Automated Information Retrieval in Science and Technology

Tamas E. Doszkocs, Barbara A. Rapp, Harold M. Schoolman

Summary. The rapid advances in computer and communication technology in the 1970's have enabled large interactive scientific and technical information retrieval systems to be implemented. Major search services today offer on-line access to millions of bibliographic citations and an increasing number of "electronic handbooks." In addition, development of knowledge bases is well under way. Despite the impressive speed and flexibility of interactive retrieval systems, their impact has been lessened by limited awareness of their existence, uneven quality of retrieval, inadequate linkages among data bases, and reliance on specially trained intermediaries.

In the past 15 years, automated information retrieval systems for science and technology have been developed at a remarkable rate. Today more than 1000 data bases are available for computerized searching. Although more than 2 million searches of these data bases will be made this year, many scientists and technologists are either unaware of them or use them in a superficial manner. This article describes the development of data bases at the National Library of Medicine (NLM) to illustrate the evolution, present capabilities, and potential of automated information retrieval systems.

In 1979, the NLM celebrated the centennial of John Shaw Billings' launching of *Index Medicus*, derived from his monumental *Index Catalogue of the Surgeon General's Office*. Billings, with help from Robert Fletcher, indexed 20,000 articles selected from 570 medical journals throughout the world. Today, the NLM indexes more than 20,000 articles a month from approximately 3,000 journals selected from the more than 20,000 received.

In the late 1950's F. B. Rogers, then director of the NLM, recognized that the usefulness of the library's bibliographic publications was being threatened by the ever-increasing volume of medical literature to be processed. He therefore initiated the mechanization that became MEDLARS (Medical Literature Analysis and Retrieval System) to support the library's bibliographic publications. MEDLARS became operational in 1964. Aided by GRACE (Graphic Arts Composing Equipment) (*1*), the August 1964 *Index Medicus* was electronically typeset in 18 hours.

The automated bibliographic data base used for publications was also available for machine searching. The retrieval system permitted complex searches to be processed in a batch mode (that is, without direct interaction with the computer). In 1970, at the height of this activity, approximately 18,000 searches were made. The turnaround time to users was 40 to 60 days.

In 1970 and 1971, the NLM conducted the first experiments in on-line access to its bibliographic data base (*2*). In 1971, MEDLINE (MEDLARS on-line) became operational. Interactive searching from remote terminals in 10 regional libraries and 14 large academic medical libraries was now possible. Today the service supports on-line access by more than 1000 domestic and foreign centers conducting about 1.5 million searches of the NLM's 19 data bases each year (Table 1).

By 1973, the NLM's MEDLINE experience (*3*) had demonstrated the potential for commercial development of other on-line scientific search systems and services. Rapid advances in computer and communications technology made widespread availability and use of on-line retrieval systems feasible.

The development of telecommunications networks has greatly reduced the cost of on-line systems. In the United States, these networks provide links to computers throughout the country. The connection is typically established by a local telephone call to the nearest network access node.

Although most on-line data bases contain bibliographic citations to the published literature, many data bases also provide numeric data that can be used to answer a specific question. "Knowledge" bases, which contain an analysis and synthesis of published information in a given field, are also beginning to emerge.

Bibliographic data bases are roughly analogous to printed indexes such as *Index Medicus* or to a library's card catalog; data banks may be thought of as automated reference manuals; and knowledge bases are rough equivalents of textbooks or state-of-the-art reviews. Each of these data base types is discussed in greater detail below.

Bibliographic Data Bases

Currently there are 528 publicly available bibliographic or bibliographic-related data bases (*4*). They contain more than 70 million citations or records and span many subject areas, including life sciences, chemistry, agriculture, energy, the environment, engineering, electronics, physics, geoscience, astronomy, toxicology, and pharmacology. The majority can be searched on-line.

