Original Investigations

JAMIA

Research Paper ■

Lessons Learned from a Pilot Implementation of the UMLS Information Sources Map

PERRY L. MILLER, MD, PhD, SANDRA J. FRAWLEY, PhD, LAWRENCE WRIGHT, MA, NANCY K. RODERER, MLS, SETH M. POWSNER, MD

Abstract **Objective:** To explore the software design issues involved in implementing an operational information sources map (ISM) knowledge base (KB) and system of navigational tools that can help medical users access network-based information sources relevant to a biomedical question.

Design: A pilot biomedical ISM KB and associated client-server software (ISM/Explorer) have been developed to help students, clinicians, researchers, and staff access network-based information sources, as part of the National Library of Medicine's (NLM) multi-institutional Unified Medical Language System (UMLS) project. The system allows the user to specify and constrain a search for a biomedical question of interest. The system then returns a list of sources matching the search. At this point the user may request 1) further information about a source, 2) that the list of sources be regrouped by different criteria to allow the user to get a better overall appreciation of the set of retrieved sources as a whole, or 3) automatic connection to a source.

Results: The pilot system operates in client-server mode and currently contains coded information for 121 sources. It is in routine use from approximately 40 workstations at the Yale School of Medicine. The lessons that have been learned are that: 1) it is important to make access to different versions of a source as seamless as possible, 2) achieving seamless, cross-platform access to heterogeneous sources is difficult, 3) significant differences exist between coding the subject content of an electronic information resource versus that of an article or a book, 4) customizing the ISM to multiple institutions entails significant complexities, and 5) there are many design trade-offs between specifying searches and viewing sets of retrieved sources that must be taken into consideration.

Conclusion: An ISM KB and navigational tools have been constructed. In the process, much has been learned about the complexities of development and evaluation in this new environment, which are different from those for Gopher, wide area information servers (WAIS), World-Wide-Web (WWW), and MOSAIC resources.

■ *J Am Med Informatics Assoc.* 1995;2:102-115.

Affiliation of the authors: Yale University School of Medicine, New Haven, CT.

Supported in part by NIH Contract N01 LM13537 and NIH Grant G08 LM05366 from the National Library of Medicine, and by an equipment grant from the Sun Microsystems Corporation.

Correspondence and reprints: Perry L. Miller, MD, PhD, Center for Medical Informatics, Yale University School of Medicine, New Haven, CT 06510. e-mail: perry_miller@yale.edu

Received for publication: 7/19/94; accepted for publication: 11/03/94.

There are a large and rapidly growing number of network-based information sources relevant to biomedicine. These sources provide information of many different types, including bibliographic citations, full text, scientific facts, clinical and research images, clinical expert advice, and computer-assisted instruction. Certain of these sources are provided within an institution's local area network. Others are provided internationally via the Internet, either free of charge or on a subscription basis. Since the set of available

sources is constantly growing and changing, there is no way that a user can personally keep up to date with all the sources potentially relevant to a biomedical question.

To assist students, clinicians, researchers, and staff in locating and accessing network-based information sources, the National Library of Medicine (NLM) is developing a biomedical information sources map (ISM) as part of its multi-institutional Unified Medical Language System (UMLS) project.¹ To help explore the many design issues involved in building and using such an ISM, we have implemented a pilot version of the UMLS ISM knowledge base (KB), with associated client-server software (ISM/Explorer), which is operational at Yale and currently contains information about 121 sources.² (For simplicity, this article uses the term ISMS to refer to the whole ISM system: the ISM KB and the ISM/Explorer software.)

This article describes the current status of the pilot ISMS, a number of underlying design issues, and preliminary experience with the ISMS's use. The article then focuses particular attention on the lessons learned in the process of building the pilot ISMS.

Background

It is clear that the biomedical user will need assistance in finding and accessing the rapidly growing set of biomedical information sources. The ISMS represents one approach to providing this help. There have been a number of other approaches taken to facilitate user access to network-based information sources.

1. One approach has been to develop a unifying front end to multiple information sources. Examples of this approach include BioSYNTHESIS,³ Willow,⁴ the Knowledge Workstation,⁵ and NetMenu.⁶
2. A more difficult task is to provide integrated automated querying of a variety of information sources. The Intelligent Query Workstation⁷ is an example of this approach, which accommodates the existing interfaces of different sources. The development of the Z39.50 protocol⁸ provides a vehicle to unify the functionality of multiple information servers, so that a single client can retrieve information from many sources in a uniform and potentially integrable fashion.
3. There are many activities in the national networking and library communities, not focusing specifically on the health sciences, that are dealing with Internet directories and resource discovery.⁹ Approaches to assisting the user in network navigation and resource discovery include Gopher, the World-Wide-Web (WWW) and its MOSAIC inter-

face, and wide area information servers (WAIS).^{10,11} Examples of directories of Internet-based databases and resources include the Internet Resources Meta-Index maintained by the National Center for Supercomputing Applications (NCSA) on WWW, the InterNIC Directory maintained by AT&T, and the OCLC Catalog of Internet Services.¹²

Later in the article, we discuss how the approach taken by the ISMS differs from certain of these approaches.

An Overview of the Pilot ISMS's Design

Figure 1 gives a schematic overview of the two major components of the pilot ISMS's system architecture: the ISMS server and the ISMS client.

1. The pilot ISMS server is written in LISP and runs on a Sun SparcStation-2 (Sun Microsystems, Mountain View, CA) at the Yale Center for Medical Informatics. It contains the ISM KB.
2. The ISMS client is written using the Dynacomm communication package (FutureSoft, Inc., Houston, TX) and currently runs on MS-DOS Windows (Microsoft, Redmond, WA) machines. (A Macintosh version of Dynacomm is under development.) The client machine may be located anywhere on the Internet. To ease the problem of software updating and maintenance at our insti-

Figure 1 A schematic overview of the pilot information sources map system's (ISMS) system architecture.

