Focus on The UMLS JAMIA

Technical Milestone ■

The Unified Medical Language System:

An Informatics Research Collaboration

BETSY L. HUMPHREYS, MLS, DONALD A. B. LINDBERG, MD, HAROLD M. SCHOOLMAN, MD, G. OCTO BARNETT, MD

Abstract In 1986, the National Library of Medicine (NLM) assembled a large multidisciplinary, multisite team to work on the Unified Medical Language System (UMLS), a collaborative research project aimed at reducing fundamental barriers to the application of computers to medicine. Beyond its tangible products, the UMLS Knowledge Sources, and its influence on the field of informatics, the UMLS project is an interesting case study in collaborative research and development. It illustrates the strengths and challenges of substantive collaboration among widely distributed research groups. Over the past decade, advances in computing and communications have minimized the technical difficulties associated with UMLS collaboration and also facilitated the development, dissemination, and use of the UMLS Knowledge Sources. The spread of the World Wide Web has increased the visibility of the information access problems caused by multiple vocabularies and many information sources which are the focus of UMLS work. The time is propitious for building on UMLS accomplishments and making more progress on the informatics research issues first highlighted by the UMLS project more than 10 years ago.

■ **JAMIA.** 1998;5:1–11.

Over the past 10 years the Unified Medical Language System (UMLS)*[1] has captured the time, talents, and attention of many informatics investigators from a broad range of disciplines. The project is focused on overcoming two important barriers to the development of information systems that can help health professionals make better decisions. These barriers are the disparity in the terminologies used in different information sources and by different users, and the sheer number and distribution of machine-readable information sources that might be relevant to any user inquiry. The UMLS supports the development of user-friendly systems that can effectively retrieve and integrate relevant information from disparate machine-readable sources. To accomplish this objective, the UMLS project has produced and widely disseminated four multipurpose knowledge sources designed for system developers: the Metathesaurus, the Semantic Network, the Information Sources Map, and the SPECIALIST Lexicon and associated lexical programs.[2] These knowledge sources have been tested and ap-

Affiliation of the authors: National Library of Medicine, Bethesda, MD (BLH, DABL, HMS); Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA (GOB).

Correspondence and reprints: Betsy L. Humphreys, MLS, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894. e-mail: ⟨blh@nlm.nih.gov⟩.

Received for publication: 8/15/97; accepted for publication: 9/16/97.

*Unified Medical Language System, UMLS, Metathesaurus, Medline, MeSH, and Grateful Med are registered trademarks of the National Library of Medicine.

plied in many different systems and environments. A substantial and growing body of published literature documents, discusses, and assesses the results of UMLS-related research and development.[3]

The UMLS has progressed as a large-scale distributed research and development project through a decade of immense change in information technology and health care delivery. Beyond its tangible products and its influence on informatics research and practice, the UMLS is an interesting case study in collaborative research and development. The purpose of this paper is to describe how the UMLS project was initiated, implemented, and managed and to discuss key factors that have contributed to its longevity, achievements, and unfinished business.

Genesis of the UMLS Project

> The objective of this program . . . is to solve what is the most fundamental barrier to the application of computers in medicine; namely, the lack of a standard language in medicine. We will attempt to build that vocabulary, a language that will cross between the biomedical literature and the observations on the patient, as well as the educational applications in the school, a language which allows those areas to be interrelated.[4]
>
> —Donald A. B. Lindberg, M.D., March 19, 1985

These words introduced the UMLS project to the U.S. Congress at the National Library of Medicine's FY 1986 Appropriations hearing. The statement reflected the first level of refinement of a concept that Dr. Lindberg brought with him when he assumed NLM's directorship in August 1984. The specific UMLS approach to solving the vocabulary problem would evolve over the next few years. In March 1985, NLM was seeking Congressional support for a long-term research and development program targeted at one underlying cause of the difficulty encountered by medical informaticians in their attempts to integrate advanced computer-based decision support into routine patient care. The task of building the UMLS—and especially of maintaining it—was described as exceeding what any academic department or collection of short-term research grants could be expected to achieve. The initiation, funding, and ultimately the maintenance of such a system were therefore proposed as a reasonable undertaking for a federal research agency.

The UMLS program was presented as complementary to and supportive of IAIMS: "A unified medical language system will also contribute greatly to the ultimate success of the Integrated Academic Information Management Systems (IAIMS). It is inevitable that the various IAIMS systems now under development will be hampered by the lack of such a language as they attempt to integrate and link existing information resources in the clinic and the hospital, the classroom, the library, the administrative center, and in remote networks and databases. . . ."[5]

The case for developing the UMLS proved persuasive, and Congress added $1 million to NLM's FY 1986 budget for UMLS-related research and development. That money was the first and only addition to NLM's base appropriation specifically earmarked for UMLS work. In FY 1986, NLM allocated a comparable amount to the UMLS effort from the related intramural research budget of the Lister Hill National Center for Biomedical Communications. The UMLS research budget has continued to be roughly $2 million annually, excluding the cost of the NLM personnel working on the project. Since the early 1990s, UMLS-related research has also supported some of the objectives of NLM's health services research information program and of the NLM initiative in High Performance Computing and Communications.

Selecting and Funding Extramural Collaborators

By the time Congress appropriated FY 1986 funds for the UMLS, an internal NLM UMLS management team[6] had been conducting background investigations and evaluating various options for organizing the project for about a year. The NLM team exemplified a key characteristic of UMLS research—its dependence on contributions from many disciplines, including medicine, biomedical science, medical informatics, computer science, library and information science, and linguistics. Team members brought complementary and synergistic views to the complex issues under discussion. When any member had to miss a project meeting, the lack of a key perspective tended to slow progress.