In the United States, on-line access to a number of data bases is provided by three major commerical vendors: Lockheed Information Systems, Bibliographic Retrieval Services (BRS), and Systems Development Corporation (SDC). There is considerable overlap among these vendors. Of the 121 data bases, 112 are offered by Lockheed, 60 by SDC, and 30 by BRS. The cost of these services ranges from about $8 to $120 for each hour the user is connected to the vendor's computer. An additional charge is made for the use of the telecommunications network. This charge varies from approximately $3 to $6 per hour and is not affected by the distance involved. Access to some data bases is offered only by their producers. The NLM serves as both data base producer and distributor and offers access to 19 different data bases (*5*).

Dr. Doszkocs is Chief of the Technical Services Division, Ms. Rapp is a librarian in the Bibliographic Services Division, and Dr. Schoolman is Deputy Director for Research and Education at the National Library of Medicine, Bethesda, Maryland 20209.

0036-8075/80/0404-0025$01.50/0

Bibliographic data bases contain references to the published literature and are most often used as tools that guide one to a journal report. For the most part, these systems are on-line versions of existing indexing and abstracting services such as *Engineering Index*, *Chemical Abstracts*, *Index Medicus*, *Science Citation Index*, and *Government Reports Announcements Index*. Some combine information from several sources and others provide access to a part of only one data base.

Boolean logic is used in searching bibliographic data bases. Several search terms can be combined by using the set operators AND, OR, and NOT so that sophisticated searches can be made. Printed indexes rely on human or machine-aided indexing with carefully controlled vocabularies (thesauri). Although conventional indexing is used in many on-line data bases, some can also be searched by using words or parts of words that actually appear in the text. For instance, while some automated data bases provide only title, author, and source information, many include a short abstract of the article or report. In this case, the text words in the abstract as well as in the title may be searchable.

In MEDLINE, access is possible through authors, index terms from a thesaurus, or words in the title or abstract. A search can be limited to certain publication years, languages, or journal titles. The searcher may also specify sex, age or research with human or animal subjects. Either specialized or general bibliographies can be prepared. In addition, the current month's additions to the data base may be searched as a separate file. This file is called SDILINE (Selective Dissemination of Information on-line).

Complex MEDLINE searches are ordinarily performed by trained intermediaries. The NLM provides a comprehensive training program for its users, including a computer-aided instruction package. Figure 1 shows a typical MEDLINE interaction between the searcher and the retrieval program.

The Institute for Scientific Information provides on-line access to its *Science Citation Index* and *Social Science Citation Index* through the Lockheed Information Systems' DIALOG system. These two data bases, which have the same features as other bibliographic data bases, can also be searched for authors and documents that cite or are cited by the individual journal articles. This capability represents an important enhancement of subject-based retrieval, since it provides rich associative pathways among scientists and scientific research developments.

Although bibliographic data bases similar to those mentioned above are by far the most numerous, others are rapidly developing. Particularly useful are those that provide summaries of ongoing research projects and names and addresses of the principal investigators. The NLM, in cooperation with the National Cancer Institute, maintains two such files relating to cancer research: (i) CANCERPROJ, containing approximately 16,000 summaries of ongoing cancer research projects in many countries and (ii) CLINPROT, containing summaries of clinical investigations of new anticancer agents and treatments. Two other such files are the Smithsonian Science Information Exchange data base (available through SDC), which includes information about ongoing government-funded research projects, and the Current Research Information System (available through the Lockheed DIALOG system), which contains data on research in many fields.

Table 1. Data bases of the NLM and their contents.

