Evaluation of the Information Sources Map

Eneida A. Mendonça, M.D., James J. Cimino, M.D.
Columbia University, New York, NY, USA

As part of preliminary studies for the development of a digital library, we have studied the possibility of using the UMLS Information Sources Map (ISM) database to provide the means to connect and map different terminologies, as well as to facilitate access to available information sources. The main issues discussed are the indexing of and connection to relevant online sources. We found the features of the ISM to be consistent with the need to support automated source selection and retrieval. However, attention should be paid to three aspects of the information: granularity, completeness, and accuracy. We found the ISM to be potentially useful; however, significant modifications will be required if the ISM is to be able to support automated source selection and retrieval.

INTRODUCTION

A number of studies have shown the need of health care providers and patients for access to information pertinent to clinical practice and health–related issues.[1,2,3] Patients and their families need information that will help them to understand their personal situations and make better decisions. Health care providers need clinical relevant information related to the individual patient under supervision, including information on diagnosis, therapy, and patient care. Evidence–based medicine focuses on questions related to the central tasks of clinical work: clinical findings, etiology, differential diagnosis, prognosis, therapy, prevention, and self-improvement.

There is a large and rapidly growing number of online information sources relevant to health care. Using the Internet, an increasing number of health care providers and patients gain free access to an expanding volume of information that previously was inaccessible. The sources provide information of different types, including full text, abstracts, scientific facts, images, bibliographic citations, patient education, and computer-assisted instruction. Seeking useful and valid information on the Internet can be difficult because of the speed and lack of control with which the information is accumulating.[4] Filtering the information is a complex and time-consuming task. Judging if the information is applicable and credible is challenging.

There is a need for tools that can facilitate the access to large amounts of information and provide appropriate interactivity. The effective use of technology can be an important facilitator of quality, and utility, in reviewing medical information on the Internet.[5]

Previous research has used different approaches to facilitate user access to online information sources.[6] One approach suggested is the use of the Unified Medical Language System (UMLS) Information Sources Map (ISM) as a tool for locating and classifying biomedical information sources.[6-9] In this paper, we analyze the ISM to determine its suitability for providing appropriate identification of information resources as well as the means to automatically access them.

THE INFORMATION SOURCES MAP

The ISM is a component of the Unified Medical Language System (UMLS)[10] project at the U.S. National Library of Medicine. It is intended to support a system in which the user can pose a biomedical query, and in response receive a summary of pertinent information, with facilitated access to the full content of the information.[11]

The 1998 version contains 75 information sources. The sources are varied and include major medical bibliography databases, expert systems, and databases on medical images, toxicology, drugs, environmental health, genetics, DNA, chemicals, and protein sequences, among others.

For each source, the ISM includes a narrative description of the database, the type of information it contains, the probable uses of the database, an indication of who the intended audience might be, the organization that provides and maintains the database, the frequency the updates occur, and the name and address of contact individuals. Four elements are used to index the scope of the information in the source: relevant MeSH terms[12], MeSH subheadings, semantic types from the UMLS Semantic

UI:	**E00033**
Full name:	**Cancer Literature**
Acronym:	CANCERLIT, CANCERLINE
Source type:	Publication database
Description:	NLM bibliographic database: cancer-related items, 83% w/ abstracts
Update frequency:	Monthly
Text:	CANCERLIT contains bibliographic records for cancer-related documents
Data type:	ASCII Text
Covers from:	1963-01-01
Covers to:	Current
Content covers:	Government publications, journal articles, technical Reports, dissertations, meeting reports, books / monographs
MeSH terms:	Cell Membrane, Chromosomes, Tumor Stem Cells, Adenoviridae, Herpesviridae, Tobacco, HTLV-I Infections, Radiation Injuries, Carcinogens (partial list)

Figure 1: Source Sample (Partial content)

Network, and semantic links. It provides the name of the host system and information on the access. Figure 1 shows an example of a partial description of an information source.