At the outset, NLM assumed that a strong set of external research collaborators would be needed to make progress in overcoming the fundamental barriers to information access that the UMLS intended to address. The mechanism for selecting and funding these collaborators was less certain. After considering various alternatives, the NLM UMLS team selected the "task-order" research contract as the most appropriate vehicle for funding extramural UMLS research collaborators. Task-order research contracts involve a series of research tasks that are defined and negotiated throughout the life of the contract. This mechanism permitted NLM and the collaborators jointly to adjust research questions and methods to take advantage of improved understanding of the problems involved or changes in the health information environment. In contrast, "regular" research and development contracts require up-front definition of objectives and methods at a level of specificity that was inappropriate to the distributed and evolutionary research effort envisioned, especially in the initial exploratory phase of the project. Grant mechanisms, including regular research grants and cooperative agreements, afforded NLM too little latitude to coordinate the efforts of multiple research collaborators.

There were some drawbacks to the task-order research contract mechanism for the university-based informatics research groups who were likely to bid on the UMLS contracts. In 1986, most medical informatics investigators were unfamiliar with the rules and regulations that governed research contracts of any type, and even their business offices were unfamiliar with task-order contracts. The task-order contract also limited the investigator's autonomy and raised the specter that NLM might issue tasks that did not fit the research interests and agendas of the successful bidders. The Library did not intend to do this, but the contract mechanism would have permitted it.

About a year after the UMLS concept was first presented to Congress, NLM issued a request for proposals (RFP) for task-order research support contracts". . . for the development of the logical models and structures required to create a UMLS and its related products."[7] The RFP reiterated that "the principal barrier to effective integrated access to biomedical information is the tremendous array of classification and representation schemes used in major information sources: the published biomedical literature, patient records, medically related data banks, and medical knowledge bases."[7] It also reflected additional refinement in NLM's thinking about the approach to be taken: "The solution to this fundamental medical information problem is the development of conceptual links among disparate classification schemes. . . . The practical result will be a metathesaurus, linking MeSH and other medically relevant thesauri, as well as related products that assist in the classification of and access to the medical information available in the wide range of information sources."[7]

The RFP outlined several areas in which tasks would be assigned to successful bidders. These included user information needs, sources of machine-readable information relevant to these needs, functional requirements for UMLS components, alternative structures and development approaches for these components, and tools to aid the research effort. Proposers were required to present their own ideas for research related to the overall UMLS goals and—incomprehensibly to many bidders—to propose a specific approach to a sample task that they probably would never be asked to carry out. Despite confusion about the goals of the UMLS and reservations about the contract mechanism, NLM received a number of excellent proposals in response to the RFP. Some of these reflected pre-existing work that had influenced NLM's thinking about the UMLS project, such as Massachusetts General Hospital's MicroMeSH,[8] a user-friendly graphical browser for the Medical Subject Headings vocabulary.

As is required for all NIH research contracts, the proposals were evaluated by a specially convened technical review group consisting primarily of non-government experts. Following a round of technical negotiations and related budget trimming, in August 1986 NLM awarded four two-year UMLS research and development contracts, involving seven different medical informatics research groups. These were the first of three rounds of competitively awarded task order research contracts issued specifically for general UMLS research and development; the others were awarded in 1988 and 1991. In 1992 and 1995, NLM issued smaller competitive purchase orders to facilitate application of the UMLS Knowledge Sources in additional institutions. Beginning in 1993, NLM has supported some UMLS-related research and development projects under research contracts and cooperative agreement grants issued as part of its High Performance Computing and Communications program. Through these various mechanisms, many investigators and institutions have received NLM funding for UMLS research.[9]

Since 1986 the UMLS project has benefited from the talents of a succession of medical informatics fellows at a number of institutions with NLM Research Training Grants. In some cases, UMLS work has significantly influenced their subsequent careers and research interests. Several who first worked on UMLS research as fellows in training programs went on to become principal investigators or co-investigators on subsequently awarded UMLS contracts or grants. As envisioned at the outset of the UMLS project, the UMLS components provide some of the infrastructure needed for integrated information systems that are the focus of IAIMS. A number of IAIMS institutions have participated directly in UMLS research.

Establishing a Framework for Collaboration

On September 12–13, 1986, the new UMLS contractors met in Bethesda with NLM's internal UMLS project team for the first of what became a series of general UMLS project meetings that were held at 6-

month intervals over the next 8 years. (These were supplemented from time to time by smaller sessions involving subsets of UMLS collaborators who were investigating particular topics.) Some thought they were beginning an exciting and important endeavor. Some wondered what they had gotten themselves into. The excitement and the uneasiness were not mutually exclusive. The NLM contingent was pleased with the initial outcome of the contracting process and eager to get on with substantive work on the UMLS project. They also knew that it was likely to be difficult to keep such a disparate and high-powered group headed in roughly the same direction—especially since the exact destination was not yet clearly defined.

