine teeth|noun|count(thr_plur)|canine tooth|canine tooth|
E0014937|canine tooth|noun|count(thr_sing)|canine tooth|canine tooth|
E0017902|colors|noun|count(thr_plur)|color|color|
E0017902|color|noun|count(thr_sing)|color|color|
E0017902|color|noun|uncount(thr_sing)|color|color|
E0017903|colored|verb|past_part|color|color|
E0017903|colored|verb|past|color|color|
E0017903|coloring|verb|pres_part|color|color|
E0017903|colors|verb|pres(thr_sing)|color|color|
E0017903|color|verb|infinitive|color|color|
E0017903|color|verb|pres(fst_sing,fst_plur,thr_plur,second)|color|color|
E0051632|quickly|adv|positive;periph|quickly|quickly|
E0055585|she|pron|thr_sing|she|she|

::::::::::::::
lrcmp.sam
::::::::::::::
E0014876|can|verb|tran=np|
E0017903|color|verb|cplxtran=np,adj|

E0017903|color|verb|cplxtran=np,np|
E0017903|color|verb|intran;part(in)|
E0017903|color|verb|intran;part(up)|
E0017903|color|verb|intran|
E0017903|color|verb|tran=np;part(in)|
E0017903|color|verb|tran=np|

::::::::::::::
lrmod.sam
::::::::::::::
E0007127|acute|adj|attrib(1),attrib(3),pred|stative|
E0051632|quickly|adv|verb_modifier;manner||

::::::::::::::
lrnom.sam
::::::::::::::
E0007121|acuity|noun|E0007127|acute|adj|
E0021126|deduction|noun|E0021123|deduce|verb|
E0021126|deduction|noun|E0021124|deduct|verb|
E0061851|transportation|noun|E0061850|transport|verb|

::::::::::::::
lrprn.sam
::::::::::::::
E0030918|he|thr_sing|pers(masc)|subj||||
E0036100|it|thr_sing|neut|subj,obj||||
E0055585|she|thr_sing|pers(fem)|subj||||

::::::::::::::
lrprp.sam
::::::::::::::
E0007127|acute|acute|adj|stative|
E0004825|Parkinson|Parkinson|noun|proper|
E0014877|can|can't|modal|negative|
E0014877|can|can't|modal|negative|
E0014877|can|couldn't|modal|negative|