Figure 2 The initial screen of the information sources map system (ISMS), which allows the user to define a search, as described in text.

tution, we keep a single copy of the ISMS client program on a network-based file server from which it can be run on many different client machines.

The ISMS server contains a variety of information about each ISM source (Appendix I contains a sample ISM KB entry), including:

1. Coded information describing the subject contents of the source. This information is coded in two biomedical coding languages: 1) MeSH, the Medical Subject Headings of the NLM,¹³ and 2) Meta-1 semantic types from the UMLS Metathesaurus.¹⁴
2. Other coded information. A range of additional information is encoded describing each source, including the type of source, its availability, and its likely use, as described in the next section.
3. Descriptive (textual) information. An English-prose description of the source, its contents, its available, its producer, etc., is also included for display to the user on request.
4. Communication access information. The ISMS server also contains scripting information to allow automatic connection from the ISMS client to a chosen source.

Of the 121 sources currently included in the pilot ISM KB, 36 source entries were prepared by the NLM and 85 were prepared by Yale.

An Example of the ISMS in Operation

This section gives an example of the pilot ISMS in operation. In this example, a user is interested in

information relevant to cancer. Figure 2 shows the initial screen by which the user formulates a search. At the top of the screen are two boxes into which the user may enter search terms. If two terms are entered, they may be ANDed or ORed together. In this example, the user enters a single term, "cancer."

In addition to entering search terms, the user may elect to further constrain the search for relevant information sources using the checkboxes in the main portion of the screen, indicating:

1. Desired types of sources to be retrieved (e.g., journal citations, full text, teaching programs). Table 1 shows the range of source types currently available through the ISMS.
2. Desired availability of the retrieved sources. Using these checkboxes, the user can request only freely available sources and/or only sources available through the NLM, BRS, or DIALOG. Table 2 shows the modes of availability of the current set of ISMS sources.
3. Desired focus on a particular "likely use." For example, if the query involves clinical patient care rather than basic research or health services research, these checkboxes allow retrieval of sources most likely to be relevant to such use.

The input screen shown in Figure 2 is displayed by the ISMS client program using Dynacomm, which includes a programmable graphic user interface. When the user finishes describing a desired search, the but-

Table 1 ■

The Range of Information Source Types Available in the Current Yale Version of the Information Sources Map Knowledge Base*

Journal citations	51
Book citations	29
Textbooks online	5
Other full text online	13
Reference databases	24
Teaching programs	21
Clinical advisors	2
Images	7
Software information	7
Government documents	10
Patents	5
Grants/research	4
News groups	2
Directories	6
Audiovisuals	5
Miscellaneous	9

*There are 121 sources, of which many fall into more than one source type.

Table 2 ■

Modes of Availability of the Information Sources in the Current Yale Version of the Information Sources Map Knowledge Base

Free via network	19
Free at Yale	50
NLM*	36
BRS	30
DIALOG	36

*NLM = National Library of Medicine.

ton "Find Information Sources" is selected. At this point the ISMS client sends a message over the network to the ISMS server.

When the ISMS server receives the user's search request from the ISMS client, it first looks up the search term(s) in the UMLS Metathesaurus dictionary. (The details of the ISMS's search algorithm can be found in Appendix II.) If the term is not found in this dictionary, then a list of alphabetically close dictionary terms (including permutations) is returned for display to the user, who may then either select a term or return to the first screen to try again. If the term is found in the dictionary, then the ISMS server performs a search of its coded sources. When the user's term can be readily mapped to MeSH, then the MeSH coding is used for this search, otherwise Meta-1 semantic types are used. MeSH is preferred because it is much more detailed and therefore allows much more specific coding than do Meta-1 semantic types. In this example, "cancer" is a Meta-1 term and is closely related (by Meta-1 "related term" links) to

the MeSH term "neoplasms," which is then used in the search for relevant sources.

In this fashion, the ISMS server compiles a list of sources that match the user's search terms and any other search constraints. This list is returned to the user as shown in Figure 3. In this example, the sources returned are grouped to reflect how closely they match the MeSH term "neoplasms." The first group of sources displayed in Figure 3 were (in general) coded specifically with the MeSH term "neoplasms." These sources would be expected to focus specifically on cancer, although the search algorithm (shown in Appendix II) has the ability to promote certain general sources such as MEDLINE into this group based on special coding. The next group was coded with a higher-level term in the MeSH hierarchy than "neoplasms", such as "diseases." These sources would be expected to include information about cancer, but not to be specifically focused on cancer. The third group of sources are those coded 1) with lower-level MeSH terms, e.g., the name(s) of specific cancers (which therefore might or might not be of interest to the user), or 2) with MeSH terms that are coded with a small number of subheadings, indicating that the source covers only limited aspects of the subject. A number of AIDS-related sources appear in this group because they are coded with the more specific term "Kaposi's sarcoma," an AIDS-related cancer.

Once the sources are displayed, the user can use the graphic interface of the ISMS client to request further information about a particular source. This request results in a message's being sent to the ISMS server,

Figure 3 On these two screens (screen A scrolls into screen B), the information sources map system (ISMS) lists the sources that have matched the user's search. At this point the user may request 1) further information about a source, 2) automatic connection to a selected source, or 3) that the set of sources be regrouped, as described in text.

which returns the desired information for display. Alternatively, the user may request automatic connection to a chosen source, in which case the ISMS server is queried for the required communication script information. All connections from the client ISMS to a chosen source are made directly via the network (not through the server), and typically happen in two to three seconds.