Inevitably, the agenda included discussion of technical aspects of communication and collaboration among the participants. Even in 1986 the Internet was the reasonable choice for routine communication among UMLS collaborators, but some investigators at nearly all UMLS research sites became Internet e-mail users for the first time as part of their UMLS participation. Guidelines for exchanging machine-readable data between sites (at that time on diskettes) were also established. In that pre-Web era, the problem of finding common hardware and software platforms loomed large. The group of collaborators included users of MS-DOS personal computers, Macintosh personal computers, and Unix-based workstations. Several UMLS participants would soon use NLM funds to purchase "alien" machines so they could use applications developed by other participants. Some UMLS contractors were asked to use particular software packages so their results would be more compatible with research efforts at NLM. In at least a few cases, this approach would hamper a contractor's progress on the substance of a task.

There was also limited discussion of intellectual property rights, particularly in relation to contractors' existing intellectual property, such as software or knowledge bases, that might be applied to UMLS research. NLM and its collaborators shared an interest in ensuring that the collaborators' previously developed intellectual property did not automatically pass into the public domain by virtue of its use in the UMLS effort. However, the special rights in data clause in the UMLS contracts ensured that the content of the central UMLS components generated as a result of NLM-funded research would belong to the U.S. government and therefore be in the public domain.

The most complex task was to determine how to collaborate on the substance of UMLS research. By the end of the second UMLS contractors' meeting in March 1987, the general parameters for substantive collaboration had been established. In essence, each UMLS collaborator engaged in two types of work: tasks jointly developed with NLM and other UMLS collaborators to assist in defining, building, and testing central UMLS components and individually designed and motivated projects related to UMLS research goals and local research priorities. An example of a joint UMLS task was the definition of the initial set of semantic types and relationships for the UMLS Semantic Network.[10] All UMLS research groups participated in this activity, some by conducting specially focused studies to identify potential sets of semantic types and relationships, and all by reviewing and critiquing several draft sets of types and relationships. There were differences of opinion about most aspects of the network, including whether more types or more relationships would be needed. (The first version of the UMLS Semantic Network was released with 133 semantic types and 37 relationships.) One example of a UMLS project defined and carried out by an individual collaborator was the University of Pittsburgh's study of physicians' information needs arising in the context of medical rounds.[11] Undertaken to improve understanding of UMLS requirements, this project used ethnographic techniques to identify a wide range of information needs. Nearly half could be met by knowledge-based information sources or the synthesis of patient- and knowledge-based information. In another example of an individually defined project, NLM intramural staff developed a test collection of user queries and Medline citations evaluated for relevance to these queries for use in bibliographic retrieval experiments.[12]

The majority of UMLS contractor effort was devoted to individually defined research projects, but the levels of effort assigned to the two types of activities varied at different stages in the UMLS project and for different collaborators. For example, from 1989 on, Lexical Technology, Inc., focused almost entirely on research related to the design, development, and testing of the UMLS Metathesaurus and received a number of more specific competitive research contracts for this work. Because the first stage of the UMLS project was purposely devoted to exploring all potential approaches identified by UMLS collaborators, there was heavy initial emphasis on individually defined projects. As the general outlines of the UMLS Metathesaurus and Semantic Network began to emerge, all contractors devoted considerable effort to tasks that helped to define the initial structure and content of these components.

Distinctions between joint and individual projects blurred as the UMLS effort progressed. All participants contributed to the evolution and better definition of the UMLS objectives and methods and were influenced by the lively and provocative exchange of ideas and opinions at the UMLS project meetings. As a result, individual contractors designed projects that addressed local interests while also contributing directly to the development and testing of UMLS hypotheses. After the first edition of the UMLS Knowledge Sources was released in 1990,[13] UMLS collaborators funded by NLM focused significant effort on testing these components in a variety of applications. The goal was to assess their utility for various purposes, to generate feedback on desirable changes and enhancements to their structure and content, and to identify the need for any additional centrally developed UMLS tools. To cite one of many examples, the Yale School of Medicine's Psychtopix[14] provided an early successful demonstration of automated Medline searches to assist psychiatry residents in performing clinical consultations on the in-patient services. This project prompted the incorporation of vocabulary from the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental and Behavioral Disorders* (DSM) into the Metathesaurus and also tested the feasibility of linking special purpose interfaces to NLM's Grateful Med search engine.

Making Key Project Decisions

From project inception, NLM anticipated that the UMLS would result in some regularly distributed and maintained products, analogous to the Library's Medical Subject Headings (MeSH) or Medline. Primarily for this reason, NLM reserved for itself final decision-making power regarding the characteristics of centrally developed UMLS components. Decisions were often the result of an iterative process. First, NLM would select a basic approach following an assessment of elements of the work and recommendations of all UMLS participants. This would then be discussed, refined, and inevitably improved by all collaborators, including those who may have strongly favored a different basic approach. Fortunately, NLM was blessed with collaborators who combined persistence in pressing their recommendations, even in the face of obvious disagreement with NLM, with a willingness to apply their best efforts to improving and supporting NLM's final decisions, even if these did not coincide with their own preferences.

An important example of iterative decision-making was the initial definition of the general parameters of the UMLS Metathesaurus and Semantic Network. At the time these decisions were made, the work of investigators from the University of California, San Francisco (later from Lexical Technology, Inc.), on the use of automated lexical processing methods to exploit the content of existing sources of machine-readable medical information[15] was closely aligned with NLM's interest in scalable, maintainable approaches. Early work by other UMLS collaborators focused on the creation of new formal representations of concepts as potential "canonical forms" to which terms from existing vocabularies could be mapped.[16–18] The hypothesis that more robust machine-readable representations of medical concepts would assist in the retrieval and integration of information from disparate sources was (and is) attractive. In NLM's view, however, the methods tested in early UMLS work were not scalable to the extent needed for the UMLS Metathesaurus, nor was it clear that the result would necessarily be more generally acceptable than previous efforts to systematize medical concepts. A University of California, San Francisco, proposal that recommended a step-by-step approach to building the Metathesaurus starting with the computation of a lexicon of all words in selected biomedical vocabularies appeared more practical to NLM. When the discussion and refinement of this proposal by all UMLS collaborators was finished, the definition of the initial UMLS Metathesaurus included significantly more semantic content than was presented in the first step of the original proposal. The general parameters for a second UMLS Knowledge Source, the Semantic Network, had also been defined.