::::::::::::::
lrspl.sam
::::::::::::::
E0017902|colour|color|

E0017903|colour|color|

E0330019|anaesthetic|anesthetic|

E0354094|anaesthetic|anesthetic|

::::::::::::::
lrtrm.sam

:::::::::::::::
E0412633|Actinex|meso-nordihydroguaiaretic acid|

E0302749|Halcion|triazolam|

E0302640|Tespamin|thiotepa|

E0523571|Somavert|pegvisomant|

:::::::::::::::
lrtyp.sam
:::::::::::::::
E0007127|acute|adj|inv;periph|

E0014875|can|noun|reg|

E0014876|can|verb|regd|

E0014937|canine tooth|noun|irreg(canine tooth,canine teeth)|

E0017902|color|noun|reg|

E0017902|color|noun|uncount|

E0017903|color|verb|reg|

E0051632|quickly|adv|inv;periph|

:::::::::::::::
lrwd.sam
:::::::::::::::
color|E0017902|

color|E0017903|

color|E0017913|

color|E0017914|

color|E0017915|

color|E0017916|

color|E0017917|

color|E0017918|

color|E0065135|

color|E0205800|

color|E0215092|

color|E0220891|

color|E0220892|

color|E0220987|

color|E0237464|

color|E0321442|

color|E0322071|

color|E0330531|

color|E0339331|

color|E0339717|

color|E0374934|

color|E0418710|

color|E0420428|

color|E0428116|

color|E0430071|

color|E0431208|

color|E0504891|

color|E0509680|

color|E0516052|

color|E0519343|

color|E0523712|

color|E0533351|

color|E0552362|

color|E0568432|

color|E0572664|

color|E0572665|

color|E0579060|

color|E0580178|

color|E0582267|

color|E0582516|

color|E0582518|

color|E0582525|

color|E0582528|

color|E0582540|

color|E0583066|

color|E0586414|

color|E0586415|

color|E0587163|

color|E0587164|

color|E0587511|

color|E0587953|

color|E0588132|

color|E0588134|

color|E0588135|

color|E0603013|

color|E0610103|

color|E0610104|

color|E0618392|

color|E0624624|

color|E0627004|

color|E0631243|

color|E0634071|

color|E0635238|

color|E0637227|

color|E0669219|

color|E0669361|

color|E0669362|

color|E0669363|

color|E0670608|

color|E0670610|

color|E0675138|

can|E0014875|

can|E0014876|

can|E0014877|

can|E0562457|

6.8. The SPECIALIST Lexical Tools

The SPECIALIST Lexical Tools package consists of three primary programs -- a normalizer, a word index generator, and a lexical variant generator, together with a set of ancillary programs for normalization. This package is implemented in Java.

The SPECIALIST Lexical Tools and the SPECIALIST Lexicon are distributed as one of the UMLS Knowledge Sources and along with the SPECIALIST NLP Tools as open source resources subject to these terms and conditions.

Updates and bug fixes can be found in the release notes on the Download Lexical Tools Web page.

The distributions come with install programs (for Solaris, Linux, and Window) and a ReadMe.txt file describing how to install and configure the Lexical Tools and providing a brief description of each program.

The **docs** directory contains user guides, Java API documents, and design documents describing in detail the use of Lexical Tools. This document is a general introduction to the programs in the lexical variant generation package.

The compressed Lexical Tools are as follows*:

lvg2008.tgz

- The official distribution of lvg. This includes the source code for the programs, the data and tables in a pure Java embedded database (Instant DB) the programs use, full documentation, installation instructions, and jar files of the programs. See the documents contained within this distribution for a more complete description of this product.

*File names for the 2008 release are shown.

Normalization (norm)

The lexical program **norm** generates the normalized strings that are used in the normalized string index, MRXNS. Thus norm must be used before MRXNS can be searched.

The normalization process involves stripping possessives, replacing punctuation with spaces, removing stop words, lower-casing each word, breaking a string into its constituent words, and sorting the words in alphabetic order. The uninflected forms are generated using the SPECIALIST Lexicon if words appear in the Lexicon, otherwise they are generated algorithmically. When a form could be an inflection of more than one base form, the new normalization process returns multiple uninflected forms. If a string to be normalized contains multiple ambiguous forms, and the permutation of these ambiguous forms offer more than 10 output forms, the input form lowercased, with punctuation replaced, word order sorted, but not uninflected, is returned. The upper limit of permutation number (10) is configurable by modifying the configuration file. The program **luiNorm** has the behavior of prior year's normalization, and is distributed for those who need it.

Norm reads its standard input and writes to standard output. It expects input lines to be records separated into fields. The field separator is **|**. The string to be normalized is identified to norm using the **-t** option. **-t** takes a numerical argument which denotes the field in which the input string is to be found. If no **-t** option appears, norm assumes that the input string is in the first field (**-t:1**). There need not be more than one field, so lines consisting only of input strings are properly understood.

Norm output records include all the fields of the input record with an additional field to the right containing the normalized form of the input string.

For example, if the user had a list of terms to be looked up via the normalized string index in a file called **terms**, he or she could use **norm -i:terms -o:terms.nrm** to get the normalized form of each term. If the input file **terms** contained the following:

2, 4-Dichlorophenoxyacetic acid
Syndrome, anterior, compartment
Abnormal, weight, gain
Anemia, Refractory, with Excess of Blasts
left atriums

the file **term.nrm** would contain:

2, 4-Dichlorophenoxyacetic acid|2 4 acid dichlorophenoxyacetic
Syndrome, anterior, compartment|anterior compartment syndrome
Abnormal, weight, gain|abnormal gain weight
Anemia, Refractory, with Excess of Blasts|anemia blast excess refractory
left atriums|atrium left
left atriums|atrium leave

The string in the second field of each line of terms.nrm is now suitable for matching to MRXNS.

Word Index (wordInd)

The lexical program **wordInd** breaks strings into words for use with the word index in MRXW. Users of the word index should use wordInd to break strings into words before searching in the word index. This assures congruence between the words to be looked up and the word index.

Word for this purpose is defined as a token containing only alphanumeric characters with length one or greater. The wordInd program lowercases the output words.