To connect to a chosen source, the ISMS client launches a communication script, or a source-specific access program (e.g., for CD-ROMs), that logs onto the source over the network. In the future this capability could be enhanced to allow a variety of different ways to access sources, such as via X-Window display or by actually downloading a program from the source over the network to run locally on the client machine.

When retrieving sources with the ISMS, a number of potentially relevant sources are typically found (as long as the search includes types of sources that are sufficiently represented in the pilot ISMS). As a result, there is seldom the "no hit" problem that is often a concern with many bibliographic retrieval systems.

Figure 4 shows an additional capability that the ISMS offers. Using the checkboxes at the bottom of the screen, the user may request that a set of retrieved sources be regrouped along a variety of different axes, in addition to the "closeness of search term" grouping shown in Figure 3 (labeled "grouping by priority"). Figure 4 shows the same set of sources grouped by "subject focus." Notice that some of the sources are quite broadly focused, but others focus on specific

Figure 4 These screens show the same set of sources seen in Figure 3 but regrouped by different criteria. (Screen A scrolls into screen B which scrolls into screen C which scrolls into screen D.)

topics, e.g., "chemically induced," "drug therapy," "genetics," and "history." In addition, the sources can currently be grouped by their "availability" (divided into groups labeled "free," "NLM," "BRS," etc.), by their likely use, or by their source type.

The purpose of this regrouping capability is to let the user look at the retrieved sources from several different perspectives to get as full as possible an appreciation of that set of sources as a whole. Thus, in addition to getting descriptive information about an individual source, the user can obtain a "gestalt" of the entire set of retrieved sources by viewing them from multiple perspectives.

The Current Status and Preliminary Use of the Pilot ISMS

The pilot ISMS is currently operational on roughly 40 PC/Windows workstations, including 15 in public areas within the Yale Medical Library and others on a growing number of staff desktops within the School of Medicine. It is accessible through NetMenu,⁶ our institutional network menu that has also been developed as part of the UMLS project. NetMenu currently provides direct access to roughly 30 different information sources, and includes the sources most frequently accessed from the public workstations at the medical library.

NetMenu itself was recently installed (in early 1993). Prior to the installation of NetMenu, there was no unifying approach to network-based information access for our medical library resources. Access to highly used sources (miniMEDLINE, Current Contents, and the Yale University Library catalog) from the medical library had been provided by separate character-based terminals dedicated to each source. NetMenu represents a major improvement in functionality over the previous capabilities, and meets most of the current information-seeking needs of the user community. Usage of NetMenu from the medical library is running at roughly 15,000 to 20,000 connections per month to the various information sources.

Usage of the ISMS currently runs at about 200 to 300 connections per month. The difference between the use of NetMenu and the use of ISMS reflects the fact that NetMenu satisfies most of the information source retrieval needs of our user community today. NetMenu has been structured to contain most of the publicly available sources that our biomedical community is likely to use, as well as those sources licensed or purchased for mounting on Yale's local area network. Thus, just looking through NetMenu is often an easy way for our users to find relevant sources. As a much broader and richer set of sources becomes available,

and as users become accustomed to accessing these sources, we anticipate that usage of the ISMS will increase.

In addition, as many more information sources become available to the Yale Community, these will no longer fit comfortably into the two-level hierarchy offered by NetMenu. One approach to this proliferation might be to change NetMenu into a multilevel hierarchical menu. The problem with this approach is that 1) sources may get buried in the hierarchy and become difficult to find (so that some kind of search capability becomes important), and 2) different user groups will likely want differently structured hierarchical menus. The ISMS provides an alternative solution, since it allows the user to retrieve a set of potentially relevant sources and then flexibly structure them in a variety of different shallow hierarchies tailored to the user's needs, using the regrouping capability. As a result, as an increasing number of sources become available, we foresee that our library will need to move to the more flexible type of capability provided by the ISMS. Thus, NetMenu can be seen as a temporary step toward a future that will require the broader types of capabilities offered by the ISMS.

To get more specific user feedback concerning NetMenu and the ISMS, we recruited a "test panel" of 12 individuals who were comfortable and active users of network-based information sources, and who would therefore not need extensive training in how to search the individual sources. We made NetMenu/ISMS available to these users for approximately one month, with use of NLM-based sources available free of charge. After the one-month test period, we interviewed these users to get their detailed, subjective reactions. The comments below are representative of the feedback we received concerning the ISMS from the test panel.

Test user 1: "I'm so stuck in my own grooves that the extra stuff that the ISMS revealed was interesting. There's been such an advance in information access in the last few years. It has helped my research. This is icing on the cake."

This comment typifies the positive feedback received. One clear lesson, however, was that for users who were already comfortable with network-based searching, the ISMS provided an extra layer of interface that they did not always need. The following two reactions illustrate this effect.

Test user 2: "There are very few free sources that one can't already get to via Gopher, news readers,

or Telnet. (Scientific American of Medicine Consult seems to be the only exception.) The search terms are limited. There's no way to keep one's own defaults in the ISMS."

Test user 3: "If you know lots about sources, the ISMS isn't useful. It's making assumptions about search terms and information that I'm not sure I feel comfortable with—it's subjective material."

These comments suggest that building an ISMS that will fully satisfy experienced users could be a challenging task. The questions raised about search terms are discussed in more detail later in this article. This preliminary user experience with the pilot ISMS suggests that it is a potentially valuable tool, but that it will be more valuable when more users start trying to access a broader set of information sources that are currently accessed by most users today.

