The UMLS Knowledge Sources: Tools for Building Better User Interfaces

Donald A. B. Lindberg, M.D. Betsy L. Humphreys

National Library of Medicine
Bethesda, MD

Abstract

The current focus of the National Library of Medicine's Unified Medical Language System (UMLS) project is the development, testing, and evaluation of the first versions of three new knowledge sources: the Metathesaurus, the Semantic Network, and the Information Sources Map. These three knowledge sources can be used by interface programs to conduct an intelligent interaction with the user and to make the conceptual link between the user's question and relevant machine-readable information. NLM is providing experimental copies of the initial versions of the UMLS knowledge sources in exchange for feedback on ways they can and should be improved. The hope is that the results of such experimentation will provide both immediate improvements in biomedical information service and useful suggestions for enhancements to the UMLS.

INTRODUCTION

The effective practice of medicine is dependent on the ability of health professionals to locate relevant information quickly and to interpret it correctly. Over the past 30 years a variety of computer databases and systems have been developed with the express purpose of assisting health professionals to identify and obtain pertinent information [e.g., 1, 2, 3, 4, 5]. For many health professionals, however, the use of machine-readable information sources is not yet perceived as sufficiently convenient, easy, and useful [6] to warrant even indirect access to such sources through a health sciences librarian or other intermediary. Fortunately, a number of forces are at work that may increase the perceived convenience and utility of automated information systems. They include: the increasing computer-literacy of those entering the health professions; the development of user-friendly search software for health professionals [7]; wider availability of machine-readable sources that provide evaluated information [2] and assist with diagnostic or treatment problems [3, 8]; the IAIMS initiative [9]; the emerging national research communications network; efforts to increase health professionals' awareness of currently available information services [10]; and the Unified Medical Language System (UMLS) project [11].

The Unified Medical Language System (UMLS) project is addressing the critical problem of establishing a conceptual link between the user's expression of an information need and relevant information in different machine-readable sources. The sources of interest include the biomedical literature, clinical records, factual databases, knowledge bases, and directories of information resources of various types. The variety of ways the same concepts are expressed in these different machine-readable sources and the lack of an efficient mechanism for determining which of many available databases are likely to contain information relevant to particular questions are severe deterrents both to the users of information systems and to the developers of interface programs that might aid these users. The UMLS approach expects continued diversity in the terminology employed in different automated systems and by the users themselves. The goal of the UMLS project is to develop products that can help to compensate for differences in terminology and therefore to make an effective conceptual connection between users and the machine-readable information they need.

The UMLS effort is a long-term project involving internal research and development at NLM, contracts with several university-based research groups and a commercial company [12], advice and input from private sector organizations coordinated by the American Medical Association, and collaboration with other governmental and related agencies, such as the newly created Agency for Health Care Policy and Research and the Institute of Medicine. The current focus of the project is the development, testing, and evaluation of the first versions of three new knowledge sources believed necessary to a fully functioning UMLS: the Metathesaurus [11,13], the Semantic Network [14, 15], and the Information Sources Map. The initial versions of these knowledge sources are designed primarily for use by system developers. They are meant to be consulted and used by search interface programs

0195-4210/90/0000/0121$01.00 © 1990 SCAMC, Inc.

to interpret and refine user queries, to select databases appropriate to the user's question, and to map the user's terms to appropriate controlled vocabularies and classification schemes. They are also useful as browsable reference tools for database builders and information professionals such as medical librarians. In a subsequent UMLS development phase, these knowledge sources will be incorporated into actual end-user searching systems such as GRATEFUL MED [7].

CONCEPTUAL OVERVIEW

The UMLS knowledge sources encompass three different but related types of information. Meta-1, the first version of the UMLS Metathesaurus, contains information about specific concepts, e.g., Acquired Immunodeficiency Syndrome, HIV-2, including the way they are expressed in a variety of vocabularies, how they have been used in selected databases, and the basic semantic categories or types, e.g., "Disease or Syndrome," "Virus," to which they belong. The UMLS Semantic Network has no information about specific concepts, but represents a variety of relationships among the semantic types or categories of concepts to which all individual terms in Meta-1 have been assigned (e.g., "Virus" CAN-CAUSE "Disease or Syndrome"). The UMLS Information Sources Map will contain information about databases as a whole (e.g., AIDSLINE, DXPLAIN): their content and coverage, vocabulary, location, access conditions, relationship to other databases, etc.

A simple example illustrates how the UMLS knowledge sources might be used by an interface program to help guide the interaction with a user and to make the conceptual link between the user's question and relevant machine-readable information.