The wordInd program reads its standard input and writes to its standard output. Like norm and lvg, it expects each input line to be a record separated into fields by **|**. The field containing the input string is identified using the **-t** option. The numerical argument of **-t** denotes the field in which the input string may be found. If no **-t** option is given, the input string is expected to be in the first field (**-t:1**). There need not be more than one field, so lines consisting only of input strings are properly understood.

The wordInd program outputs one line of output for each word found in the input string. Input fields are not repeated in the output unless specified in a **-F** option. Applying wordInd to the input string **Heart Disease, Acute** would result in three output lines:

heart
disease
acute

The numerical argument of **-F** indicates an input field to be repeated in the output. A numerical argument for **-F** option is required for each input field that is to be repeated. Fields are repeated in the order in which the numerical argument of **-F** options appear. The output words always appear as an additional field to the right of any repeated input fields. For example, applying **wordInd -t:2 -F:2:1** to a record of the form **UI23456|tooth, canine|definition.....**; would result in the following output:

tooth, canine|UI23456|tooth
tooth, canine|UI23456|canine

The third field of each of those records contains a word extracted from the input term in the first field (**-t:2** ,**-F:2**). The **-F:1** option repeats the UI numbers from the first field of input. The fact that **-F:2:1** placed the UI numbers (field 1) after the input string (field 2).

Lexical Variant Generation (lvg)

The lvg program generates lexical variants of input words. It consists of several different flow components that can be combined in various ways to produce lexical variants. The user of lvg chooses combinations of flow components and combines them into a **flow**. (The normalizer program, norm, is essentially the lvg program with a pre-selected flow option: **lvg -f:N**.) The arguments of the **-f** flag are used to specify a flow. Each flow can be thought of as a pipeline with each flow component feeding the next. For example, the flow **-f:i** simply generates inflectional variants and **-f:l:i** generates lowercase inflectional variants. Each of the flow components options is discussed on the documents for lvg.

The lvg program reads from its standard input and writes to its standard output. Input records may be typed in at the keyboard, after typing the command on the command line (**lvg -f:i**) or input lines may be read from a file (**lvg -f:i -i:file**) or piped to lvg from another command (**COMMAND|lvg -f:i**). Output records may be directed to the screen (default), send to a file (**lvg -f:i -i:INFILE -o:OUTFILE**) or piped to another command (**lvg -f:i -i:infile | COMMAND**).

Input

The lvg program is designed to work with one line input records divided into fields. The default field separator is **|**. The field separator can be changed using the **-s** option. The field in which the input term, whose variants are to be generated, can be specified with the **-t** option. In the absence of a **-t** flag the input term is assumed to be in the first field of the input. So both **dog** and **dog|canine|UI4567** would generate variants of **dog**. With the **-t** flag set to **2**, **dog|canine|UI4567** would generate variants of **canine**. In the case of single field input (**dog**), lvg generates variants from the only field regardless of the setting of **-t**.

The lvg program can read category (part of speech) and inflection information from the input record. The numerical argument to the **-cf** option indicates the field in which category information is located. In the input record, category information needs to be encoded as a number according to the scheme described on the documents for lvg. The numerical argument to the **-if** option indicates the field in which inflection information is located. In the input record, inflection information needs to be encoded as a number according to the scheme described on the documents for Lexical Tools.

Output

The lvg program adds five new fields to the input record and outputs a record for each variant generated. For example, if **dog|canine|UI4567** is given to the standard input of **lvg -f:i** the output sent to standard out will be:

dog|canine|UI4567|dog|128|1|i|1| dog|canine|UI4567|dog|128|512|i|1| dog|canine|UI4567|dogs|128|8|i|1| dog|canine|UI4567|dog|1024|1|i|1| dog|canine|UI4567|dog|1024|262144|i|1| dog|canine|UI4567|dog|1024|1024|i|1| dog|canine|UI4567|dogs|1024|128|i|1| dog|canine|UI4567|dogged|1024|64|i|1| dog|canine|UI4567|dogged|1024|32|i|1| dog|canine|UI4567|dogging|1024|16|i|1|

The first three fields of each record above are identical to the input record, the rest are supplied by lvg. The first additional field is the variant form lvg has generated. The second additional field is the syntactic category of the variant encoded as a number. The third additional field is the inflection of the variant encoded as a number. The fourth additional field indicates the flow that was selected. The fifth field is the number of the flow which generated this variant. Output category (parts of speech) and inflection information are encoded in the same scheme used for input category and inflection information.