Lessons Learned in Building the Pilot ISMS

The purpose of building the pilot ISMS was to gain a better understanding of the various design issues involved in creating such a tool. In the process of building the pilot ISMS described here, we have gone through several design and implementation iterations. First, we built a prototype nonoperational version of the ISM KB (which did not include the capability to connect to sources) to let us test different approaches to coding the information sources.¹⁵ Next, we implemented an initial version of the current client-server version.² This version was then refined in many ways over more than a year of operational use. For example, one of the major refinements was the incorporation of the source regrouping capability described previously in this article. In this section, a number of the major lessons learned in the process of building, deploying, and refining the pilot ISMS are discussed.

Achieving Seamless, Cross-platform Access to Heterogeneous Sources Is a Major Undertaking

The primary focus of the project has been to explore the conceptual design issues involved in building an ISMS, rather than to build a product. At the same time, it was clear that we needed to provide our users with a polished system if we expected them to use it comfortably and give us the type of feedback we desired. For example, we felt it was important to be able to connect automatically to a requested information source. The need to strike a balance between 1) a primary focus on conceptual design issues and 2) the desire to present as polished a system as pos-

sible to our users resulted in a number of implementation decisions.

To handle network communication, we used Dynacomm, a commercial communication package that runs on a PC/Windows client machine. Dynacomm includes communication scripting, which handles communication between the ISMS client and server and between the ISMS client and most of the sources. Dynacomm also provides a graphic user interface (Figs. 2, 3, and 4) for formulating a search and for viewing results. In a fully functional product, it would be desirable to provide cross-platform compatibility so that the ISMS could run on a variety of platforms, e.g., PC/Windows machines, Macintoshes, and X workstations.

To achieve such product-oriented goals would require a great deal of additional work. With the use of Dynacomm and with a focus on the PC/Windows platform, we were able to provide our users with a polished system, while concentrating on the design of the ISMS itself and not getting bogged down in issues of portability and communication programming. The significant amount of work required to implement the pilot ISMS makes it clear that building a fully functional product would be a major undertaking. Looking to the future, however, a variety of commercial software tools are becoming available to assist with cross-platform compatibility and network communication, and over time such tools will become increasingly powerful.

Coding the Subject Content of an Information Source Is Different from Coding That of an Article or a Book

In building the pilot ISMS, it quickly became clear that the process of coding an information source is different in several important ways from the process of coding a specific article or book. With a specific article, an indexer can read it through to determine exactly what is discussed. When coding an information source, this is usually not possible, in part because of the large amount of material involved and also because most current ISMS sources (e.g., MEDLINE) continue to receive new material on an ongoing basis. As a result, the coding can at best reflect the general topic areas for which consulting the source would likely be productive. A major problem, however, is that a coder may find it very difficult to anticipate all the potential topics relevant to a source.

A related issue concerns the granularity of coding. For example, if an expert system contains knowledge about 900 diseases of internal medicine, should it be coded specifically with a listing of all those diseases

or with a broader term such as "diseases"? If the broader coding is chosen, then the source will be suggested even when a disease not included in the source is entered as a search term. On the other hand, as described above, coding many sources (for example, MEDLINE) explicitly and exhaustively with exactly what they contain is impractical, if not impossible. In the pilot ISMS, most sources are coded with a relatively modest number of quite broad terms.

As a result of these factors, the coding of ISM sources may at times be inherently less precise than is the coding of specific articles. The user of the ISMS must therefore understand that the sources retrieved are offered as *potentially* relevant to the query.

The inherently inexact nature of the coding can also cause a potential problem when two terms are ANDed together in a search. The difficulty is that any imprecision in retrieval may be compounded, and that some useful sources may match one but not both terms. As a result, when ANDing two terms together, it may be useful to show the user all sources retrieved in three groups: 1) sources matching both terms, 2) sources matching term 1 only, and 3) sources matching term 2 only. This approach would allow the user to take advantage of the increased focus provided by the AND, while still preserving the ability to view useful sources that might have matched only one term.

Access to Different Versions of a Source Should Be as Seamless as Possible

Another problem arises because there are multiple versions of certain information sources (e.g., versions provided by different vendors), and these versions may differ. For example, there are multiple versions of MEDLINE, including an institutional version at Yale that is provided free to Yale users, as well as versions available through various online services such as BRS or DIALOG.

One approach might be to treat each of these access routes to MEDLINE as a separate source. This approach might be justified by the fact that the versions differ, usually in relatively small ways but nevertheless in ways that might be important for certain users. The problem with treating each version as a separate ISM KB entry is that it can be confusing and potentially overwhelming to a user whose search retrieves many versions of one or more sources. The user does not want to see ten versions of MEDLINE listed as potentially relevant to a query. This approach could also greatly increase the number of "sources" retrieved by a query.

As a result, we found that it was necessary to create a "generic" ISM KB entry for such a source, e.g., a single entry for MEDLINE as a whole, which provides a basic description of the source. The source description might also highlight any significant differences between different versions of the source. Using this approach, when the user asks to be connected to such a source, the ISMS first asks which version is desired and then activates the appropriate connection script.

One result of taking this integrated approach is that when a new version of a source is incorporated into the ISM KB, it cannot be added as a separate, independent entry. The generic description of the sources may need to be modified, and the new version's connection script will need to be added to the generic source entry. The problem of maintaining such a generic description of a source is made even more complex when multiple versions of the ISM KB need to be kept, each customized to a particular institution, as discussed below. Each institution may have a different set of source versions available and may therefore even need a different generic entry for such a source.