* * * * *

A physician is interested in knowing the current status of research underway to test the efficacy of AZT in preventing the onset of AIDS in persons such as health care workers who are exposed to the disease via the blood of AIDS patients, but are not yet HIV-positive. Using a formatted screen search interface like GRATEFUL MED, she enters the two principal terms of interest:

AIDS and AZT

In Meta-1, AIDS appears as a synonym to "Acquired Immunodeficiency Syndrome" and AZT as an indexed word fragment of the synonym "AZT (Antiviral)". The search interface easily translates both terms to the MeSH headings: "Acquired Immunodeficiency Syndrome" and "Zidovudine." The interface also finds in Meta-1 the MeSH tree numbers of these two terms, their semantic types (i.e., "Disease or Syndrome" and "Pharmacologic Substance") and the fact that the two terms co-occur very frequently in MEDLINE citations.

Turning to the Information Sources Map, the interface program uses the MeSH tree numbers of the two terms and their semantic types to determine that there are at least three databases that appear to be particularly relevant to the user's inquiry. Using information from Information Sources Map records for these databases, the interface presents the following tailored screen to the user:

Available relevant information includes:
(1) recent citations to the literature and abstracts in AIDSLINE
(2) descriptions of ongoing clinical trails in AIDSTRIALS
(3) descriptions of drugs in AIDSDRUGS
Enter or point and click on the number(s) of interest.

At this point the physician selects 1 and 2.
As the AIDSLINE record in the Information Sources Map indicates that much of the database is derived from MEDLINE, the interface program knows that the terms "Acquired Immunodeficiency Syndrome" and "Zidovudine" will also co-occur very frequently in AIDSLINE. Drawing on information in the UMLS Semantic Network, the interface presents the following screen to the user to obtain guidance on narrowing the search:

You have selected terms for a "Pharmacologic Substance" (AZT) and a "Disease or Syndrome" (AIDS). Possible relationships between Pharmacologic Substances and Diseases or Syndromes are:

(1) TREAT (3) CAUSE
(2) PREVENT (4) COMPLICATE

Enter or point and click on the number(s) of particular interest.

Here the user selects 2. Based on the user's input, the interface can now construct the following MeSH search strategy:

***Acquired immunodeficiency syndrome/prevention & control AND *Zidovudine/therapeutic use**

which will retrieve 10 recent AIDSLINE citations relevant to the user's question and descriptions of 2

relevant clinical trials from AIDSTRIALS.

* * * * *

A basic UMLS hypothesis is that the UMLS knowledge sources can be used by search interface programs to interact more effectively with users and to retrieve a variety of machine-readable information relevant to user queries. We expect that the initial versions of these new knowledge sources will permit modest improvements in the retrieval and integration of machine-readable biomedical information. Feedback from experimentation with the early versions can guide the subsequent expansion and enhancement of these knowledge sources. While it will probably always be necessary for system developers to add some information of local importance to these sources, the goal is centrally maintained knowledge sources that can reduce the cost of developing effective medical information systems in a variety of contexts.

METATHESAURUS

The Metathesaurus is the central vocabulary tool of the UMLS. Its role is to facilitate the essential step of relating the user's terms to the variety of vocabularies and classifications used in biomedical databases. Meta-1 [16], the first version of the Metathesaurus, was issued for experimental use in the fall of 1990. It is a large (200+ megabyte) database encompassing 66,000 concepts and about 100,000 terms. Meta-1 was created through a combination of: automated processing and lexical matching of its source vocabularies [16], review and editing of the resulting preliminary records by subject experts [17], explicit labelling by subject experts of the hierarchical relationships between selected terms in NLM's MeSH (Medical Subject Headings) vocabulary, and computation of the occurrences of Meta-1 terms in several databases, e.g., MEDLINE, PDQ, DXPLAIN.

Meta-1 includes as its base vocabulary all terms in MeSH and the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, 3rd ed., revised (DSM-IIIR), a set of clinical terms heavily used at three COSTAR [18] ambulatory care sites, and a small set of terms for frequently ordered laboratory procedures. Terms from the College of American Pathologists' *Systemized Nomenclature of medicine* (SNOMED), the *International Classification of Disease's 9th edition. Clinical Modification* (ICD-9-CM), and the American Medical Association's *Current Procedural Terminology* 1989 edition (CPT) that could be related to this base set by automated lexical matching techniques are also included. Selected terms from the *Library of Congress Subject Headings* (LCSH) mapped to MeSH terms by the NLM staff also appear in Meta-1.

The objective of Meta-1 construction was to preserve the meanings and contexts of terms in these source vocabularies--not to describe the "complete" range of meanings or interterm relationships that might be applicable to each Meta-1 term. An accurate representation of how terms are used in different vocabularies and classifications is essential to the effective retrieval of information from databases that employ these terminologies. Expansion of the Metathesaurus to encompass a broader range of meanings and relationships will occur as additional vocabulary sources are incorporated in future editions.

