The UMLS Metathesaurus®: representing different views of biomedical concepts

By Peri L. Schuyler, M.L.S.
Head, Medical Subject Headings Section
Library Operations

William T. Hole, M.D.
Research Medical Officer
Computer Science Branch
Lister Hill National Center for Biomedical Communications

National Library of Medicine
8600 Rockville Pike
Bethesda, Maryland 20894

Mark S. Tuttle

David D. Sherertz

Lexical Technology, Inc.
1000 Atlantic Avenue, Suite 106
Alameda, California 94501

The UMLS Metathesaurus® is a compilation of names, relationships, and associated information from a variety of biomedical naming systems representing different views of biomedical practice or research. The Metathesaurus is organized by meaning, and the fundamental unit in the Metathesaurus is the concept. Differing names for a biomedical meaning are linked in a single Metathesaurus concept. Extensive additional information describing semantic characteristics, occurrence in machine-readable information sources, and how concepts co-occur in these sources is also provided, enabling a greater comprehension of the concept in its various contexts.

The Metathesaurus is not a standardized vocabulary; it is a tool for maximizing the usefulness of existing vocabularies. It serves as a knowledge source for developers of biomedical information applications and as a powerful resource for biomedical information specialists.

INTRODUCTION

The UMLS Metathesaurus® is the UMLS Knowledge Source that represents biomedical concepts derived from a variety of controlled vocabularies, classifications, and other biomedical terminologies, such as collections of terms used in ambulatory care or clinical records systems. In accordance with the definition of *meta* in Webster's—"more comprehensive, transcending" [1]—the Metathesaurus transcends its individual sources by encompassing and combining their scopes, by organizing the resulting collection by meaning, and by adding useful information and interconcept relationships not present in any of its source vocabularies.

The Metathesaurus reflects the philosophy of the English physician, Peter Mark Roget, whose *Thesaurus* is a collection of words organized by meaning [2]. Each meaning in a Metathesaurus source is represented as a single concept linked to the names for that meaning in any Metathesaurus source. The notion of concept rather than term is central to the purpose of the UMLS Metathesaurus. By linking different terms used to express the same concept, the UMLS Metathesaurus transcends specific vocabularies, conveys meaning, and reduces ambiguity.

Each concept is linked to associated information from its sources and to information added in construction of the Metathesaurus. Examples of added information are the UMLS Semantic Types assigned to each concept; the labeling of relationships between concepts in the MeSH® "trees," or hierarchies; and definitions from independent sources.

The Metathesaurus is a knowledge source for system developers, to be used in building applications for the retrieval and integration of biomedical information from diverse machine-readable information sources. It is also a knowledge source of interest to those concerned with biomedical naming and information retrieval. The UMLS Information Sources Map identifies and describes available machine-readable information sources; the UMLS Semantic Network defines potential relationships among the broad categories or semantic types to which all Metathesaurus concepts are assigned. When the Information Sources Map and Semantic Network are combined with the Metathesaurus, the result is a rich resource for improving information retrieval systems in the field of biomedicine.

The 1992 edition of the UMLS Metathesaurus contains more than 270,000 names for more than 130,000 identified concepts from the twenty source vocabularies listed in Appendix A. It also includes a great deal of associated information, including source vocabulary hierarchies, more than 46,000 definitions and scope notes, and in excess of 350,000 established relationships between different concepts. The Metathesaurus also indicates which of its concepts appear in different databases and systems, pointing toward further information. Nearly seven million co-occurrences of concepts in databases such as MEDLINE® are included, giving additional insight into the meaning and the usage of these concepts.

PHILOSOPHY

The Metathesaurus is established as a "closed world," representing only those concepts or "meanings implicit in the sources from which it is constructed" [3]. This approach is at once restrictive and comprehensive. It is restrictive in the sense that the number of meanings is limited to the number expressed in the component vocabularies. It is also comprehensive, in that all meanings expressed in component vocabularies will eventually be represented in the Metathesaurus. The Metathesaurus is not intended to be a single coding scheme, an overarching classification system, or a standard vocabulary system. What is being developed is a mechanism for identifying the meaning of biomedical names for translating, interpreting, and resolving ambiguity. The goal is to transcend the manner in which names are expressed, the purpose for which they have been established, or the nature and location of the information they are used to access.