Customization of the ISM KB to an Institution Is Surprisingly Complex

There are currently two versions of the pilot ISM KB: 1) a generic version that contains information about sources on the Internet that can be accessed from any institution, and 2) a Yale version that includes all the sources in the generic ISM KB plus a number of sources that are local to Yale. (We have also collaborated with the University of Utah School of Medicine to explore the development of a customized version of the ISM KB for its use.) The two versions are needed in part 1) because the use of many Yale sources is limited by licensing restrictions, and 2) because we have added Yale-specific information, e.g., places and hours of availability, to many sources. Customization is also needed based on specific machine addresses, e.g., certain resources are available only to machines located within our medical library.

An institution often has its own version of many commonly used information sources that run on local machines or are available locally over the network from CD-ROM. Since these versions of the sources have been specifically mounted for local use, and are often prepaid and therefore available free of charge to local users, it is especially important that users be informed about them. (Not surprisingly, users, including those in our evaluation, prefer sources that are available free of charge.) As a result, when cus-

tomizing the generic ISM KB to a particular institution, a variety of issues arise.

1. Locally available sources not already in the ISM KB may need to be added.
2. Locally available versions of sources already in the ISM KB may need to be incorporated into the existing source entries.
3. In the process, it may be necessary to change the textual description of a source in the ISM KB (and/or other fields of the ISM KB entry) to reflect certain features of the local version of the source, and to add the communication script needed to access the local version of the source.

Keeping all these different versions of an ISM KB entry for multiple institutions, each involving a different subset of source versions for many different sources, would require keeping an organized record of many different permutations of an entry. If such customization were to be done on a wide scale, it would be essential to develop tools to facilitate the creation and maintenance of these customized versions of the ISM KB.

There Are Important Design Trade-offs between Search Specification and the Viewing of a Set of Retrieved Sources

The problem of ANDing two search terms, which was discussed previously, is an example of a much more general issue; there are design trade-offs between the degree to which the user should be encouraged to restrict a search and the degree to which the user should be encouraged to selectively partition the sources after they have been retrieved.

1. On the other hand, the ISMS can provide tools that let the user constrain a search and thereby limit the number of sources retrieved. As discussed above, because of the nature of source content coding, this approach runs the risk of not retrieving some potentially relevant sources.
2. On the other hand, the ISMS can let the user retrieve sources fairly liberally, and then provide tools to let those sources be grouped and viewed in a variety of ways to help the user focus on those that might be most helpful. In general, this approach seems to be most promising and, as discussed above, we have already incorporated a capability into the pilot ISMS to group the retrieved sources in various ways, allowing them to be viewed from different perspectives.

Another example of this trade-off is seen in deciding how best to allow modifiers (such as topical sub-

headings) to be used in the ISMS. Topical subheadings are modifiers that can be applied to search terms. For example, when applied to a drug, a modifier might be "diagnostic use" or "therapeutic use." Modifiers are widely used in indexing articles for MEDLINE. Such modifiers are often more difficult to apply to an information source than to a specific article because the contents of many sources (such as MEDLINE) are quite broad and cover many, if not all, appropriate modifiers. In the future, however, as more focused sources are included in the ISMS (such as expert systems or clinical guidelines that focus on specific aspects of a clinical problem), modifiers may be easily assigned to such sources.

When modifiers are included in the ISM KB coding (as many are now), the question is how best to use them. One approach is to allow the user to specify modifiers when formulating a search. A second approach is to let the user use coded modifiers to group a set of sources only after they have been retrieved. An advantage of the second approach is that the user does not have to anticipate ahead of time what modifiers might be relevant to a query. After a set of sources is retrieved, the computer can inspect the coding and make a list containing only those modifiers that apply, perhaps indicating how many sources were indexed by each modifier. The computer can present this list to the user, who can then indicate which subset(s) of the retrieved sources to display. Alternately, if the list of retrieved sources is relatively short, the computer could display the entire list of retrieved sources grouped by modifier (Fig. 4).

Although we favor the second approach for the present pilot ISMS, it is clear that there will always be a balance between providing tools to constrain a search and providing tools that allow the results of a search to be viewed in a selectively focused fashion. This is an important design trade-off and there is no simple, general solution. As the number of sources in an ISM KB grows to be quite large, the issue will become even more important and the best solution to this problem may evolve over time.

Promising Areas for Future Work

The implementation of the pilot ISMS has allowed us to recognize a number of ways in which the overall approach could be enhanced. Two interesting areas for future work are discussed.

Graphic Viewing of "ISM Space"

When the user defines an ISMS search, the current Dynacomm interface returns a linear list of grouped

sources. As described previously, the user can request that this list be regrouped in several different ways to get a better appreciation of the set of sources as a whole. This capability, however, is currently limited since it allows the display of only a *linear* list of grouped sources.

We anticipate that a more powerful graphic interface could allow the viewing of a set of sources to be performed in a much more flexible fashion.

1. The set of retrieved sources could be spread out in two dimensions along different axes.
2. The various groups of sources could be presented as nodes in a directed graph with any relationships between the groups indicated by labeled arcs between the nodes. Clicking on a node would then produce either a list of the sources in that group, or a graphic outline of subgroups of sources within that group. This type of capability will be particularly useful as the total number of ISMS sources grows, and as the set of retrieved sources for many queries becomes unmanageably large for convenient presentation as a single list.
3. In (1) and (2) above, the computer could examine the coded descriptions of the retrieved sources to suggest axes or groupings that would be most informative.
4. Icons could be used to represent different sources or groups of sources. Characteristics of each icon (e.g., its size, color, shape, and graphically depicted features) could reflect certain characteristics of the source.

In addition to such a 2D interface, there also exist "3D" interfaces¹⁶ for information access, which might be very useful for the ISMS.