NLM plans to issue new editions of the Metathesaurus at least annually. New versions will incorporate regular updates to MeSH, modifications to other source vocabularies, and the more extensive expansions and enhancements that will be made to the database over the next several years. Both the format and content of Meta-1 are considered experimental. If the feedback received from those attempting to use it indicates that major changes would be beneficial, Meta-2 may differ substantially in structure from its predecessor.

NLM's UMLS collaborators and advisers have already identified a range of possible enhancements and expansions that could be made to future versions of the Metathesaurus. Specific enhancements discussed include: (1) incorporation of additional broad groups of terms and concepts from vocabularies and classifications already partially represented in Meta-1, e.g., ICD-9-CM, (2) addition of terms from other vocabularies such as Universal Medical Device Nomenclature System; (3) addition of definitions to undefined Meta-1 terms (4) addition of more synonyms of Meta-1 terms and (5) expansion of coverage in specific clinical problem areas, such as those that are the focus of practice guideline development by the Agency for Health Care Policy and Research. We plan to base decisions on the order and priority of various enhancements on the experiences and feedback of those attempting to apply Meta-1 to a variety of information problems.

SEMANTIC NETWORK

The Semantic Network defines the likely relationships

among the types or categories of terms in the Metathesaurus. The first version of the Semantic Network was issued with Meta-1 in 1990. It contains records for each of the semantic types or categories to which the specific terms in Meta-1 were assigned. Each record contains the name of the type, the number of its location within the hierarchy of types, a definition, and an enumeration of its relationships to other semantic types within the network. The network represents relationships among types that are particularly significant and potentially useful in focusing information retrieval strategies. It does not attempt to include all possible relationships among its types.

The set of Semantic types assigned to concepts in Meta-1 was developed by a combination of top-down and bottom-up experiments involving NLM staff and all of the UMLS collaborators. The actual assignment of types to the concepts in Meta-1 served as a large scale test of the comprehensiveness of the set of types and of the clarity and precision of their definitions. During the editing process, additional refinements and changes in the names and definitions of the Meta-1 types were made.

Now that semantic types have been assigned to a large body of concepts, a variety of analyses are being conducted to review the reasonableness of the relationships among the types based on the specific concept names that are linked by those relationships. The results of these investigations will affect both the categorization of concepts in Meta-2 and the content of the second edition of the UMLS Semantic Network. The Semantic Network is in part an authority file for Semantic types used to categorize terms in the Metathesaurus. New versions of Network will therefore continue to be issued simultaneously with new versions of the Metathesaurus.

Although it remains to be proven, we believe that the Network's knowledge of the significant and useful relationships among the general categories of concepts represented in a user's query will be helpful in interacting with users and translating their information needs into effective search strategies. As with Meta-1, expansion of the Semantic Network will be based on feedback from those who attempt to use it.

INFORMATION SOURCES MAP

The third UMLS knowledge source, the Information Sources Map, is currently in the prototype development phase. The eventual goal is a knowledge source that can assist computer programs: (1) to determine which machine-readable information sources are likely to be relevant to a particular user inquiry; (2) to supply human-readable information to users about the scope, probable utility, and access conditions of information sources; (3) to make an automatic connection to sources likely to be relevant to a particular user inquiry; and (4) to conduct automatically a successful retrieval session on relevant information source(s). For information sources with specific term occurrence data in the Metathesaurus, a combination of information in the Metathesaurus and the Information Sources Map will be used to determine whether a particular information source is relevant. For information sources without corresponding term occurrence data in the Metathesaurus, the general scope information in the Information Sources Map will be used independently to determine whether a particular information source is likely to be relevant.

The prototype currently under development is designed to support selection of appropriate sources, display of useful information about the sources to the user, and automatic connection to the sources. For the command-driven sources represented, the prototype Information Sources Map will also provide the information needed for a program to conduct the actual search session. The prototype will contain records for a limited number of sources, including examples of bibliographic, factual, and full-text databases and knowledge bases. It will also contain records for the specific computer systems, telecommunications networks, and software necessary to support automatic connection to these sources.

Each Information Sources Map record will contain both human readable and computer processable descriptive information. Information sources will be described on a variety of levels including: type of information present (e.g., bibliographic citations, abstracts, descriptions of research protocols), subject scope (defined in terms of applicable MeSH categories, semantic types, and, in some cases, specific subject headings), language, dates of coverage (if applicable), nature of links and relationships to other databases, size, update schedule, actual data element definitions, etc. Each Information Sources Map record will also include scripts for use in automatic connection to the source it describes.