The Metathesaurus is being built by successive approximation. For this reason, it does not have as many sources, concepts, or relationships as are planned ultimately. Since the scope of the Metathesaurus is determined by the scope of its sources, its breadth and depth will grow as the number and the diversity of its sources increase.

Additions and changes to the Metathesaurus are guided by user evaluation and feedback. For example, vocabulary sources added in 1992 were suggested by users or selected as a result of requests for increased coverage of clinical medicine. A key to appropriate growth and direction is widely based evaluation and effective feedback [4].

ORGANIZATION AND CONSTRUCTION OF THE METATHESAURUS

Metathesaurus construction begins with the collection of the names, relationships, and associated information found in each source vocabulary. Names from all sources are collected into three successive classes. Identical character strings in the same language, disregarding case, are grouped into string classes; lexical variants are grouped into term classes; and synonymous names are joined in a concept. Classes are constrained by meaning. In cases where a string has multiple meanings, there is a separate string, term, and concept class for each distinct meaning.

Names are characterized as lexical variants if they differ only by spelling (hematology/haematology); word order (tuberculosis, pulmonary/pulmonary tuberculosis); punctuation (T-lymphocytes/T lymphocytes); number (dog/dogs); or abbreviated status (organometallic compounds/organometallic cpds). Lexical variation is the tightest of all relationships in the Metathesaurus because it is open to the least degree of interpretation.

Synonyms are lexically dissimilar names which carry the same meaning; sets of synonyms form equivalence relationships to each other. Because it is open to a greater degree of interpretation, synonymy is not as tight a conceptual relationship as lexical variation.

The complete set of preferred names, lexical variants, and synonyms form a concept or conceptual group. The larger the set becomes, the greater will be the opportunity for concept recognition, interpretation, and understanding.

An example of string classes and term classes forming a concept is shown in Table 1.

To construct the Metathesaurus, candidate classes are computed. String classes and term classes are created by programs which match strings and recognize lexical variants. Lexical similarity is used to propose candidate relationships. Synonymy and other rela-

tionships are also derived from information present in the source vocabularies.

Additional information, such as candidate definitions or default semantic types, are added through source analysis and automated techniques. After automated processing, the candidate concepts are reviewed and corrected by subject matter experts who examine the names for a concept, the meanings in the source vocabularies, the correctness of the relationships, and the accuracy of other linked concept information.

The meaning of a name within a source vocabulary is judged from all available evidence, including scope notes or definitions; contexts, either explicit (such as those within hierarchical structures) or implicit (such as those implied by the purpose of the source, e.g., the naming of laboratory procedures); relationships to other names within the source; the perspective or philosophy of the creator, for example, the scope or relationships for a concept as viewed by biochemists may differ from that of practicing physicians; and the usage of the source in real applications.

A meaning not found in one of the sources is not created for the Metathesaurus. Until the Metathesaurus contains sources representing all possible biomedical usage, there will be cases where a name's meaning, in some contexts and from some viewpoints, will not be included. As the Metathesaurus grows to encompass a greater diversity of sources, these cases become less frequent.

SEMANTIC CHARACTERIZATION

A primary device for characterizing the meaning of concepts in the Metathesaurus is the Semantic Type. There are 134 Semantic Types, or categories, such as "Disease or Syndrome" and "Pharmacologic Substance," in the 1992 edition of the Semantic Network, which is described in greater detail in the paper by McCray et al. in this symposium [5]. Semantic Types are assigned to concepts based on the intrinsic and functional properties of each concept.

Semantic Types help to distinguish different meanings associated with a single name. "Meaning" is understandably relative. It is relative first to the scope and granularity of the Metathesaurus; it is relative second to the source vocabulary from which the name originated. That names alone are often not sufficient for unambiguous identification of concepts can be illustrated by the term *osteopathy*. In MeSH, "OSTEOPATHY" is the discipline of osteopathic medicine. In SNOMED II, "osteopathy" is a general expression for bone disease. Thus in the Metathesaurus, "osteopathy" denotes two different concepts, each with its own definition and each labeled with its own semantic type and source designation [6]. In this case, assignments "Biomedical Occupation or Discipline" and "Disease or Syndrome" are sufficient to discriminate between the two meanings of *osteopathy*, even in the absence of definitions.