This exemplifies what has been a continuing debate between NLM and many of its UMLS collaborators regarding the amount of new intellectual content that should be created for the UMLS Metathesaurus and Semantic Network. In general, most UMLS collaborators favor more original content; NLM, with an eye to long term maintenance and resource issues, favors the minimum needed to achieve the UMLS goals. In line with its desire for iterative refinement of the UMLS Knowledge Sources, the Library is interested in more experiments to determine which content enhancements are likely to be most beneficial. Collaborators have contended, not unreasonably, that some experiments make no sense until additional content is available. They have been fairly consistent in recommending the creation of detailed clinical vocabulary expressly for the Metathesaurus and the addition of more specific semantic types and relationships to the Semantic Network for use in clinical systems.

A number of UMLS collaborators disagreed with NLM's early decision to define the scope of the Metathesaurus as essentially equal to the combined scopes of its source vocabularies. They argued, with good cause, that none of the machine-readable vocabularies available in 1988 provided adequate coverage of many of the detailed clinical concepts needed for patient records. However, NLM took the position that it should not address the development of extensive new controlled vocabulary expressly for the Metathesaurus until existing relevant vocabularies had been incorporated. Similarly, NLM could not itself undertake to develop the entire vocabulary necessary to support computer-based patient record systems throughout the U.S. The Library has, however, been able to encourage the assumption of ever greater responsibility in this area by the Department of Health and Human Services.

In the short term, NLM's approach led to early editions of the UMLS Metathesaurus with content seen as insufficient for a number of important clinical applications. The longer term impact is less certain. Some would contend that the informatics field would have been better served by the development of a new UMLS canonical representation of medical concepts. Others would say that the development and release of *SNOMED International* and the *Read Clinical Classification* and their incorporation into the Metathesaurus provide some validation for the Library's decision not to generate a new clinical vocabulary, as do the results of the recent NLM/AHCPR Large Scale Vocabulary Test.[19] Current work to implement the administrative simplification provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA), to which NLM is contributing, may also address the need for ongoing Federal support and coordination for the maintenance and distribution of clinical vocabulary[20]—which has been an important subtext in all UMLS discussions.

Most, if not all, UMLS decisions have been less controversial than those that defined the basic characteristics of the Metathesaurus and the Semantic Network, but others have also illustrated both the difficulties and the benefits of multidisciplinary research. This is particularly true of many decisions related to the content, format, and maintenance systems for the UMLS Metathesaurus. In general, NLM and Lexical Technology, Inc., have been chiefly involved in making these decisions, although many changes to Metathesaurus content and format have been based on suggestions from other UMLS collaborators. Format changes have also been routinely circulated to the larger group for comment and revision before final implementation.

Staff at NLM and Lexical Technology, Inc., almost always agree on the end goals for Metathesaurus maintenance. They sometimes differ on immediate tradeoffs between the simplicity and maintainability of underlying software systems, the need to reduce the cognitive load on the human editors who are responsible for final review of Metathesaurus content, the desire to add substantial additional content to the Metathesaurus each year, and the importance of meeting release schedules. When disagreements arise, Lexical Technology, Inc., represents the advanced computer science perspective; the NLM viewpoint reflects experience in the development of software systems to support knowledge workers and the production and distribution of database products, as well as a concern for face validity. The different perspectives can lead to miscommunication. In one memorable early misunderstanding, Lexical Technology, Inc., assumed that "unique identifier" had a limited technical meaning, whereas NLM was in fact advocating permanent context-free identifiers for concepts in the Metathesaurus. When the smoke clears, the interplay of the two viewpoints has usually achieved a result acceptable to both sides and often better than either's original position.

The UMLS and the Advance of Information Technology

An important assumption underlying the UMLS effort is that "information systems must be used if they are to improve. To ensure UMLS components get actual use as soon as possible, they will be developed through a series of successive approximations of the capabilities ultimately desired. Early versions of UMLS components will be relatively simple structures, offering modest enhancements to current systems with respect to their representation of the interrelationships among biomedical terms and concepts. Complexity will be added in subsequent versions as actual use shows it to be necessary. To facilitate broad use and feedback, all versions of the UMLS components will be distributed in formats compatible with a wide variety of hardware and software."[21] From the first edition in 1990, the UMLS Knowledge Sources have been available free-of-charge to all interested domestic and international users, from 1990–96 under the terms of an experimental agreement and beginning in 1997 under a regular license agreement. As previously explained, NLM has also provided funding for UMLS applications via a variety of mechanisms.