Data base	Total records	Dates covered
AVLINE	7,680	Through 1979
BACK66*	501,802	Jan. 1966 to Dec. 1968
BACK69*	668,258	Jan. 1969 to Dec. 1971
BACK72*	669,109	Jan. 1972 to Dec. 1974
BACK75*	642,953	Jan. 1975 to Dec. 1978
BIOETHICS	7,733	Jan. 1963 to Sept. 1979
CANCERLIT	183,433	1976 to 1979
CANCERPROJ	18,641	1965 to 1979
CATLINE	191,053	N/A†
CHEMLINE	425,112	N/A
CLINPROT	1,520	N/A
EPILEPSY	25,635	1945 to present
HEALTH	125,608	Jan. 1975 to Aug. 1979
HISTLINE	37,256	N/A
MEDLINE	573,960	Jan. 1977 to Oct. 1979
MESH VOC	14,819	1979
NAME AUTH	105,873	1979
RTECS	36,851	1978
SDILINE	21,543	October 1979
SERLINE	33,122	1979
TDB	2,515	N/A
TOXLINE	628,743	1950 to 1979
CBAC	326,675	1974
TOXBIB	123,868	1974 to Oct. 1979
IPA	35,265	1974 to July 1979
HEEP	75,876	1974
PESTAB	16,063	1974 to June 1979
EMIC	24,290	1960 to June 1979
ETIC	15,301	1950 to Mar. 1979
RPROJ	9,567	Nov. 1979
TOXBACK	379,299	1940 to 1975
CBAC	167,668	1965 to 1973
TOXBIB	127,104	1968 to 1973
IPA	19,188	1970 to 1973
HEEP	24,680	1971 to 1973
HAPAB	12,816	1966 to 1973
HAYES	10,043	1940 to 1966
TMIC	4,552	1971 to 1975
TERA	13,248	1960 to 1974

*MEDLINE back files. †N/A, not available.

A major problem in performing on-line searches is that of choosing appropriate search terms. The searcher must think of all the possible ways to express a concept in anticipation of the words chosen by the author and indexer. An on-line dictionary file can be very useful in this respect, allowing synonyms for one concept to be pulled together. This is especially important in specialized areas such as chemical nomenclature, in which there are more than 7 million chemical substances and 40 to 50 ways of identifying each compound.

The NLM has two on-line, interactive dictionary files: CHEMLINE (chemical dictionary on-line) and MESH (medical subject headings). CHEMLINE provides a mechanism whereby more than 760,000 chemical names representing nearly 415,000 compounds can be searched and retrieved on-line. This file contains Chemical Abstracts Service (CAS) registry numbers, molecular formulas, preferred chemical index nomenclature, generic and proprietary names derived from the CAS registry nomenclature file, and a locator designation that points to other files in the NLM system containing information on a particular chemical substance. Where applicable, each registry number record in CHEMLINE contains ring information including the number of component rings within a ring system, ring sizes, ring elemental compositions, and component line formulas. The user searches CHEMLINE by entering either a chemical name, generic name, trivial name, commercial name, molecular formula, or even a part of a name. The registry number information or other nomenclature can then be used in performing more thorough searches of the other NLM files (for example, TOXLINE, a file of more than 600,000 journal citations and abstracts pertaining to toxicology and the environment).

The MESH vocabulary file is an on-line thesaurus of medical terminology used in indexing journal citations for MEDLINE. It is thoroughly cross-indexed (for example, GUINEA WORM, see DRACUNCULUS MEDINENSIS; AVOIDANCE LEARNING, see related ESCAPE REACTION; MIREX, see under INSECTICIDES, ORGANOCHLORINE).