METHODOLOGY

As part of the preliminary studies for the development of a digital library, we have studied the possibility of using the ISM database to provide the means to connect and map different terminologies, as well as to allow access to available information sources. For this study, we used the 1997 version because the 1998 version was not available.

During the evaluation process, we looked at the percentage of completeness and type of the information provided in the ISM. We also reviewed the use of indexing terms and the ability to connect to an information source as stated in the access information and scripts.

RESULTS

We found 72 information sources in the 1997 version, all containing basic information such as name of the source, and a characterization (including an overview, a description, and a sample record). Source identification was also provided for all sources including a source type and a short description. We found references to the update frequency in 66 (91.7%) of them. Sixty-four (88.9%) had information on the geographic origin, and 47 (65.3%) had a description on the content language(s).

Each source was described in terms of content, and each source content had a more detailed description. The contents included were: journal articles, vocabulary, chemicals and substances, treatments, book chapters, books/monographs, technical reports, book chapters, dissertations, multimedia, legislation and laws, computer software, government publications, organizations, series/journals, newspaper articles, manuscripts, databases, collection servers, treatments, genetic sequences, patients, anatomy, and pictures.

Table 1 – Fields of Activities

Field description	C	P	U
Basic research	33	12	27
Clinical research	29	20	23
Consumer information	16	28	28
Emergency response	4	16	52
Environment monitoring	20	14	38
Health services research	17	13	42
Historical research	5	33	34
Library & information service	41	6	25
Patient care	31	14	27
Teaching	9	36	27

* Where: C = commonly, P = possible, U = unlikely

The database also provided estimates of the general likelihood that a source would prove useful for various specified fields of activity. The likelihood was described as unlikely, commonly, and possibly. Ten fields of activities were described. All sources were described in terms of field of activities and estimates of use. Table 1 shows the number of times each field of activities is defined as commonly, possible or unlikely.

We found that one or more MeSH terms were assigned to all sources, as well as MeSH subheadings, reflecting areas of significant coverage. Semantic types and semantic relation index term were also assigned to all sources.

The assignment of MeSH terms is quite broad in some sources. For example, the

Directory of Information Resources Online (DIRLINE) is indexed only by very broad MeSH terms such as “anatomy”, “diseases”, “physical science”, and “health care”, among others. Some of the semantic types associated with DIRLINE are “pathological function”, organism attribute”, “organization”. There are 15 other sources indexed with “diseases”, and 18 with “health care”.

Other sources descriptions had only a small number of very broad indexing terms, such as Integrated Risk Information System (Chemicals and Drugs and Environmental Exposure), and Toxicology Information Online (Chemicals and Drugs, Environmental Pollution). Others have a more detailed indexing. For example, there are 116 MeSH terms associated with QMR, including some specific diseases.

We found that 64 (88.9%) descriptions had information about accessing the source. Some sources had two or more connection protocols available for access. Fifty-four sources were available by telnet, 23 via web-based applications, 4 by gopher, 1 by WAIS, and 1 by ftp. Only 18 (25%) of the sources allowed free access to the information.

The majority of the free access protocols (15 – 83.4%) did not require login/password. Only 6 (33.3%) contained scripts for accessing the source. When trying to connect such information sources, we found that only a few connections would go directly to the search page and no information retrieval could be done automatically using only the information stored in the ISM database. All sources, but one, are said to connect to the “top” of the application. However, we found that pointers to sources were often to a Web page or a menu that required additional steps to actually reach the source content. No information is provided about the client-side environment and requirements. Table 2 describes the results of our attempts to access free information sources.

DISCUSSION

The primary focus of this review was to explore the issues involved in the use of the UMLS Information Sources Map as a tool to provide an appropriate identification of information resources as well as the means to automatically access them.

One of the main challenges in building a digital library is the resource location. The main issues discussed in this paper are the indexing and the connection to relevant online sources. Searching mechanisms have been evaluated[13] and are not the objective of this project, although indexing terms are an essential part of the searching mechanisms.