The dual strategy of free UMLS distribution to all interested parties and targeted funding of focused research and development efforts has been followed since 1990. It has become increasingly successful as the information technology available to system developers has improved, as NLM has used new technology to make the UMLS Knowledge Sources more accessible, and as the content of the UMLS Metathesaurus and lexical tools has matured. Although the term is newer than the project, in essence the objective of the UMLS effort is to build ''middle-ware'' that enables advanced capabilities in many different health information systems. Until quite recently, the UMLS project was building this middle-ware for a future that had not yet arrived—a future with substantial growth in clinical information systems, a sharp reduction in hardware and software compatibility issues, an explosion in machine-readable knowledge-based information sources, and increasing availability of high-speed computing and communications capabilities. Before this future became reality for appreciable numbers of system developers, it was hard to grasp the goals of the UMLS project and even harder to build, deploy, and test prototype applications of the UMLS Knowledge Sources.

In this regard, as in many others, early difficulties faced by the UMLS contractors were excellent predictors of problems that would be even more acute in the general informatics community. With the software systems and techniques then in use, it was difficult, if not impossible, to test the use of UMLS Knowledge Sources in combination with existing operational systems. As already mentioned, at the outset of the project, hardware and software incompatibility was a significant barrier to many types of collaboration among UMLS participants. Many sites did not yet have local area networks that could be used to deploy resource-intensive UMLS applications to user sites. Few institutions had Internet connections that would support efficient use of applications on remote servers. Hardware and software incompatibility was often an issue for those who did have good Internet access. These factors discouraged many potential UMLS application developers. They also hampered, or even precluded, meaningful testing at other sites of the prototype applications developed by UMLS contractors.

The UMLS contractors correctly predicted that the size, complexity, and unfamiliarity of the UMLS Knowledge Sources would also discourage use. Some collaborators thought that NLM should not release the Knowledge Sources without an associated set of tools that facilitated their use. Here again hardware and software platform issues complicated the issue. In the early days of the UMLS project, platform independent code was ''technically possible'' but often not practically implementable. The Library was reluctant to devote resources to developing tools for multiple platforms, especially given the difficulty of predicting what tools would be needed and the certainty of significant changes to the format of the UMLS Knowledge Sources. Early attempts to apply the Metathesaurus showed that its initial relational format was unclear and unwieldy. The relational format was therefore substantially simplified for the 1992 edition.[22] All early users of the Metathesaurus devoted considerable time and effort to building indexes to its terms. These varied widely in quality and sophistication, and they naturally led to experimental results that were inconsistent and not comparable. To address this problem, in 1994 NLM began to distribute word, normalized word, and string indexes with the Metathesaurus, along with the SPECIALIST lexicon and lexical programs that were used to generate these indexes.[23] By applying the same tools used to construct the Metathesaurus indexes to any input term, system developers optimize their chances of linking external terms to related information in the Metathesaurus. By themselves and in combination with the Metathesaurus and Semantic Network, the SPECIALIST lexicon and lexical programs are powerful tools for natural language processing. Their inclusion as part of the UMLS Knowledge Sources was strongly recommended by linguists in the UMLS research group at Columbia University.

While work was proceeding on streamlining the format of the UMLS Metathesaurus, expanding its content, and adding lexical resources to the UMLS Knowledge Sources, access to the Internet and then the World Wide Web and its platform-independent browsers was increasing. These developments have both heightened interest in the UMLS Knowledge Sources and simplified their use. The Web provides a readily accessible vehicle for distribution of current UMLS fact sheets and documentation, and many new UMLS users now ''discover'' the UMLS on the Web. Internet access to the UMLS Knowledge Source Server[24] offers an easy way for users to explore the content and format of the UMLS Knowledge Sources, to download subsets for incorporation in local software, or to embed access to NLM's server in local applications. The Library also continues to distribute all UMLS files on CD-ROMs. Some users still do not have good Internet connections, and even those with high-speed Internet access may be reluctant to ftp more than 800 megabytes of Metathesaurus files.

Fortunately, the large and growing size of the UMLS Metathesaurus has become much less problematic to the deployment of end user applications. One illustration of this phenomenon is NLM's ability to make the advanced search capabilities originally developed for the DOS-based Coach expert search assistant[25] available to a much broader audience through Internet Grateful Med.[26] Medline users who loaded and tested the Coach application on individual DOS workstations or LAN servers were enthusiastic about its functionality; largely because it incorporated information from the Metathesaurus, however, it was just too big to be mounted at most sites. With Internet

Grateful Med, the large Metathesaurus files and advanced search functionality reside on high-performance servers at NLM, and the user needs only a Web-capable workstation and browser. New versions can be made instantly available to all users. Hundreds of people assisted in beta-testing Internet Grateful Med. Such broad testing was prohibitively resource-intensive when copies of software had to be distributed and loaded at many locations. In the current favorable environment, a growing number of significant Web applications make use of the UMLS Metathesaurus. Examples include Internet DXplain,[27] CliniWeb,[28] and Medical World Search.[29]

Once thorny technical issues associated with connecting and interacting with disparate machine-readable information sources have been simplified by Java and Web technology, Web-based interfaces to legacy systems, and the rapid growth in health-related information on the Web. Internet developments highlight and increase the importance of achieving the UMLS objective of helping users to locate and retrieve relevant information from the sea of available sources. The Internet has also encouraged new approaches to scanning and retrieving potentially relevant information. The UMLS collaborators began work on an Information Sources Map of human-readable and machine-interpretable descriptions of online information sources before the dramatic rise in Internet connectivity or the invention of the Web,[30,31] but they were quick to see that the new developments offered great potential for progress on the problems the UMLS Information Sources Map was designed to address.[32] Nonetheless, these dramatic changes have had a temporarily disruptive effect on UMLS efforts to define a method of describing available machine-readable information sources that would facilitate automated selection and retrieval of information from relevant sources. Staff at NLM have been re-examining their assumptions about description and access to machine-readable information sources as they and many others explore the ramifications of the explosive growth in Web-based information.