Data Banks

Data bases that contain numeric and analytic data derived from the published literature and references to the source of the information are beginning to appear. These systems are comparable to handbooks. The NLM offers two on-line data bases of this type: RTECS (Registry of Toxic Effects of Chemical Substances) and TDB (Toxicology Data Bank). The RTECS file is the on-line version of a compilation published annually by the National Institute for Occupational Safety and Health, and contains toxicity data for approximately 36,000 substances known by more than 125,000 different names. The file also contains threshold values, recommended standards in air, aquatic toxicity, some listings of toxicological effects, and the

```
PROG:
YOU ARE NOW CONNECTED TO THE MEDLINE FILE.

SS 1 /C?
USER:                                   (Enter medical subject headings to express primary concept.)
thromboembolism or cerebrovascular disorders
PROG:
SS (1) PSTG (1631)

SS 2 /C?
USER:
contraceptives, oral                    (Enter medical subject headings to express primary concept.)
PROG:
SS (2) PSTG (760)

SS 3 /C?
USER:                                   (Match retrieved sets of records.)
1 and 2
PROG:
SS (3) PSTG (40)

SS 4 /C?
USER:                                   (Enter next major concept.)
middle age
PROG:
SS (4) PSTG (54386)

SS 5 /C?
USER:
3 and not 4                             (Exclude the latter item from the retrieved set.)
PROG:
SS (5) PSTG (28)

SS 6 /C?
USER:
print 1 include abstract skip 11        (Print the 12th record with full abstract.)
PROG:

12
AU  - Hilliard GD
AU  - Norris HJ
TI  - Pathologic effects of oral contraceptives.
AB  - The pathologic effects of oral contraceptives have been described
      in this paper and in other reviews [1, 5, 23, 35, 44, 48, 59, 80,
      103, 114]. Approximately 10 million women currently use oral
      contraceptives in the United States. These drugs are beneficial
      both to the users and for population control. It is their effect
      on the health status of women who take them that must continue to
      have well-organized investigation so that more meaningful
      conclusions concerning their safety will permit continued use. In
      some instances, the pathologic effects of oral contraceptives
      make it necessary that new methods of contraception be found.
      Intensive research in this area is needed and judicious use of
      oral contraceptives must be maintained. A national registry
      should be formed to record and investigate the cases of women who
      die or have adverse reactions while taking these agents. A
      registry might identify associations no previously known to exist
      in patients taking oral contraceptives. It would serve to
      concentrate the data in one area so that more material would be
      available for the study of pathogenetic mechanisms. It would
      heighten patient and physician awareness of the untoward effects
      and increase the responsibilities of the women who take them to
      monitor their own health.
SO  - Recent Results Cancer Res 1979;66:49-71

SS 6 /C?
USER:
stop
PROG:
DONE? (YES/NO)

USER:
yes
TIME  0:01:58     NLM TIME 17:44:01
. . .

PROG:

GOOD-BYE!
```

Fig. 1. MEDLINE search on the topic "thromboembolism or cerebrovascular disorders due to the use of oral contraceptives by persons who are not middle-aged."

sources from which the data were taken. The file can be used to retrieve information on a particular chemical substance, species, route of administration, and effect. Figure 2 shows a sample RTECS record.

The TDB contains chemical, pharmacological, and toxicological information and data from some 80 major textbooks and handbooks. The data were reviewed by scientists knowledgeable in the subject matter. The file contains data on 1122 chemical substances and information on some 1500 additional substances for which data are being collected. Thus approximately 2600 substances are in the TDB file. The TDB contains more than 60 different data elements (for example, synonyms, molecular formulas, and mechanisms of action).

A variant is the recently developed Laboratory Animal Data Bank (LADB). It obtains data directly from participating laboratories rather than from published literature. Numeric data bases such as LADB typically permit interactive statistical analysis of the stored data and have computational and report-generation capabilities. They are often used by policymakers, planners, and researchers in making decisions. Using LADB, a scientist may (i) select and examine physiologic and pathologic baseline data for various groups of animals; (ii) determine the environmental and husbandry conditions for each animal group selected; (iii) statistically analyze the obtained data; and (iv) cause the data to be printed as distributions (such as histograms) or as complete reports.

SOURCE IDENTIFICATION	NIOSH/AB1060000
PRIME NAME	ACENAPHTHENE, 5-NITRO-
CAS REGISTRY NUMBER	602-87-9
CLASS OF COMPOUND	CARCINOGEN-NEOPLASTIGEN
TOXICOLOGY/CANCER REVIEW	CARCINOGENIC DETERMINATION:ANIMAL POSITIVE IARC** IARC MONOGRAPHS ON THE EVALUATION OF CARCINOGENIC RISK OF CHEMICALS TO MAN. 16,319,78
STATUS	NCI CARCINOGENESIS BIOASSAY COMPLETED AS OF SEPT 1978
SYNONYMS	ACENAPHTHYLENE, 1,2-DIHYDRO-5-NITRO-
SYNONYMS	1,2-DIHYDRO-5-NITRO-ACENAPHTHYLANE
SYNONYMS	5-NAN
SYNONYMS	NCI-C01967
SYNONYMS	5-NITROACENAPHTHENE
SYNONYMS	5-NITROACENAPTHENE
SYNONYMS	5-NITRONAPHTHALENE ETHYLENE
MOLECULAR FORMULA	C12-H9-N-02
MOLECULAR WEIGHT	199.22
WISWESSER LINE NOTATION	L566 1A LT&&J HNW
TOXIC DATA SOURCE	TJIDAH Tokyl Jikeikai Ika Daigaku Zasshi. Tokyo Jikeikai Medical Journal. 89,475,74
TOXDATA KEYWORDS	ORAL;RAT;RODENTS;TDLo;120 gm/kg/17W-C;TOXIC EFFECTS:CARCINOGENIC
TOXIC DATA SOURCE	BJCAAI British Journal of Cancer. 30,481,74
TOXDATA KEYWORDS	ORAL;HAMSTER;RODENTS;TDLo;504 GM/kg/24W-C TOXIC EFFECTS;NEOPLASTIC