Table 1
An example of grouping of biomedical names in Metathesaurus classes

String classes	Term classes	Concept
Atrial fibrillation* Atrial fibrillations Fibrillation, atrial Fibrillations, atrial	Atrial fibrillation*	Atrial fibrillation
Auricular fibrillation* Auricular fibrillations Fibrillation, auricular Fibrillations, auricular	Auricular fibrillation (Synonym)	
Fibrillation auriculaire*	Fibrillation auriculaire (Translation, French)	

* Preferred name at each level.

Semantic Types provides a general indication of meaning which may enable accurate matching of concepts from various sources and the discrimination among them regardless of lexical similarities. The types serve to complement or reinforce definitions in conveying meaning.

CONCEPT NAMING IN THE METATHESAURUS

The Metathesaurus identifies the preferred name for a concept in each component vocabulary, and hence allows translation between source vocabularies. As a convenience for managing and maintaining the Metathesaurus, one name is selected as the preferred Metathesaurus name for each concept. In establishing this preferred name, a rank was assigned to all sources. For example, if a concept exists in MeSH, the preferred form of the concept is the form of the main heading in MeSH. All other terms, phrases, and strings used to express the concept are mapped to this Metathesaurus name. The collection of names is considered identical in meaning, either through synonymy or through lexical variation, so system developers are free to select different preferred names for their particular applications.

RELATIONSHIPS BETWEEN CONCEPTS

In addition to specifying relationships between different names for the same concept, the Metathesaurus also represents a variety of relationships among different concepts. These relationships are derived from information present in the source vocabularies; from relationships suggested by computer programs and then reviewed by editors; or from external sources, such as a special effort to label hierarchical relation-

ships in parts of MeSH. In general, different concepts are linked because they share properties or are similar along some dimension. The majority of the interconcept relationships come from hierarchical contexts present in the source vocabularies, such as parent, child, and sibling relationships. The ability to identify clusters of closely related concepts, to determine the nature of the relationships within the cluster, and to identify the sources from which the relationships are derived offers a significant potential advantage in identifying similarities and differences in meaning in disparate sources [7].

Other kinds of relationships are indicated by the occurrence of concepts in systems or databases. For example, the fact that the concept *Lyme disease* occurs in the AI Rheum, DXPlain, and QMR systems as well as in MEDLINE is recorded in the Metathesaurus occurrence data. The frequency with which specific qualifiers have been used in MEDLINE in conjunction with a specific concept is also noted. For example, *breast neoplasms,* as a principal concept in indexed articles, was qualified by pathology 3,043 times; by drug therapy, 1,322 times; by diagnosis, 995 times; and by etiology, 395 times in a particular time span of the MEDLINE database.

The co-occurrence of concepts in selected databases is also indicated. Continuing with the example of *breast neoplasms,* "antineoplastic agents, combined" co-occurred with "breast neoplasms" 438 times; with "receptors, estrogen," 437 times; and with "Tamoxifen," 306 times, when both were identified as principal concepts of the same article.

USE OF THE METATHESAURUS

The goal of the UMLS project is to facilitate the retrieval and integration of information from machine-readable biomedical sources—transparently, whenever possible. The Metathesaurus can be used also to help searchers interactively by asking which meaning of an ambiguous name is desired. Queries which are likely to be unsatisfactory can be identified before they are executed, such as by looking at the relationships between semantic types of the specific concepts in the query or by checking concept occurrence and co-occurrence information. Searchers may also browse Metathesaurus concepts and relationships to formulate a query. The paper by Kingsland et al. in this symposium provides specific examples of these and other approaches that use Metathesaurus information to provide MEDLINE search assistance [8].

It is also possible to use the Metathesaurus in an unindexed environment. Instead of translating concepts into controlled vocabularies, alternative concept names can be collected from the Metathesaurus for use as text words in free-text searches. A search can be enriched automatically by collecting all lexical variants and synonyms for a concept and including the names of related concepts as well. The degree to which the query is expanded can be controlled by the user by selecting from kinds of relationships or from lists of related concepts.