As advances in information technology improved the prospects for building, testing, and deploying some UMLS applications and changed the nature of the problem for others, they also facilitated the development and maintenance of the Metathesaurus. The chief parties involved in building the Metathesaurus reside on opposite coasts of the United States—NLM in Washington, DC, and Lexical Technology, Inc., in Alameda, California. For a two-site project the size of Metathesaurus construction, the logical approach was to select and implement a common platform for the underlying maintenance system. There was essentially no debate about the use of relational database technology and Unix machines for the back end of Metathesaurus construction. Ingres was easily chosen as the relational database management software based on the experience that Lexical Technology, Inc., had with this software. When NLM began development of the interface for Metathesaurus editors, a fourth-generation language (Windows 4GL) was selected as providing a basis for editing from different types of workstations and for accommodating the eventual migration of the back end system to a different relational software package. Given the distance between the collaborators and the interdependence of tasks undertaken at each site, reliable high-speed communications rapidly became essential to building the Metathesaurus. In early 1993, Lexical Technology, Inc., installed a T1 connection to allow more efficient transfer of initial source vocabularies and updates, preliminary Metathesaurus records, and final edited data to and from NLM, as well as to facilitate shared remote access to computing resources. Advanced communications have also allowed NLM to support the Metathesaurus editing interface at distributed sites, although to date only on a limited scale. The NLM/AHCPR Large Scale Vocabulary Test recently demonstrated the current potential for wider distribution of tasks associated with the development and maintenance of vocabularies. Building upon the capabilities of the UMLS Knowledge Source Server, a special test interface allowed more than 60 participants to search more than 41,000 terms in existing controlled vocabularies and to submit the results to NLM in a standard format during a 5-month period.[33]

The Impact of the UMLS

The published literature documents that the UMLS has helped to shape the medical informatics research agenda since 1986.[3] This is due in large part to its budget, which has funded work by more than 100 U.S. investigators from many disciplines, but there are other contributing factors. The UMLS objectives continue to be compelling. Technological advances have improved the methods for investigating UMLS research questions without diminishing the importance of these questions. Over the years, the UMLS Knowledge Sources have matured into significant research and development tools. The UMLS project has also offered the incentive and opportunity for stimulating inter-institutional collaboration. Many informaticians (e.g., Evans et al.[34]) received their first exposure to the rewards and frustrations of substantive research collaboration as part of UMLS participation and have extended their collaboration outside the UMLS project.

Since 1990 the UMLS project has produced annual editions of tangible products that are now regularly used by their intended audience. Although the value of the UMLS products must be assessed by more disinterested observers, an increasing array of operational systems employ one or more of the UMLS Knowledge Sources or lexical programs. As of August 1997, about 500 institutions and individuals around the world had signed the new license agreement required to receive the 1997 edition of the UMLS Knowledge Sources; NLM continues to receive new signed agreements each week. The Library itself is a significant UMLS user, both in its production information retrieval services and its research programs.

The number of UMLS users engaged in the development of commercial clinical information systems is substantial and growing. HL7 has recently selected the UMLS Metathesaurus as an appropriate vehicle for recording and distributing its planned decisions about the vocabularies that are valid for specific parts of the HL7 clinical messaging standard.[35] These two developments reflect the fact that the Metathesaurus provides access to a large and increasing number of important vocabularies in a common and explicit database format. The 1997 version contains 331,756 biomedical concepts named by 739,439 different terms from more than 30 source vocabularies. The Metathesaurus obviously builds on the strengths of its source vocabularies. Some source vocabulary producers have in turn used feedback from Metathesaurus construction or connections present in the Metathesaurus to enhance the format and content of their terminologies. The next generation of the Metathesaurus maintenance system[36] should provide a better automated infrastructure for symbiotic relationships between the Metathesaurus and the vocabularies it encompassses.

Differences of opinion about its purpose and utility aside, the development of the UMLS Metathesaurus has increased interest in controlled clinical vocabulary. It has expanded understanding of desirable vocabulary features, including concept organization, multiple hierarchical perspectives, and unique concept identifiers without embedded meaning. The UMLS project has also raised consciousness about the need to represent changes in vocabularies explicitly and about the problems associated with keeping local systems synchronized when changes to externally developed vocabularies occur.[37]

Despite these accomplishments, some UMLS objectives are yet to be achieved. The overall goal of the UMLS is to make it easier to develop sophisticated information systems that can help users retrieve and integrate relevant biomedical information from disparate machine-readable sources. Ten years ago this was hard to explain and justify. Today the retrieval problems encountered when searching many different information sources are familiar to every Web user and a major focus of Digital Library research. The UMLS project can claim credit for its early recognition of the fact that advances in computing and communications technology would increase the importance of effective retrieval from multiple databases. In addition, UMLS research has made progress on some of the many research issues associated with interpretation of user queries, mapping between the language of different information sources, and medical natural language indexing and retrieval techniques. After a promising start, there has been less progress in determining the extent to which machine-interpretable descriptions of information sources are needed for effective identification and retrieval of information from multiple information sources—and in deciding how such descriptions should be structured, created, and maintained. This is the object of current research at NLM, at other health sciences institutions, and in the general library and information science community.