Fig. 2. Sample RTECS record, as it appears on the on-line file.

Knowledge Bases

The newest form of automated information retrieval system is based not on bibliographic records or numeric data, but on analyzed and synthesized "knowledge." In the Hepatitis Knowledge Base, a prototype of this kind of system, knowledge pertaining to viral hepatitis is synthesized from recent reviews by experts in the field (*6*). Relevant information is selected, placed in a highly organized hierarchical arrangement to permit easy retrieval, and encoded into a minicomputer. The content of the data base is validated by consensus of a group of ten experts in the field of viral hepatitis. Dissenting points of view are included when they reflect current understanding. This knowledge base is updated monthly.

The Electronic Information Exchange System, an experimental computer conference network developed by the New Jersey Institute of Technology and supported by the National Science Foundation (NSF), serves as the principal medium of communication linking experts with one another and with the NLM staff (*7*).

Limitations of Retrieval Systems

Use of these systems is increasing as they are improved and become better known. However, the systems are not yet widely used. Even when used, they are frequently not exploited to their fullest capacity. Part of the difficulty is that even though many of the systems do not require that the users be formally trained, most searches are performed by trained "search analysts." Such delegation of the search function is inevitable given the complexity of and differences among existing services. As Williams (*8*) noted,

> the data bases vary with respect to subject coverage, source types (journals, monographs, patents, theses, book reviews, etc.), file format, record format, data elements included, and indexing or vocabulary practices. . . . On-line systems vary with respect to command languages, protocols, and system responses. Retrieval systems vary with respect to search features, system features, and output formats.

This variability is a serious hindrance, since it is unlikely that users will become familiar with all the files and systems they might need. Even trained searchers find it difficult to be fully conversant with several data bases and on-line systems.

There are other barriers to widespread direct use of these systems by scientists. Paradoxically, the superhuman speed and flexibility of computer searching often overwhelms the searcher with instant bibliographies of hundreds or sometimes thousands of citations, creating a formidable reading burden. Critical perusal may be made difficult by the lack of informative text and data in retrieved records.

Moreover, it is likely that the search will not yield 100 percent recall (the proportion of relevant documents retrieved from all the potentially relevant references) or provide 100 percent precision (the proportion of relevant documents in

the actually retrieved set of records). In practice, precision and recall are inversely related and are rather elusive measures in operational retrieval systems because of the subjectivity of relevance judgments. The scientist can be further frustrated when he or she cannot locate the articles listed in the printout.

Admittedly, none of these problems is unique to computer searching. Nonetheless, until on-line retrieval systems surpass their conventional manual counterparts in overall ease, quality, and cost-effectiveness, many scientists will continue to rely on the more traditional means of gathering information.

Directions in Research and Development

Much of the recent research and development in the field of information science has been directed toward overcoming the remaining barriers to effective use of large computerized retrieval systems. Significant accomplishments can be noted in three critical areas: (i) data base and terminology selection, (ii) common access and retrieval protocols, and (iii) development of natural language–user interface techniques.