Further description of the SPECIALIST Lexical Tools is available at the SPECIALIST Lexical Tools Web site: http://lexsrv3.nlm.nih.gov/Specialist/Home/index.html.

7. Using the UMLS Terminology Services (UTS) via the Internet

The UMLS Terminology Services (UTS) is a computer application that provides Internet access to the UMLS Knowledge Sources. The purpose of the UTS is to make the UMLS data more accessible to users and, in particular, to systems developers. The system architecture is based on the client server model, allowing remote site users (individuals as well as computer programs) to send requests to a centrally managed server at the U.S. National Library of Medicine. Access to the system is provided through a command line interface, through the World Wide Web, an Extensible Markup Language (XML)-based socket programming interface, and through an Application Programmer Interface (API).

Users are encouraged to consult the UTS Web site for the most current UTS documentation, including the Developer's API Guide and information on downloading the UMLS release files.

7.1. Downloading the UMLS Knowledge Sources

UMLS licensees may access the UTS and create an account with a username and password of their choosing. Licensees can download the current UMLS Knowledge Sources from the UTS. Archives of UMLS releases are kept and made available for several previous years. For detailed technical specifications and installation instructions refer to the README.TXT file available on the Knowledge Sources downloads page.

7.2. System Architecture

The UTS, made available in December 2010, replaced the UMLS Knowledge Source Server (UMLSKS). The UTS provides access to the UMLS Knowledge Sources through a browser-based application and a Web services client. The UTS uses the same underlying domain model as other UMLS systems, including MetamorphoSys, which allows for integrity at the data store level. The UTS features the following enhancements:

- Highly responsive search engine for quick retrieval
- Increased capacity by an order of magnitude
- Scalability to accommodate past and future releases
- High availability via co-location redundancy
- 508 compliance
- More intuitive graphical user interface
- Real-time monitoring and collection of statistics
- Integration of licensing and reporting requirements

7.3. Querying the UTS

7.3.1. Metathesaurus Browser

The UTS allows the user to request information about particular Metathesaurus concepts, including attributes such as the concept's definition, its semantic types, and the concepts that are related to it.

Basic concept information includes the Metathesaurus unique identifier of the concept, the preferred name for the concept, and the names and sources of all terms that comprise that concept. Additional concept information often includes a definition and the source of that definition. Semantic type information is also included. Information about the hierarchical contexts of Metathesaurus concepts is readily available in the system. Related concepts are easily found.

An important perspective on the Metathesaurus is source specific data. It is possible to query the server by limiting the query to a particular vocabulary. The user may wish to see the ancestors or descendants for a term in

just a particular vocabulary, or the user may wish to see just the synonyms for a particular term in a particular vocabulary.

The Metathesaurus Tree tab allows the user to navigate the hierarchy of a selected source vocabulary in order to browse both its content and structure. *Note: Some source vocabularies in the Metathesaurus are not organized into a hierarchical arrangement and cannot be browsed using the Metathesaurus Tree tab.*

7.3.2. Semantic Network Browser

The Semantic Network contains information about semantic types and their relationships. The implementation of the network module computes the relationships between semantic types using the inheritance property of the network type hierarchy. Information in the Semantic Network can be browsed for semantic types and the relationships between them.

It is possible to retrieve all the relations between a pair of types. For example, "treats", "prevents", and "complicates" would be listed, among others, as potential relationships between drugs and diseases. It is also possible to retrieve an exhaustive list of all related types in the network. Queries can be made about the definition, unique identifier, tree number, ancestors, parents, children, descendants, and siblings of a semantic type or relation.

7.3.3. SNOMED CT Browser

The UTS provides access to SNOMED CT content, as included in the Metathesaurus. Users may request information about particular SNOMED CT terms, including descriptions, relationships, hierarchical tree positions, etc. Users may query by SNOMED CT term, SNOMED CT ConceptID, or SNOMED CT DescriptionID.

Concept information includes the SNOMED CT unique identifier of the concept, concept status, and whether or not a concept is primitive. Additional concept information includes its descriptions, parents and children of the concept, relationships to other concepts, and its hierarchical tree positions. The unique identifier(s) and semantic type(s) of the UMLS Metathesaurus concept(s) to which the SNOMED CT concept belongs are also included with the SNOMED CT concept information.