The first issue is how to determine the best source to answer a particular question. The question is “does the indexing cover the general topic areas of which the source would be probably searched?” One important point to consider is that each source has much information available, and it is probably not possible to index all of it. At the same time, with rapid growth of information, is it possible to anticipate the potential topics relevant to a particular source?

A second and very important point is the granularity of indexing. Some sources are quite broadly characterized in the ISM. For example, how does one characterize an expert system such as QMR or DxPlain? In the ISM QMR is defined with more granularity, including many diseases, syndromes, group of diseases, and certain drugs, among others. DxPlain, on the other hand, is defined with less granularity, including MeSH terms such as “diseases” and “therapeutics”. The use of broad indexing such as “diseases” will always lead to problems such as the suggestion of possible source which does not contain a certain disease, although it is identified with the term. At the same time, if all information will be described explicitly, it will be possible to index an expert system such as QMR, but will probably be impractical to index huge database sources such as MEDLINE. Similar issues have been discussed by Miller at al.[6]

The selection process also involves the type of questions asked and the “fields of activities”. The ISM provides estimates of the general likelihood that a source would prove useful for various specified fields of activity (11 fields in the 1997 version). The more specific the indexing and “fields of activities”, the more useful it will be for source selection.

Another crucial problem is the automated connection to online sources. Once we have identified a potential source, an automated connection is desired. The current version of the ISM does not provide sufficient information to allow an automated connection. The automated connection requires not only information on the protocols available but specific scripts or templates. For example, the Online Mendelian Inheritance in Man source, described in the 1997 version of ISM, does not have a script associated with the WWW protocol. However, when we connected to the URL associated with the protocol, we found that the page contained all the information on how to create scripts for automatic queries to that database using their web-search engine. We believe that such information should be added to the ISM.

A second issue related to the automated connection problem is the presence of multiple versions of the same source (e.g. MEDLINE, Toxic Chemical Release Inventory). This problem raises questions such as "should one version be used or all versions be used?" Miller et al.[6] suggest an implementation of a "generic source" to solve multiple versions of the same source.

The features of the ISM are consistent with the need to support automated source selection and retrieval. However, attention should be paid to three aspects of the information:

a) granularity: levels of indexing must be set to ones which are consonant with the needs for selecting the most appropriate information source;
b) completeness: although many fields are complete, additional work is need to achieve 100% completeness;
c) accuracy: with the rapid changes of information resources, maintenance becomes more difficult, making achieving accuracy almost impossible. We believe each provider should be responsible for maintaining their records, and either send updates to the UMLS or make them available in some standardized way on their Web sites.

CONCLUSION

This paper describes a review we have done of the 1997 ISM, as part of the preliminary studies for the development of a digital library. We found the ISM to be potentially useful. However, significant modifications will be required if the ISM is to be able to support automated source selection and retrieval.