Potential Metathesaurus uses in other retrieval applications include query formulation, query screening through examination of semantic type relationships, query evaluation [9], the use of occurrence and co-occurrence information to select information sources or to predict the success of searches, or the use of various data elements for the integration and the ranking of retrieved information for display.

Broader uses of the Metathesaurus are also possible, including its use in building or maintaining other naming systems [10]; in indexing clinical records [11–12] as well as biomedical journal articles; and in many areas of medical informatics research.

CHALLENGES

The Metathesaurus is a complex example of a knowledge source composed of multiple, independent vocabulary sources, each updated and maintained on its own schedule. If one considers what is involved in mapping between only two vocabularies, such as MeSH and Library of Congress Subject Headings (LCSH), and then increases the number of sources by an order of magnitude, it can be seen that the increase in complexity is exponential. Changes made in one source which are not represented correctly in the Metathesaurus may undermine the integrity of the relationships in the Metathesaurus and create a schism between the Metathesaurus representation and a new version of the source vocabulary. Tuttle et al. consider the maintenance of synchrony between local systems and each subsequent release of the Metathesaurus to be of prime importance [13]. The ultimate approach to representing abrupt and gradual changes in biomedical meaning over time is a future challenge.

In the near future, Metathesaurus releases must coincide with annual update schedules, particularly the yearly release of MeSH, if the Metathesaurus is to be used effectively as a tool for retrieval from bibliographic files indexed or cataloged with MeSH. Each release must contain recomputed links between concept names in all Metathesaurus sources. As the number of sources increases, this process becomes increasingly formidable. Fortunately, developments in information science and computer technology are allowing increasing automation of these tasks and are maximizing the productivity of the human specialists who must guide the process and edit the results.

The UMLS Metathesaurus is a large, complex knowledge source which integrates diverse views of biomedicine. The power of the Metathesaurus results from its organization by meaning and from the com-

bined scope of the differing views of biomedicine in its source vocabularies. The National Library of Medicine is adding source vocabularies and is developing automated systems to maintain, enlarge, and enhance future editions of the Metathesaurus in response to user evaluation and feedback.

REFERENCES

1. Webster's third new international dictionary of the English language, unabridged. Springfield, MA: G. & C. Merriam, 1969:3a.
2. ROGET PM. Thesaurus of English words and phrases classified and arranged so as to facilitate the expression of ideas and assist in literary composition. London: Longmen, Brown, Green, and Longmans, 1852.
3. TUTTLE M, SHERERTZ D, OLSON N, ERLBAUM M ET AL. Using Meta-1—the first version of the UMLS Metathesaurus℠. In: Miller RA, ed. Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care: standards in medical informatics, November 4-7, 1990, Washington, DC. New York: IEEE Computer Society Press, 1990: 131-5.
4. HUMPHREYS BL, LINDBERG DAB, HOLE WT. Assessing and enhancing the value of the UMLS knowledge sources. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care: assessing the value of medical informatics, November 17-20, 1991, Washington, DC. New York: McGraw-Hill, 1992:78-82.
5. MCCRAY AT, ARONSON AR, BROWN AC, RINDFLESCH TC ET AL. UMLS® knowledge for biomedical language processing. Bull Med Libr Assoc 1993 Apr;81(2):184-94.
6. HUMPHREYS BL, SCHUYLER PL. The Unified Medical Language System℠: moving beyond the vocabulary of bibliographic retrieval. In: Broering NC, ed. High performance medical libraries: advances in information management. Westport CT: Meckler Publishing. In press.
7. IBID.
8. KINGSLAND LC III, HARBOURT AM, SYED EJ, SCHUYLER PL. Coach℠: applying UMLS Knowledge Sources in an expert searcher environment. Bull Med Libr Assoc 1993 Apr;81(2): 178-83.
9. CIMINO C, BARNETT GO. Analysis of physician questions in an ambulatory care setting. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care: assessing the value of medical informatics, November 17-20, 1991, Washington, DC. New York: McGraw-Hill, 1992:995-9.
10. CIMINO, JJ. Representation of clinical laboratory terminology in the Unified Medical Language System℠. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care: assessing the value of medical informatics, November 17-20, 1991, Washington, DC. New York: McGraw-Hill, 1992:199-203.
11. WAGNER MM, COOPER GF. Evaluation of a Meta-1-based automatic indexing method for medical documents. Comput Biomed Res 1992 Aug;35(4):336-50.
12. CHUTE CG, YANG Y, EVANS DA. Latent semantic indexing of medical diagnoses using UMLS semantic structures. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care: assessing the value of medical informatics, November 17-20, 1991, Washington, DC. New York: McGraw-Hill, 1992: 185-9.
13. TUTTLE MS, SHERERTZ DD, ERLBAUM MS, SPERZEL WD ET AL. Adding your terms and relationships to the UMLS Metathesaurus®. In: Clayton PD, ed. Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care: assessing the value of medical informatics, November 17-20, 1991, Washington, DC. New York: McGraw-Hill, 1992:219-23.