Dynamically Matching a Source to a Client Machine

Most of the sources currently coded in the ISMS have simple character interfaces, and can therefore be readily accessed from any machine on which the ISMS runs. Looking to the future, however, there will be an increasing number of more exotic sources available. These include image databases, and image-containing teaching programs, with a variety of different types of static images and dynamic images. Also, certain programs will require specific hardware features to operate efficiently or at all. In addition, sources programmed to use a particular graphic user interface (e.g., X Window System, Macintosh, PC/Windows) will run only on certain machines.

As a result, in the future it will be important for the ISM KB to contain coded information indicating the

hardware and software configuration(s) that will support each source. Each client machine will similarly need to indicate its capabilities, so that the ISMS can retrieve only those sources that are appropriate for that client. Alternatively, the ISMS might show all the sources but use grey-scale shading to indicate which sources are available only on other client platforms.

Discussion

To help put the design of the ISMS into perspective, it is useful to compare it with Gopher, a network navigation tool with somewhat similar overall goals. Since it is easy to mount a Gopher server on the Internet and to access such a source from a range of Gopher clients, the use of Gopher has grown rapidly. One fundamental way in which the ISMS differs from Gopher is that ISMS sources are coded with information that can be searched for potential relevance to a particular biomedical question. Gopher allows only a very limited ability to perform such a search, since the equivalent Gopher capability (called Veronica) performs only a free-text search of the names of Gopher sources.

One way to characterize the difference between the ISMS and Gopher is to use the analogy of trying to view a forest versus seeing the trees. In this analogy, Gopher provides a very powerful tool to enfranchise different groups to make their own trees (Gopher sources and servers) available on the Internet, and to let users browse the Internet by moving from one tree to another. In contrast, the ISMS tries to give the user a more comprehensive view of the whole forest as it relates to a particular biomedical question (or at least that portion of the forest coded in the ISMS). In addition, using the pilot ISMS's regrouping capability, the user can view the forest from several different perspectives. The ISMS can provide these capabilities only because a great deal of time and energy has been invested in coding all the ISM KB sources and in keeping that coding up to date.

This coding of information sources also differentiates the ISMS from approaches such as WAIS and WWW. WAIS allows free-text searching of potentially relevant documents or server descriptions, but does not control the vocabulary that is used. WWW implements a network-based hypermedia approach with hand-coded links among documents, but does not implement global coding of the documents.

As a result of focusing on a selected set of sources and enforcing a centralized control on the contents of its entries, the ISMS may be better able to maintain

the integrity of its access links. In contrast, by enfranchising the community so broadly, Gopher allows different groups to create, modify, and remove their own Gopher sources as they chose. As a result, pointers in Gopher-space may point to Gopher servers or sources that no longer exist or that have been renamed or modified. There is currently no organized way to propagate changes so that pointers within Gopher-space can be updated when such changes occur. This problem also occurs with WWW.

Indeed, the features of the ISMS are quite complementary to those of Gopher and WWW. In the future, it is likely that these features will be increasingly combined.

Summary

The purpose of the pilot implementation of the UMLS ISM is to explore the design issues involved. In the near future, the ISM will need to evolve as the world of network-based information itself changes. As the number of available sources increases dramatically, the issues of filtering and displaying those sources will be more complex than they have been with a relatively small number of available sources. In addition, the types of available sources will become much more diverse and potentially esoteric. The interconnectivity among sources will also become more complex. Many network-based resources will come to be seen less as independent entities to be queried, and more as components that can be dynamically connected in a variety of ways to answer different types of questions. The current pilot ISMS is a first step to exploring certain of these issues, and provides a base from which to investigate issues that will be confronted in the future.

References ■

1. Lindberg DAB, Humphreys BL. The UMLS knowledge sources: Tools for building better user interfaces. In: Miller R, ed. Proceedings of the Fourteenth Symposium on Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1990:121-5.
2. Clyman JI, Powsner SM, Paton JA, Miller PL. Using the UMLS information sources map to access online reference materials. *Bull Med Libr Assoc.* 1993;8:207-16.
3. Broering NC, Gault HR. BioSYNTHESIS: Bridging the information gap. *Bull Med Libr Assoc.* 1989;77:19-25.
4. Ketchell DS, Fuller SS, Freedman MM. Collaborative development of a uniform graphical interface. In: Frisse M, ed. Proceedings of the Sixteenth Symposium on Computer Applications in Medical Care. Washington, DC: AMIA, 1992: 251-5.
5. Lucier RE, Matheson NW, Butter KA, Reynolds RE. The knowledge workstation: An electronic environment for knowledge management. *Bull Med Libr Assoc.* 1988;78:248-55.
6. Shifman MA, Clyman JI, Paton JA, Powsner SM, Roderer NK, Miller PL. NetMenu: Experience in the implementation of an institutional menu of information sources. In: Safran C, ed. Proceedings of the Seventeenth Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1993: 554-8.
7. Cimino C, Barnett GO, Hassan L, Blewett DR, Piggins JL. Interactive query workstation: Standardizing access to computer-based medical resources. *Comput Methods Programs Biomed.* 1991;35:293-9.
8. Lynch CA. The client-server model in information retrieval. In: Dillon M, ed. Interfaces for Information Retrieval and Online Systems: A State of the Art. Westport, CT: Greenwood Press, 1991:301-18.
9. Obraczka K, Danzig PB, Li SH. Internet resource discovery services. *Computer.* 1993;8-22.
10. Krol E. The Whole Internet: User's Guide and Catalog. Sebastopol, CA: O'Reilly & Associates, 1992.
11. Powsner SM, Roderer NK. Navigating the Internet. *Bull Med Libr Assoc.* 1994;82:419-25.
12. Dillon M. Assessing information on the Internet: Toward providing library services for computer-mediated communication. *Internet Res.* 1993;3:54-69.
13. Medical Subject Headings—Annotated Alphabetical List. Bethesda, MD: National Library of Medicine, 1993.
14. Tuttle M, Sherertz D, Olson N, et al. Using Meta-1: The 1st version of the UMLS metathesaurus. In: Miller R, ed. Proceedings of the Fourteenth Symposium on Computer Applications in Medical Care. Washington, DC: IEEE Computer Society Press, 1990:131-5.
15. Miller PL, Wright LW, Frawley SJ, Clyman JI, Powsner SM. Selecting relevant information resources in a network-based environment: The UMLS information sources map. In: Lun KC, Degoulet P, Piemme TE, Rienhoff O, eds. MEDINFO-92. New York: North Holland, 1992:1512-17.
16. Robertson GG, Card SK, Mackinlay JD. Information visualization using 3D interactive animation. *Comm ACM.* 1993;36(4): 57-71.