The SNOMED CT Tree tab allows the user to navigate the hierarchy of SNOMED CT in order to browse both its content and structure.

7.4. Gaining Access to the UTS

Access to the UTS is available to anyone who has signed the UMLS Metathesaurus License Agreement and activated a UTS account. First time users should click “Sign Up” on the UTS homepage to begin the license request and UTS account activation process. Any questions or problems should be addressed via e-mail to NLM Customer Support.

7.5. UTS Documentation

UTS users should always consult the documentation on the UTS for the most current information.

The following are publicly available on the site, under Documentation:

- Source documentation pages.
- UMLS database query diagrams.
- Developer's API Guide -- documentation generated using the javadoc facility that includes the object model, interfaces and some examples.

- Information on validating UMLS licensees for third party applications.
- A link back to the UMLS Reference Manual.
- A link to RxNorm and SNOMED CT documentation.

8. MetamorphoSys - The UMLS Installation and Customization Program

MetamorphoSys is the UMLS installation wizard and Metathesaurus customization tool included in each UMLS release. It installs one or more of the UMLS Knowledge Sources. When the Metathesaurus is selected, it enables you to create customized Metathesaurus subsets. Please use only the version of MetamorphoSys distributed with the release.

Users customize their Metathesaurus subsets for two main purposes:

1 To exclude vocabularies from output that are not required or licensed for use in a local application.

The Metathesaurus consists of a number of files, some of which are extremely large; excluding sources can significantly reduce the size of the output subset. Given the number and variety of vocabularies reflected in the Metathesaurus, it is unlikely that any user would require all, or even most, of its more than 100 vocabularies. In addition, some sources require separate license agreements for specific uses, which a UMLS user may not wish to obtain. These are clearly indicated in the License Agreement.

2. To customize a subset using a variety of data output options and filters.

To identify vocabularies that may not be needed in a customized subset, read the License Agreement, and refer to UMLS Source Vocabulary Documentation page of the current UMLS release documentation.

The RRF Browser, which allows users to find a term within a customized Metathesaurus subset or any vocabulary in the RRF format, is also included in MetamorphoSys.

There are no license restrictions on the MetamorphoSys code. We hope that users will acknowledge the NLM source, in the spirit of the GNU Public License (GPL).

8.1. MetamorphoSys Requirements

MetamorphoSys has been tested on the following operating systems:

- Windows 7 Enterprise, Vista, XP, 2000, NT
- Linux, all releases are fully tested under Red Hat Workstation Linux; other Linux releases may work equally well
- Solaris 9
- Macintosh OS X (Leopard, Snow Leopard)**

It is implemented in Java and requires the run-time JRE version included in the release (except for the Macintosh, which licenses its own JRE). Solaris and Windows Java Runtime Environment — http://javasoft.com.

Macintosh note: **MetamorphoSys requires Java 1.6.

You may use a high-speed Internet connection to download the UMLS files from the UMLS Web site downloads page. To ensure proper functionality users should download and extract all UMLS data and zip files to the same directory.

To use the DVD, which was discontinued as of the 2012AB UMLS release, you must have a DVD reader and at least 30 GB of free disk space. Multiple runs that create multiple subsets of the Metathesaurus will need even more space. For reasonable performance, we suggest these minimum requirements:

- A CPU of 2GHz or higher
- 6x (or better) DVD drive
- 2 GB of RAM, preferably more

- A DVD reader which can read the standard UDF DVDs

DVD options allow you to (1) install the UMLS Knowledge Sources from the DVD, (2) copy MetamorphoSys .nlm data format files to local storage, and (3) copy the installation program and files to local storage. This may be useful for multiple runs or subsetting an existing subset, and it may improve performance time.

All file sizes are checked at installation. The Validate Distribution option allows users to verify the integrity of .nlm files downloaded from the UMLS Web site or copied from the UMLS DVD. It compares special MD5 signatures to those in the release .MD5 file. CHK file, and is a useful first step for trouble-shooting when problems occur with a UMLS installation.

If the UMLS release is downloaded from the UMLS Web site downloads page, it must include these files*, in the same directory:

- mmsys.zip (zipped MetamorphoSys application)
- 2009aa-1-meta.nlm (compressed Metathesaurus data)
- 2009aa-2-meta.nlm (compressed Metathesaurus data)
- 2009aa-otherks.nlm (compressed Semantic Network and SPECIALIST Lexicon)
- 2009AA.CHK
- 2009AA.MD5
- Copyright_Notice.txt
- README.txt

The mmsys.zip file is first unzipped to local storage and the MetamorphoSys application started. To ensure proper functionality users must unzip mmsys.zip to the same directory as the other downloaded files.