Received October 1992; accepted November 1992

APPENDIX A

Source vocabularies in the 1992 Metathesaurus

ACR92 — Index for radiological diagnoses: including diagnostic ultrasound. Rev. 3d ed. Reston, VA: American College of Radiology, 1986.
AIR92 — AI/RHEUM. Bethesda, MD: National Library of Medicine, 1992.
COS92 — COSTAR (Computer-Stored Ambulatory Records) of [Massachusetts] General Hospital, 1992. (List of terms that occur frequently at certain clinical sites)
CPT89 — Physicians' current procedural terminology: CPT. 4th ed. Chicago: American Medical Association, 1989.
CST92 — COSTART: coding symbols for thesaurus of adverse reaction terms. Rockville, MD: Food and Drug Administration, Center for Drug Evaluation and Research, 1992.
CSP92 — CRISP thesaurus. Bethesda, MD: National Institutes of Health, Division of Research Grants, Research Documentation Section, 1992.
DSM3R — Diagnostic and statistical manual of mental disorders: DSM-III-R. 3d rev. ed. Washington, DC: American Psychiatric Association, 1987.
ICD91 — The international classification of diseases: 9th revision, clinical modification: ICD-9-CM. 4th ed. Washington, DC: Health Care Financing Administration, 1991.
INS92 — Thesaurus biomedical francais/anglais. Paris: Institut National de la Sante et Recherche Medicale, 1992. French translation of: Medical Subject Headings, vide infra.
LCH90 — Library of Congress subject headings. 12th ed. Washington, DC: Library of Congress, 1989.
MCM92 — List of epidemiology terms submitted by McMaster University.
MSH92 — Medical subject headings. Bethesda, MD: National Library of Medicine, 1992.
MTH — UMLS Metathesaurus.
NAN92 — Carroll-Johnson RM, ed. Classification of nursing diagnoses: proceedings of the 9th conference, March 1990, Orlando, FL. Philadelphia: Lippincott, 1991.
NIC92 — McClosky JC, Bulechek GM, eds. Nursing interventions classification (NIC): Iowa intervention project. St. Louis: Mosby-Year Book, 1992.
SNM2 — Cote RA, ed. Systematized nomenclature of medicine. 2d ed. Skokie, IL: College of American Pathologists, 1979.

UMS92 — Universal medical device nomenclature system: product category thesaurus. Plymouth Meeting, PA: ECRI, 1992.

APPENDIX B

Number of names and concepts in the 1992 Metathesaurus

Totals:

Identified Concepts:	130,137
Names from all Sources:	270,797

Number of names contributed by each source vocabulary:

ACR92	122	
AIR92	776	
COS89	776	
COS92	735	
CPT89	543	
CSP92	5,553	
CST89	2,548	
DSM3R	450	
DXP92	603	
ICD89	520	
ICD91	9,345	
INS92	16,640	
LCH90	5,094	
MCM92	43	
MSH92	213,355	(This includes 16,641 preferred terms and 115,940 supplementary chemical terms.)
MTH	1,159	
NAN90	100	
NIC92	905	
SNM2	11,418	
UMD91	112	