Acknowledgments

This work was support by the New York State Center for Advanced Technology, USA, and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), grant 200557/95-5, Brazil.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med 1985; 103(4):596-9.
2. Timpka T, Ekstrom M, Bjurulf P. Information needs and information seeking behaviour in primary health care. Scand J Prim Health Care 1989; 7(2):105-9.
3. Shelstad KR, Clevenger FW. Information retrieval patterns and needs among practicing general surgeons: a statewide experience. Bull Med Libr Assoc 1996; 84(4):490-7.
4. Jadad AR, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? JAMA 1998; 279 (8):611-4.
5. Silberg W.M., Lundberg G.D., Musacchio R.A. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveant lector et viewor--Let the reader and viewer beware. JAMA1997; 277(15):1244-5.
6. Miller PL, Frawley SJ, Wright LW, Roderer NK, Powsner SM. Lessons learned from a pilot implementation of the UMLS Information Sources Map. JAMIA 1995; 2(2):102-15.
7. Miller PL, Clyman JI, Frawley SJ *et al.*, author. NetMenu and a prototype UMLS Information Sources Map. Proceedings - The Annual Symposium on Computer Applications in Medical Care. 1994: 957.
8. Miller PL, Paton JA, Clyman JI, Powsner SM. Prototyping an institutional IAIMS/UMLS information environment for an academic medical center Bull Med Libr Assoc 1992; 80(3):281-7.
9. Cimino C, Barnett GO, Blewett DR *et al.*, author. Interactive query workstation: a demonstration of the practical use of UMLS knowledge sources . Proceedings - The Annual Symposium on Computer Applications in Medical Care. 823-4.
10. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993; 32(4):281-91.
11. Humphreys BL, Lindberg DAB. Bull Med Libr Assoc 1993; 81(2):170-7.
12. Chute CG, Yang Y. An overview of statistical methods for the classification and retrieval of patient events. Methods Inf Med 1995; 34(1-2):104-10.
13. Sievert ME, McKinin EJ, Johnson ED, Reid JC, Mitchell JA. Beyond relevance--characteristics of key papers for clinicians: an exploratory study in an academic setting. Bull Med Libr Assoc 1996; 84(3):351-8.

Table 2- Free Information Sources in the ISM with comments describing our experience with attempts to access them.

E00031 – AudioVisuals Online (AVLINE)		
Protocol: Telnet	Script: Yes	Script is not precise: takes us into a menu after 2 additional steps.
E00034 - Catalog Online (CATLINE)		
Protocol: Telnet	Script: Yes	Script is not precise: takes us into a menu after 2 additional steps.
E00036 - Directory of Information Resources Online (DIRLINE)		
Protocol: Telnet	Script: Yes	Script is not precise: takes us into a menu after 2 additional steps.
E00039 - Serials Online (SERLINE)		
Protocol: Telnet	Script: Yes	Script is not precise: takes us into a menu after 2 additional steps.
E00048 - Online Mendelian Inheritance in Man (OMIM)		
Protocol: FTP	Script: No	Takes us into a list of directories and files. It has a *readme* file which contains some information on the content of the files and directories, as well as how to make some queries.
Protocol: Gopher	Script: Yes	Script is fine. It takes us into a search page (expects a word).
Protocol: WAIS	Script: No	Unable to connect.
Protocol: WWW	Script: No	Takes us to a page with multiple links. The page gives information on how to create scripts for automatic queries to the database using the web-search engine.
E00052 - Nucleic Acid Sequence Data Bank (GenBank)		
Protocol: Gopher	Script: No	Takes us into a search page (expects a word).
Protocol: WWW	Script: No	Opens a page with multiple links (different search engines). The ISM states that this is a search menu.
E00054 - EMBL Nucleotide Sequence Database (EMBL)		
Protocol: WWW	Script: No	Opens a page containing multiple services.
E00055 - DNA Data Bank of Japan (DDBJ)		
Protocol: Gopher	Script: No	Takes us into a search page (expects a word)
Protocol: WWW	Script: No	Opens a page with a series of services, information, and links.
E00064 - National Environmental Data Referral Service (NEDRES, NEDS)		
Protocol: Gopher	Script: Yes	Script describes a possible WAIS connection. Opens a page with a list of directories and files.
E00073 – Technology Transfer Network (TTN)		
Protocol: Telnet	Script: No	Takes us into a series of questions (steps) before the main menu.
E00075 - Health Service/Technology Assessment Texts (HSTAT)		
Protocol: WWW	Script: No	Takes us to a page containing information and also a search engine.
E00083 - Online Images from the History of Medicine (IHM, OLI/HMD, OLI)		
Protocol: WWW	Script: No	Page not found.
E00082 - Visible Human Project (Visible Woman, Visible Man, Visible Human)		
Protocol: WWW	Script: No	Opens a page with the project description and a series of links to image viewers.