APPENDIX I

*A Sample Information Sources Map (ISM)
Knowledge Base (KB) Entry*

This appendix provides a representative example of an ISM KB entry describing an information source, in this case BIOSIS PRE-VIEWS. The purpose of this example is to give the reader a concrete appreciation for the type of free text and coded information contained in the ISM KB. Most of the ISM KB entry fields originated from the National Library of Medicine. Several fields were added by Yale, either to code institution-specific information or to allow us to test capabilities that we felt were potentially desirable. The time required to create an ISM KB entry was very dependent on how easy it was to find and review the required documentation, and ranged from a half hour for a simple source with all information easily available to several hours or more.

For machine processing, the ISM KB entry is coded internally in a slightly terser format than that shown below. We have shown a more legible, machine-processed version, which is displayed when a user requests "Full Details" about an entry. Each field of the ISM KB entry starts with a field name. For clarity, these have been underlined below. We have omitted a few fields that are not used in the searching and regrouping algorithms. To help the reader understand the entry, we have added explanatory notes in *italic*, enclosed by square brackets.

[The entry starts with an identification field.]

Name: BIOSIS PREVIEWS

[The following free-text fields are displayed in full when the user asks for a description of the source.]

Description: Contains more than 7 million citations with abstracts to worldwide literature on research in life sciences.

General-Description: Covers microbiology; plant and animal science; experimental medicine; agriculture; pharmacology; ecology; biochemistry; bioengineering; and biophysics. Also covers patents (to 1989) in such areas as immunological testing, food processes, and fishing. Patent records include inventor's home and address, patent title and number, patent classes, date granted, and assignee. Approximately 9,000 periodicals, as well as books, patents, monographs, conference proceedings, research communications, and symposia, are screened.

[The following field provides institution-specific information in the Full Detail screen.]

Misc: A version of BIOSIS is available free at Yale through ORBIS. It contains Biological Abstracts from 1988 to the present, but does not include reports, proceedings, and monographs. Dial-in access is available. Contact Systems Office, Yale University Library, 432-1847. Remote access is available between the hours of 8 AM and 12 midnight, 7 days per week.

[The following free-text field provides availability and access information customized to the local institution, in this case Yale. This information is displayed in full when the user asks for a description of the source.]

Availability:

FREE COMPUTER access to BIOSIS on ORBIS, 1988 to present:

Online access available at:
Medical Library

Kline Science Library
Social Science Library
Sterling Memorial Library
EPH Library
Engineering and Applied Science Library
Seeley G. Mudd Library
Law Library
Divinity Library
(and all other Yale libraries)
Microcomputer or workstation with modem.
For information on connecting to BIOSIS, call Systems Office, 432-1847,
or pick up information sheet at any Yale Library Reference Desk.

MEDIATED online searching:

Medical Library
Kline Science Library

USER ACCOUNT required for online access:

BRS 800-955-0906
DIALOG 800-334-2564

PRINT version: "Biological Abstracts"

Medical Library
Kline Science Library

[The following fields may be used to restrict searching and to regroup search output.]

Use-Basic-Research: commonly

Use-Clinical-Research: commonly

Use-Health-Services: unlikely

Use-Patient-Care: unlikely

Type-Database:

journal citations
book citations
patents

[Medical Subject Headings (MeSH) numbers, with subheadings when appropriate, are used to code the subject content of an information source. These codes are used to select sources based on the closeness of match to the search terms. This source was coded at Yale.]

MeSH:

A	Anatomy (MeSH Category)
B	Organisms (MeSH Category)
C	Diseases (MeSH Category)
E	Analytical, Diagnostic and Therapeutic Techniques and Equipment (MeSH Category)
F	Psychiatry and Psychology (MeSH Category)
G1	Biological Sciences
G1.201	Biochemistry
G1.273.343	Genetics
G1.344	Biophysics
G2.77	Biomedical Engineering
G3	Environment and Public Health (Non MeSH)
G4	Biological Phenomena, Cell Physiology, Immunity (Non MeSH)
G5	Genetics
G6	Biochemical Phenomena, Metabolism, Nutrition (Non MeSH)
G7	Physiology, General (Non MeSH)
G8	Reproduction, Urogenital Physiology (Non MeSH)
G9	Circulatory, Respiratory Physiology (Non MeSH)
G10	Digestive, Oral, Skin Physiology (Non MeSH)
G11	Musculoskeletal, Neural, Eye Physiology (Non MeSH)
G12	Chemical Pharmacologic Phenomena (Non MeSH)

H1.181.122 Biochemistry
 H1.671.100 Biophysics
 J1.293.143 Biomedical Engineering

[MeSH "extra strength" (MES) codes are used to indicate that the source provides particularly strong coverage of particular subjects. The MES codes have the effect of promoting the source to the highest priority for searching category even when the search term(s) are lower on the tree of any of the MES codes.]