The UMLS has placed particular emphasis on developing the ability to retrieve and integrate knowledge-based information that is directly relevant to the patient conditions described in an automated clinical record. Much of the serious investigation and prototype system development involving links between automated patient data and knowledge-based information has been performed under the aegis of the UMLS,[38] often using UMLS components. This important work notwithstanding, no one associated with the UMLS would claim that all the important research issues surrounding the effective linking of clinical and knowledge-based information have been explored, let alone that such capabilities have reached the mainstream of clinical information systems.

Fortunately, the environment for making progress on UMLS objectives has never been better. Advanced computing and communications capabilities are cheaper and more generally available. The Web environment has eliminated many of the technical problems that previously slowed research and development on retrieval of information from multiple information sources. Interest in extracting information from many Internet-accessible information sources has yielded new search methods and approaches. Computerized patient record systems are beginning to reach the point where they contain enough clinical data to enable robust links to knowledge-based information sources. The content of the UMLS Metathesaurus, Semantic Network, SPECIALIST lexicon, and lexical programs has expanded to be more applicable to the problem of linking clinical and knowledge-based information, although more concerted use of and feedback on these tools are still needed. The time is propitious for renewed and expanded collaboration on the medical informatics research problems first highlighted by the UMLS project a decade ago.

The authors thank their many UMLS colleagues for their contributions to UMLS accomplishments—and for the pleasure and inspiration of their company.

References ■

1. Lindberg, DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Meth Inf Med. 1993;32:281–91.
2. Current descriptions, documentation, and information about obtaining the UMLS Knowledge Sources are available from NLM's Web site: www.nlm.nih.gov
3. For a comprehensive bibliography of papers published from 1986–96 on UMLS-related work, see Selden CR, Humphreys BL. Unified Medical Language System. Current Bibliographies in Medicine. 1997;8:96.
4. Departments of Labor, Health and Human Services, Education, and Related Agencies Appropriations for 1986: Hearings Before the Subcommittee on the Departments of Labor, Health and Human Services, Education, and Related Agencies of the House Committee on Appropriations, 99th Cong., 1st Sess. Part 4B, (857) (1985) (statement of Dr. Donald A. B. Lindberg, Director of the National Library of Medicine).
5. Departments of Labor, Health and Human Services, Education, and Related Agencies Appropriations for 1986: Hearings Before the Subcommittee on the Departments of Labor, Health and Human Services, Education, and Related Agencies of the House Committee on Appropriations, 99th Cong., 1st Sess. Part 4B, (896-7) (1985) (statement of Dr. Donald A. B. Lindberg, Director of the National Library of Medicine).
6. In addition to the first three authors, over the life of the project NLM's UMLS management team has included: *William T. Hole, MD, *Lawrence C. Kingsland, III PhD, Daniel R. Masys, MD, *Alexa T. McCray, PhD, *Stuart Nelson, MD, Roy Rada, MD, PhD, *R. P. C. Rodgers, MD, and Peri L. Schuyler, MLS. Those preceded by asterisks are current members of the team. Many other NLM staff members, particularly in the Library Operations Division's Medical Subject Headings Section and the Lister Hill Center's Cognitive Sciences and Computer Sciences Branches, have made substantial contributions to the development of the UMLS Knowledge Sources and lexical programs and to NLM's applications of these tools.
7. National Library of Medicine. Statement of Work. In: Request for Proposals for Research Support for the Unified Medical Language System. March 28, 1986.
8. Lowe HJ, Barnett GO. MicroMeSH: a microcomputer system for searching and exploring the National Library of Medicine's Medical Subject Headings (MeSH) vocabulary. Proc Annu Symp Comput Appl Med Care. 1987;717–20.
9. The institutions that have received one or competitive contracts for general UMLS research and development are: Brigham & Women's Hospital (Robert A. Greenes, MD, PhD, PI), Carnegie–Mellon University (David A. Evans, PhD, PI), Columbia University (James J. Cimino, MD, PI), Lexical Technology, Inc. (Mark S. Tuttle, PI), Massachusetts General Hospital (G. Octo Barnett, MD, PI), the University of California, San Francisco (Marsden S. Blois, MD, PhD, PI), the University of Pittsburgh (Randolph A. Miller, MD, PI), the University of Utah (Homer R. Warner, MD, PhD, PI), and Yale School of Medicine (Perry A. Miller, MD, PhD, PI). Lexical Technology, Inc., has also received a number of competitive contracts for research and development related to the construction of the UMLS Metathesaurus. Additional institutions that received one or more competitive purchase orders for specific UMLS applications are: American Lake Biomedical Research Institute (Kenric W. Hammond, MD, PI), Chicago Medical School (David Trace, MD, and Frank Naeymi-Rad, PhD, PIs), Harvard University (Elizabeth Wu, MLS, PI), Georgetown University Medical Center (Naomi Broering, MLS, PI), Johns Hopkins University (Edwin B. George, MD, PhD, and Kevin Johnson, MD, PIs), State University of New York, Buffalo (John Eisner, DDS, PI), University of Missouri, Columbia (E. Andrew Balas, MD, PhD, PI), University of Washington (Sherrilynne Fuller, PhD, and Debra Ketchell, MLS, PIs). The following institutions have done UMLS-related research under competitive contracts or cooperative agreement grants issued as part of NLM's High Performance Computing and Communications initiative: Beth Israel Hospital (Charles Safran, MD, PI), Columbia University (James J. Cimino, MD, PI), Indiana University (Clement J. McDonald, MD, PI), Kaiser Permanente (Simon Cohn, MD, MPH, PI), Mayo Foundation (Christopher G. Chute, MD, Dr PH), Oregon Health Sciences University (William R. Hersh, MD, PI), University of Pittsburgh (Henry J. Lowe, MD, PI). The University of Maryland, Baltimore (Gary Freiburger, MLS, PI) received an NLM information systems grant for UMLS-related work.
10. McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. Proc Annu Symp Comput Appl Med Care. 1990; 126–30.
11. Osheroff JA, Forsythe DE, Buchanan BG, Bankowitz RA, Blumenfeld BH, Miller RA. Physicians' information needs: analysis of questions posed during clinical teaching. Ann Intern Med. 1991;14:576–81.
12. Schuyler PL, McCray AT, Schoolman HM. A test collection for experimentation in bibliographic retrieval. Medinfo. 1989;6(Pt 2):910–2.
13. The Metathesaurus and Semantic Network were first issued in 1990, the Information Sources Map in 1991, and the SPECIALIST Lexicon, lexical programs, and indexes to the Metathesaurus in 1994. Each has been updated annually since its introduction.
14. Powsner SM, Miller PL. Automated online transition from the medical record to the psychiatric literature. Methods Inf Med. 1992;31:169–74.
15. Sherertz DD, Tuttle MS, Blois MS, Erlbaum MS. Intervocabulary mapping within the UMLS: the role of lexical processing. Proc Annu Symp Comput Appl Med Care. 1988; 201–6.
16. Cimino JJ, Barnett GO. Automated translation between medical terminologies using semantic definitions. MD Comput. 1990;7:104–9. Published erratum appears in MD Comput. 1990;7:268.
17. Masarie FE, Jr, Miller RA, Bouhaddou O, Giuse NB, Warner HR. An interlingua for electronic interchange of medical information: using frames to map between clinical vocabularies. Comput Biomed Res. 1991;24:379–400.
18. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA. Conceptual modeling for the Unified Medical Language System. Proc Annu Symp Comput Appl Med Care. 1988; 148–51.
19. Humphreys BL, McCray AT, Cheh ML. Evaluating the coverage of controlled health data terminologies: report on the results of the NLM/AHCPR Large Scale Vocabulary Test. J Am Med Inform Assoc. 1997;4:483–97.