Data Base and Terminology Selection

Because of the proliferation of data bases and the interdisciplinary nature of much scientific research, the selection of appropriate data bases is an important step in the search process. This has led the commercial data base vendors to offer data base selection aids within their individual search systems. Examples are the DIALIST printed indexes to the many files in the Lockheed DIALOG system, the Data Base Index on-line file of the SDC ORBIT system, and the BRS CROSS file index search capability.

Developmental work for generalized intersystem data base selection has been pursued at the Coordinated Science Laboratory at the University of Illinois (*9*), and a prototype Chemical Data Base Directory (CDBD) system has been implemented at the NLM. The CDBD is a switching file for the emerging chemical substances information network that is being designed to effect the logical integration of diverse data bases and search systems.

As noted earlier, it is difficult to choose all the appropriate search terms for a topic in a specific file or across different files. Thesauri are designed to assist in this task, but they are limited in scope and quickly become dated. Doszkocs (*10*) successfully demonstrated the feasibility of automatically identifying and displaying terminology associations for very large files such as MEDLINE and TOXLINE. The associations are based on the observed and expected frequency with which terms appear in retrieved sets of records. The Associative Interactive Dictionary (AID) software can also be used to capture automatically and display a variety of other associations, such as the interest profile of a given researcher or a ranked list of chemical substances associated with a specific adverse biological effect.

The NSF-sponsored automatic subject-switching work conducted at Battelle Columbus Laboratories (*11*) is aimed at establishing topical linkages among entries in major thesauri. Linkages have been developed among entries in the Thesaurus of Engineering and Scientific Terms, the Defense Technical Information Center Thesaurus, and others. The ultimate objective is to help searchers overcome the inherent variability and ambiguity of language.

Common Access and Retrieval Protocols

The multiplicity of bibliographic search systems poses a major obstacle to convenient and effective on-line access to the scientific literature. Despite a noticeable trend toward standardization of retrieval systems, important differences continue to prevail. The Conversion for Network Information Transfer (CONIT) research project at the Massachusetts Institute of Technology (*12*) demonstrated the feasibility of a network access and retrieval interface to four different operational on-line search systems. CONIT incorporates a common retrieval language, hidden access protocols, and extensive instructional dialogue.

Natural Language–User Interface

A variety of man–machine interface techniques have been developed to provide access to well-defined and highly structured computerized data bases, such as the numeric data bases and gen-

```
MEDLINE CURRENT INFORMATION TRANSFER IN ENGLISH

    PLEASE ENTER YOUR SEARCH QUESTION
:Recombinant DNA guidelines at NIH and other research institutions

     573760 RECORDS SEARCHED
       95 CITATIONS FOUND

**********************************************************************
*** RECORD NUMBER    1  WEIGHT= 20  MAXIMUM WEIGHT= 25  ***
**********************************************************************
CITATION #          = '79010565'
AUTHOR              = 'Dickson D'
TITLE               = 'NIH confirms violation of recombinant DNA research
                    = guidelines [news]'
JOURNAL TITLE       = 'Nature'
PAGINATION          = '5'
PUBLICATION DATE    = '4 May 78'
VOLUME-ISSUE        = '273'

    C O N T I N U E   P R I N T I N G  (Y/N)?
yes


**********************************************************************
*** RECORD NUMBER    2  WEIGHT= 18  MAXIMUM WEIGHT= 25  ***
**********************************************************************
CITATION #          = '78223558'
AUTHOR              = 'Dickson D'
TITLE               = 'NIH relaxes recombinant DNA guidelines.'
JOURNAL TITLE       = 'Nature'
PAGINATION          = '303'
PUBLICATION DATE    = '27 Jul 78'
VOLUME-ISSUE        = '274'

    C O N T I N U E   P R I N T I N G  (Y/N)?
no

    SEARCH COMPLETE
```

Fig. 3. MEDLINE search assisted by CITE.

eralized data base management systems commonly used in business applications. The methods range from structured menu selection (in which the user is offered a series of choices that narrow down the subject) to English-like query languages and sophisticated but restricted language-understanding systems. The special challenge to producers of bibliographic and other textual search systems lies in the fact that the text portion of the documents is written in English or some other versatile but redundant and ambiguous natural language. There is little doubt that the scientist also prefers using the same natural language in expressing search topics of interest. Experience shows that most scientists are unwilling to learn the intricacies and subtleties of access protocols, command languages, Boolean search strategy formulation, and controlled vocabularies.