MeSH "Extra Strength":

B Organisms (MeSH Category)
 D5 Environmental Pollutants, Pesticides (Non MeSH)
 D6 Hormones, Substitutes, Antagonists (Non MeSH)
 D7 Reproductive Control Agents (Non MeSH)
 D8 Enzymes, Coenzymes, Enzyme Inhibitors (Non MeSH)
 D9 Carbohydrates and Hypoglycemic Agents (Non MeSH)
 D10 Lipids and Antilipemic Agents (Non MeSH)
 D11 Growth Substances, Pigments, Vitamins (Non MeSH)
 D12 Amino Acids, Peptides and Proteins (Non MeSH)
 D13 Nucleosides and Nucleotides (Non MeSH)
 D23 Neuroregulator Blockaders (Non MeSH)
 D24 Immunologic and Biologic Factors (Non MeSH)
 D25 Biomedical and Dental Materials (Non MeSH)

[Meta-1 semantic type codes are used for matching a search term to a source only when no MeSH terms related to the search term can be found.]

Semantic Types:

A1.1 Organism
 A1.2 Anatomical Structure
 A2.1.5.3 Molecular Sequence
 B2.2.1 Biologic Function

[The following field contains script information, which is passed from the ISM server to the ISM client when the user presses the Connect button, to allow the client to connect to the source.]

Script-Information:

RS Orbis
 SS "1992 to present" NBIO
 SS "1988 to 1991" OBIO
 RS BRS
 SS "1978 to present" BIOL
 SS "1970 to 1977" BIOB
 SS "1970 to present" BIOZ
 RS DIALOG
 SS "1969 to present" 5
 SS "1981 to present" 55

[The following fields provide general descriptive information when the user asks for "Full Detail" about the source.]

Content:

Abstracts
 Citations
 Clinical protocols
 Clinical treatment protocols
 Titles of publications

Publication-Types:

Book chapters
 Books
 Computer programs
 Conference papers
 Dissertations
 Government documents

Journal articles
 Technical reports

Audience:

Biomedical researchers
 Chemical and Drug Manufacturers
 Environmental Specialists
 Health Sciences Students
 Historians
 Librarians and Information Specialists
 Science Writers

Coverage: comprehensive

Language-Of-Pubs: @ALL

Language-Of-System: ENG

First-Year: 1970

Update-Frequency: Monthly

Consensus-Level: article

[Information indicating who prepared the entry, original date, and date of last modification.]

Ism-Record-Source: SJF

Entry-Date: 19930407

Last-Modified: 19940825

APPENDIX II

Details of the Information Sources Map System's (ISMS) Search Algorithm

This algorithm outlines the search algorithm used to help users of the pilot ISMS identify sources of potential interest.

1. User Input. Users can specify their interests in two ways (as shown in Fig. 2):
 - a. Subject terms. Zero, one, or two terms of interest are entered into text boxes. If a term doesn't appear in Meta-1, a dictionary and term-detail tools help the user select one that does. Two terms can be joined by "and" or "or."
 - b. Source characteristics. Checkboxes are used for type, use, and access categories.
2. Source Evaluation. The list of sources is pruned to those qualifying under 1b, source characteristics, above. If no subject terms are specified, sources are listed alphabetically. If terms are specified, they are used to find appropriate subject codes to select and rank sources.
 - a. Term lookup. For each term, Medical Subject Headings (MeSH) or semantic type codes are found:
 - If the term's concept entry has any MeSH codes, those are used.
 - Otherwise, up to three successive levels of parent, child, and related concepts are checked for all codes (in a breadth-first fashion).
 - If MeSH codes are found at any level, the union of MeSH codes at that level is used. (This is virtually always possible, normally at the first level.)
 - If no MeSH codes are found by level 3, the original concept's semantic type codes are used (except for the 35 concepts that don't have semantic types).
 - Otherwise, the union of semantic types at the closest level available (always level 1, in practice) is used, with one exception. One concept has no codes or relationships: "masked" (from McMaster Epidemiology Terms). In this instance, a message is returned saying that no codes can

be found and asking the user to choose a different term.

- b. Source selection and ranking. The stored subject codes for each source are compared with the term codes from 2a (derived directly or indirectly from the terms entered). Matches are placed in one of the three "priority" categories, in two stages. The first stage is based on each source's normal MeSH or semantic type codes, as follows:

Priority 1 Exactly match: Some term code exactly matches some source code.

Priority 2 Partial match, source broader. The source has a code that is a parent of some term code in the relevant code hierarchy. As a special case, some sources are coded as relevant to all codes.

Priority 3 Partial match, source narrower (same principle, reversed).

In the second stage, when MeSH codes are being used (almost always the case), two other source fields created at Yale are checked:

1. MeSH strengths (MES) is a field with MeSH codes for which a source is particularly strong. Any matches that would normally rank as priorities 1 or 2 above are ranked as priority 1. Thus, CANCERLIT includes the code "C4 Neoplasms," which would result in a "priority 1" match to any term having an MeSH code starting with "C4" (e.g., C4.557.690.863 Sarcoma, Kaposi's).

2. MeSH subheads (MSH) is a field with MeSH codes qualified by one or a few MeSH subheads to which their coverage is limited. For example, OMIM includes the code "C4/GE Neoplasms (Genetics)." Any matches, even if they would normally rank as priorities 1 or 2 above, are ranked as priority 3 because the source's coverage is only partial.

For each term, the highest score found is used.

- c. Score merging. If only one term was entered, its score is used. If two terms are ORed together, the highest score is used. If ANDing two terms, the lowest score is used.