20. For a summary of current activities related to vocabulary see the report of the HIPAA Codes and Classifications Implementation Team at the following address: http://www.va.gov/meetings/hhs970709/007/Index.htm
21. Humphreys BL, Lindberg DAB. Building the Unified Medical Language System. Proc Annu Symp Comput Appl Med Care. 1989;475–80.
22. Tuttle MS, Sperzel WD, Olson NE, et al. The homogenization of the Metathesaurus schema and distribution format. Proc Annu Symp Comput Appl Med Care. 1992;299–303.
23. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc Annu Symp Comput Appl Med Care. 1994;235–9.
24. McCray AT, Razi AM, Bangalore AK, Browne AC, Stavri PZ. The UMLS knowledge source server: a versatile Internet-based research tool. Proc AMIA Fall Symp 1996;164–8.
25. Kingsland LC, 3d, Harbourt AM, Syed EJ, Schuyler PL. Coach: applying UMLS knowledge sources in an expert searcher environment. Bull Med Libr Assoc. 1993;81:178–83.
26. Lowe HJ, Lomax EC, Polonkey SE. The World Wide Web: a review of an emerging Internet-based technology for the distribution of biomedical information. J Am Med Inform Assoc. 1996;3:9.
27. Information about DXplain may be obtained from the Laboratory of Computer Science, Massachusetts General Hospital, at the following address: http://www.lcs.mgh.harvard.edu
28. Hersh WR, Brown KE, Donohoe LC, Campbell EM, Horacek AE. Cliniweb: managing clinical information on the World Wide Web. J Am Med Inform Assoc. 1996;3:273–80.
29. The Web address of Medical World Search is http://www.mwsearch.com
30. Masys DR, Humphreys BL. Structure and function of the UMLS Information Sources Map. Medinfo. 1992;7:1518–21.
31. Miller PL, Wright LW, Frawley SJ, Clyman JL, Powsner SM. Selecting relevant information resources in a network-based environment: the UMLS Information Sources Map. Medinfo. 1992;7:1512–7.
32. Rodgers RPC. Automated retrieval from multiple disparate information sources: the World Wide Web and the NLM's Sourcerer project. J Am Soc Inf Sci. 1995;46:755–64.
33. McCray AT, Cheh ML, Bangalore AK, et al. Conducting the NLM/AHCPR Large Scale Vocabulary Test: a distributed Internet-based experiment. Proc AMIA Fall Symp. 1997; 560–4.
34. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DS, for the Canon Group. Toward a medical concept representation language. J Am Med Inform Assoc. 1994;1:207–17.
35. Hammond WE. Call for a standard clinical vocabulary. J Am Med Inform Assoc. 1997;4:254–5.
36. Suarez-Munist ON, Tuttle MS, Olson NE, et al. MEME II supports the cooperative management of terminology. Proc AMIA Fall Symp. 1996;84–8.
37. Campbell KE, Cohn SP, Chute CG, Rennels G, Shortliffe EH. Galapagos: computer-based support for evolution of a convergent medical terminology. Proc AMIA Fall Symp. 1996; 269–73.
38. For a review article discussing work done by many research groups see Cimino JJ. Linking patient information systems to bibliographic resources. Meth Inform Med. 1996;35:122–6.