After the prototype has been tested and refined by NLM and its UMLS contractors, the first version of the Information Sources Map will be distributed for broader experimental use. Work will be needed to determine the most effective way to assist users in

selecting appropriate databases and to handle automatic interactions with these sources. The content and format of the Information Sources Map will evolve as such work progresses. It is hoped that the record format developed will be suitable for describing locally available sources and that system developers will be able to add local records to those that NLM will distribute. The Information Sources Map may be amenable to true distributed maintenance, with source producers taking responsibility for the currency and accuracy of the records for their own sources.

PLANS FOR FUTURE DEVELOPMENT

The initial versions of the UMLS knowledge sources represent a significant first step in the development of a total system that can facilitate the retrieval and integration of information from multiple machine-readable sources. To achieve the overall goals of the UMLS project, these knowledge sources must be linked to functional components or programs that can exploit them effectively. We expect that several implementations of these functional components will evolve, each tailored to particular computing or functional environments. NLM itself will develop these components within GRATEFUL MED and its associated expert search assistant, COACH. The knowledge sources must also be extended and enhanced as actual use and testing shows this to be desirable. To facilitate broad use and experimentation, NLM is providing experimental copies of the initial versions of the UMLS knowledge sources in exchange for feedback on ways they can and should be improved. Many research groups, medical schools, medical libraries, hospitals, and commercial companies have requested sample records and documentation for Meta-1 and the first version of the UMLS Semantic Network. We hope that the results of their experimentation will provide both immediate enhancements to biomedical information service and useful suggestions for improvements to UMLS products.

REFERENCES

1. Adams S, Taine S. Searching the medical literature. JAMA 1964 Apr 20;188(3):251-254.
2. Hubbard SM, Henney JE, DeVita CT Jr. A. computer database for information on cancer treatment N Engl J Med 1987 Feb 5;316(6):315-8.
3. Hupp JA, et al. DXplain--a computer-based diagnostic knowledge base. MEDINFO 86. Amsterdam:North-Holland, 1986:part 1,117-121.
4. Pryor TA, et al. The HELP system. J Med Syst 1983:787-102.
5. Miller RA, et al. The INTERNIST-1/QUICK MEDICAL REFERENCE project--status report. West J. Med 1986;145:816-822.
6. Shortliffe EH. Testing reality: the introduction of decision - support technologies for physicians. Methods Inf Med 1989 Jan; 28(1):1-5.
7. Haynes RB, et al. Online access to MEDLINE in clinical settings. A study of use and usefulness. Ann Intern Med 1990 Jan 1;112(1):78-89.
8. Miller PL, Black HR, Medical plan analysis by computer: critiquing the pharmacologic management of essential hypertension. Comput Biomed Res. 1984 Feb; 17(1):38-54.
9. Integrated Academic Information Systems (IAIMS) model development. Bull Med Libr Assoc 1988 Jul;76(3):221-67.
10. National Library of Medicine Board of Regents. Improving health professionals access to information: report of the Board of Regents. Bethesda, MD, 1989.
11. Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. In: Kingsland LC III, ed. Proceedings of the 13th annual SCAMC Washington, D.C. IEEE Computer Society Press, 1989;475-80.
12. UMLS contractors, June 1988-June 1991: Lexical Technology, Inc.; Massachusetts General Hospital with subcontractor, Brigham and Women's Hospital; University of Pittsburgh; University of Utah; Yale University School of Medicine.
13. Tuttle MS, et al. Using Meta-1 -- the First Version of the UMLS Metathesaurus. In: Miller RA, ed. Proceedings of the 14th annual SCAMC. Washington, D.C. 1990.
14. McCray AT. The UMLS Semantic Network. In: Kingsland LC III, ed. Proceedings of the 13th annual SCAMC. Washington, D.C. IEEE Computer Society Press, 1989;475-80.
15. McCray AT, Hole WT. The Scope and Structure of the First Version of the UMLS Semantic Network. In: Miller RA, ed. Proceedings of the 14th annual SCAMC. Washington, D.C., 1990.
16. Sherertz DD, et al. Source Inversion and Matching in the UMLS Metathesaurus. In: Miller RA, ed. Proceedings of the 14th annual SCAMC. Washington, D.C., 1990.
17. Sperzel WD, et al. Editing the UMLS Metathesaurus: Review and Enhancement of a Computed Knowledge Source. In: Miller RA, ed. Proceedings on the 14th annual SCAMC. Washington, D.C., 1990.
18. Barnett GO. The application of computer-based medical record systems in ambulatory practice. N Engl J Med 1984 June 21;310(25):1645-9.