A citation is unique and unambiguous, but numerical data or analytic statements are dependent on context for their utility. Moreover, if we hope to achieve simplified retrieval systems that are based on the use of natural language, the ambiguities of syntax and grammar will have to be dealt with. Recent work at the NLM (*13*) resulted in a prototype of an English language interface to MEDLINE, TOXLINE, and the Hepatitis Knowledge Base. No special training is required to use this system, named Current Information Transfer in English (CITE). Questions may be posed in English; the software then searches for documents that contain all or most of the key terms in the query. By using a special algorithm based on combinatorial term weights, the system then ranks the retrieved records according to their relevancy. The searcher can select documents of interest and can command the system to expand the search by finding other items that contain terminology similar to the selected citations. Figure 3 shows an actual search on MEDLINE in which CITE was used. Although the system does not perform syntactic and semantic analysis (it cannot "think"), its comparative simplicity and performance offer a genuine potential for vastly increased interactive access to the literature of science by the scientists themselves.

Conclusions

Although this article attempts to identify the many advantages of full use of existing automated data bases, it also points out some of their shortcomings. There are many problems attendant on information retrieval systems in science and technology, and we have not discussed all of them. For example, what criteria should be used to determine whether a datum or statement is a valid entry to a data base? The NLM tends to rely on the published literature. Other systems rely on patient records or expert opinions. But regardless of which is used, there immediately arises a second set of questions: What published literature? Which patient records? Whose expert opinion? These problems become even more complex when information in data bases is updated. As the number of data bases grows, more sophisticated computational methods will be required to ensure complete updating of the many data bases. The criteria need not be the same for all data bases, but they must be clearly defined in each case.

There are those who believe that the increasing amount of scientific and technical research will create a volume of information so large as to frustrate the very purpose for which it was created. If this prediction is not going to become a reality, then a larger percentage of the resources now expended on generating scientific and technical information must clearly be invested in research on how to handle the mass of information being generated.

References and Notes

1. L. Karel, C. J. Austin, M. M. Cummings, *Science* **148**, 766 (1965).
2. D. B. McCarn and R. V. Katter, in *Interactive Bibliographic Search: The User/Computer Interface*, D. E. Walker, Ed. (American Federation of Information Processing Societies, Montvale, N.J., 1971), pp. 121–141.
3. D. B. McCarn and J. Leiter, *Science* **181**, 318 (1973).
4. M. E. Williams and J. H. Rouse, *Computer-Readable Bibliographic Data Bases: A Directory and Data Sourcebook* (American Society for Information Science, Washington, D.C., 1979).
5. *Med. Inf.* **3**, 3 (1978).
6. L. M. Bernstein, E. R. Siegel, W. H. Ford, *Proc. 2nd Annu. Symp. Comput. Appl. Med. Care* (1978), p. 366; A. D. Merritt, E. R. Siegel, C. M. Goldstein, L. M. Bernstein *Proc. 3rd Annu. Symp. Comput. Appl. Med. Care* (1979), p. 752.
7. E. R. Siegel, *Proc. Am. Soc. Inf. Sci.* **16**, 124 (1979).
8. M. E. Williams, *On-line Rev.* **2**, 353 (1978).
9. _____ and J. E. Preece, *Proc. Am. Soc. Inf. Sci.* **14**, 34 (1977).
10. T. E. Doszkocs, *ibid.* **15**, 105 (1978).
11. R. T. Niehoff, *J. Am. Soc. Inf. Sci.* **27**, 3 (1976).
12. R. S. Marcus and F. J. Reintjes, *Experiments and Analysis on a Computer Interface to an Information Retrieval Network* (Rep. LIDS-R-900, Massachusetts Institute of Technology, Cambridge, 1979).
13. T. E. Doszkocs and B. Rapp, *Proc. Am. Soc. Inf. Sci.* **16**, 131 (1979).