*File names for the 2009AA release are shown.

8.2. Starting MetamorphoSys

Open a terminal window and change to the root directory of the DVD. Type the appropriate command for your platform:

- ./run_mac.sh (or click on the run_mac.command file)
- ./run_linux.sh
- ./run_solaris.sh

Press the return key.

A new window will appear. This may take a few minutes since a good deal of software must load before the Welcome screen appears.

- Windows run.bat

On Windows machines with Autorun enabled, the DVD will start automatically. If it does not, go to the root directory of the DVD and click on the file named run.bat.

8.3. MetamorphoSys Help

Help is available at the UMLS Web site. Users may also receive assistance from Webcasts, the UMLS Listserv, and the various MetamorphoSys Tutorials. We are developing additional Web resources based on user input.

9. UMLS DVD

Prior to the 2013AA release, the UMLS was available on a single disk DVD which contained the three UMLS Knowledge Sources in three compressed format files (.nlm data format) and the MetamorphoSys file. (File names for the 2009AA release are shown.)

- 2009aa-1-meta.nlm
- 2009aa-2-meta.nlm (Metathesaurus = 2 files)
- 2009aa-otherks.nlm (Semantic Network and SPECIALIST Lexicon)
- MMSYS.zip, MetamorphoSys - the UMLS installation and Metathesaurus subsetting program for PC and UNIX machines (cross-platform Java executable with JRE for each supported platform)

The DVD was discontinued as of the 2012AB UMLS release.

9.1. Hardware and Software Requirements

9.1.1. Supported Operating Systems

- Windows Vista, XP, 2000, NT
- Linux*
- Solaris 9
- Macintosh OS X (Leopard, Snow Leopard)**

*Linux note: The specific version fully validated is Red Hat Enterprise Linux WS, Version 5. Other versions should also work.

Macintosh note: **MetamorphoSys requires Java 1.6.

9.1.2. Hardware Requirements

- A MINIMUM of 30 GB of free hard disk space
- A MINIMUM of 2 GB of RAM, preferably more. Smaller memory size will cause virtual memory paging with exponentially increased processing time.
- A CPU speed of at least 2 GHz
- DVD drive

9.2. Installing from the DVD

Insert the DVD into a DVD drive. DVD drives must be mounted as UDF, the standard file system type for all DVDs. File length validation errors and other errors will result if a non-UDF drive is used. For best results the drive should be 6X or faster.

Start the MetamorphoSys install program:

Windows

- The DVD should autorun.
- If it does not (or if you are installing from the hard disk), go to the DVD root directory and click on "run.bat."

Linux, Solaris, and Macintosh

- Open a terminal window and change to the root directory of the DVD, then type the appropriate command for your platform:

./run_linux.sh

./run_solaris.sh

./run_mac.sh

- Hit the return key
- Select "Install UMLS" on the Welcome to MetamorphoSys screen
- Select destination (local) directory for files on the Install UMLS screen

The Install UMLS Metathesaurus progress monitor charts the process through:

- Initializing the CUI list
- Subsetting Content
- Subsetting Indexes
- Final Processes

If selected, Semantic Network and SPECIALIST Lexicon files are copied first. Accept the License Agreement notice to proceed with customizing the Metathesaurus.

Select "Cancel" to exit MetamorphoSys at any time. The interrupted process cannot be resumed. The configuration must be recalled (if saved), or recreated (if not saved), and subsetting must be started again.

MetamorphoSys produces an initial install.log file of the installation up to the start of Metathesaurus subsetting. If files pass validation, processing continues and subsetting begins. If validation fails, a warning is displayed and recorded in install.log.

When subsetting is complete, configuration settings are displayed on the screen and also written to "mmsys.log" in the directory containing the subsetted files. Your customized Metathesaurus files are located in the destination directory.

For more information on running MetamorphoSys see Chapter 8 of this documentation.

10. Current UMLS Release Information

The UMLS is updated twice each year in May (AA release) and November (AB release). For information about UMLS releases, licensing, and metadata, visit the current UMLS release Web page.