Metathesaurus

[2]

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 obtain a DVD of the UMLS Knowledge Sources or download them 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.

UMLS® Reference Manual

UMLS® Reference Manual

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 Source Vocabularies 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

UMLS® Reference Manual

UMLS® Reference Manual

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.

MetamorphoSys also generates special subsets referred to as Content Views. A content view may specify any pre-defined subset of the Metathesaurus that is useful for some specific purpose. The actual definition of a content view can take a variety of different forms: (1) an actual list of Metathesaurus UIs maintained over time; (2) a list of sources that participate in the view; and (3) a complex query that identifies particular sets of data.

A Content View Flag (CVF) consists of an arbitrary bit field, with each bit representing membership in a particular Content View; each Content View is documented in MRDOC.RRF. The first Content View available in the 2005AA release, the MetaMap NLP View, identifies terms that are useful for Natural Language Processing. The CVF in rows with these terms carries the value "1" in the "256" bit. MetamorphoSys users who wish to use this special subset should choose File Menu, Enable/Disable Views to implement this feature.

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 Source Vocabularies 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 Source Vocabularies 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

UMLS® Reference Manual

UMLS® Reference Manual

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 the code sets mandated for use in electronic administrative transactions in the U.S. under the provisions of the Health Insurance Portability and Accountability Act (HIPAA). With the exception of the National Drug Codes (NDC), the Metathesaurus includes all concepts and terms from these code sets. NDC codes available from the Food and Drug Administration are included as attributes of clinical drug concepts present in the FDA National Drug Code Directory (MTHFDA), which is a source vocabulary.

NLM intends to incorporate all clinical terminologies designated as target U.S. government-wide standards by the Consolidated Health Informatics (CHI) initiative and/or recommended as U.S. standards by the National Committee on Vital and Health Statistics. Several of these (e.g., LOINC, SNOMED CT, RxNorm) are already present in the Metathesaurus.

The fact that a vocabulary has been designated as a HIPAA or CHI standard is included on the Source Vocabularies page.

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.

UMLS® Reference Manual

UMLS® Reference Manual

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.

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.

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 Source Vocabularies 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.

Some of the intra-source vocabulary relationships are statistical relationships, which are computed by determining the frequency with which concepts in specific vocabularies co-occur in records in a database. For example, there are co-occurrence relationships for the number of times concepts have co-occurred as key topics within the same articles, as evidenced by the Medical Subject Headings assigned to those articles in the MEDLINE database. Co-occurrence relationships have also been computed for different ICD-9-CM diagnosis codes assigned to the same patients as reflected in a discharge summary database. In contrast to the relationships asserted within source vocabularies, the statistical relationships in the Metathesaurus can connect very different concepts, such as diseases and drugs. There are specific Metathesaurus files for the co-occurrence relationships (MRCOC.RRF and MRCOC in ORF).

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).

UMLS® Reference Manual

UMLS® Reference Manual

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.

2.5 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.

UMLS® Reference Manual

UMLS® Reference Manual

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	(C0024117, L0024117, S0058463)
Obstructive Lung Diseases	(C0024117, L0024117, S0068169)
Lung Disease, Obstructive	(C0024117, L0024117, S0058458)
Obstructive Lung Disease	(C0024117, 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

Sample Records

ENG|000003|C1273274|L3139159|S3660797|
ENG|000003|C1306276|L3139160|S3660798|

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|C0024117|L0024117|S0058458|
ENG|disease|C0024117|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|C0024117|L0024117|S0058458|
ENG|disease|C0024117|L0024117|S0058463|

ENG|disease|C0024117|L0024117|S0068168|
ENG|disease|C0024117|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|C0024117|L0024117|S0058458|
ENG|disease lung obstructive|C0024117|L0024117|S0058463|
ENG|disease lung obstructive|C0024117|L0024115|S0068168|
ENG|disease lung obstructive|C0024117|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.

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 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)

Table 2. Representation of Ambiguity in the Metathesaurus.

Concepts (CUIs)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF only
C0009264 **cold temperature**	**L0215040** cold temperature	**S0288775** cold temperature	**A0318651** cold temperature (from CSP)
	L0009264 Cold Cold	**S0007170** Cold	**A0016032** Cold (from MTH)
		S0026353 Cold	**A0040712** Cold (from MSH)
C0009443 **Common Cold**	**L0009443** Common Cold	**S0026747** Common Cold	**A0041261** Common Cold (from MSH)
	L0009264 Cold Cold	**S0007171** Cold	**A0016033** Cold (from MTH)
		S0026353 Cold	**A0040708** Cold (from COSTAR)
C0024117 **Chronic Obstructive Airway Disease**	**L0498186** Chronic Obstructive Airway Disease	**S0837575** Chronic Obstructive Airway Disease	**A0896021** Chronic Obstructive Airway Disease (from MSH)
	L0008703 Chronic Obstructive Lung Disease	**S0837576** Chronic Obstructive Lung Disease	**A0896023** Chronic Obstructive Lung Disease (from MSH)
	L0009264 COLD COLD	**S0829315** COLD	**A0887858** COLD (from MTH)
		S0474508 COLD	**A0539536** COLD (from SNMI)