Planning for On-Line Bibliographic Access by the Lister Hill National Center for Biomedical Communications

BY DAVIS B. MCCARN, *Deputy Director*

Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland

ABSTRACT

Lister Hill Center is concerned with developing a computerized information system, with a data base consisting of an expanded *Abridged Index Medicus,* using part of a large computer system, and connecting this system to the TWX network. With this in view, a study was made of the demands on such a network, and a study of the SUNY network showed that on-line access more than doubled the use of the data base.

Other alternatives to the TWX system are described.

THE explosive growth of the world's medical literature makes it virtually impossible for physicians and scientists to satisfy their information needs without efficient indexing and abstracting services which lead them to relevant portions of the literature" (1).

A study (2) of interlibrary loans indicates that the half life of the journal literature in medicine is about ten years. The rate of growth of the literature appears to have held relatively constant over the past twenty years and to have been independent of the level of federal support available for medical research. Thus, while there may be a slackening in federal support to medical research, it seems reasonable to assume that there will probably not be a slackening in the growth of the medical literature. An improvement in bibliographic access to this volume of literature would seem to be an effective way of increasing the utility of medical knowledge.

In this context, the Lister Hill Center believed that it was making a significant contribution to health practitioners, educators, and researchers by finding a way or ways of providing simple, remote, on-line access to medical bibliographic information. A process as simple as the card catalog and usable from anywhere in the nation appeared technically possible but not obviously practicable. Bibliographic access is not the same as access to the literature. Such access only helps identify desired material; other procedures and techniques would be required to actually make the literature available once the relevant articles have been identified.

The major problems in the development of such a system include assessing the benefit to justify the experiment, providing the data base, providing the appropriate computer system, and connecting the communications and access terminals. There is no doubt these problems could be solved with sufficient funds; the final problem was whether a meaningful effort could be made for very little money. Aspects of these problems are discussed below, but we do believe a way has been found to provide a useful and meaningful experimental service. The essential facets of this solution are (1) use of the NLM data base for the *Abridged Index Medicus* with slight augmentation to cover the bibliographic information on the most used medical literature, (2) use of part of a large, time-shared, computer system rather than all of a smaller one, and (3) connection of this system to the TWX network thus providing 40,000 potential access terminals at no cost to the government and shifting the communication costs to the user.

But is there really enough benefit to be gained?

DEMAND

The experience of many information systems would indicate that except in unusual circumstances where the information is of immense practical or financial value, the demand for such services is very dependent on availability and ease of access and use.

Experience with a system comparable to the National Library of Medicine's Medical Literature Analysis and Retrieval System (MEDLARS) (3) at the Foreign Technology Division of the Air Force Systems Command indicated that remote access can triple or quadruple the use of a bibliographic data base. Thus, it seemed of substantial importance to look at what the factor might be in the medical community. One operational, on-line medical literature bibliographic retrieval system exists: the system of the State University of New York (SUNY) at Syracuse. That system has now been in operation since December 1968. The author has conducted an analysis of that system in terms of use of the MEDLARS data base.

On-line searching of the MEDLARS data base at SUNY has resulted in increased use of that data base. There are a variety of factors which must be considered, however, to demonstrate such increased use. Figures available for April of 1969 indicate that there were approximately 2,400 searches run on the SUNY system and 1,036 run on all other U.S. MEDLARS demand search systems combined. (There are five such centers which actually process computer searches.)

While there were substantially more searches on the SUNY system, these searches do not cover comparable size data bases. The normal search on the MEDLARS system covered the period January 1966 to April 1969, and searched on the average about 636,000 citations. The SUNY data base, on the other hand, has to be segregated to fit on the disk file storage. The maximum data base is roughly 90,000 citations; thus, each search on the SUNY system covers a substantially smaller number of citations than does a search on the MEDLARS system. A final factor considered in assessing the benefits was the total number of doctors in New York State as compared to those throughout the nation. Approximately one-seventh (1/7.6) of the active M.D.'s reside in the state of New York. Parenthetically, it should be noted that some requests on the SUNY system do come from the National Library of Medicine and from the Countway Library in Massachusetts, and, therefore, the total fraction of the physician population may be slightly in excess of one-seventh, but the majority come from New York and this seems a reasonable estimate.

On the basis of these figures, it would seem reasonable to assess the relative utility of the two data bases as equal to the ratio of their search rate per month, adjusted for the number of citations searched, and also adjusted for the number of physicians who could reasonably be using the service, i.e.,

$$\text{Relative Use} = \frac{\text{No. of requests on 1}}{\text{No. of requests on 2}} \times \frac{\text{Coverage of Data Base 1}}{\text{Coverage of Data Base 2}} \times \frac{\text{No. of potential Users of 2}}{\text{No. of potential Users of 1}}$$

This calculation for SUNY vs. MEDLARS would read:

$$\frac{2400}{1036} \times \frac{90}{636} \times 7.6 = 2.5$$

or the relative use of the SUNY system is 2.5 times that of MEDLARS. Thus it would appear that the availability of an on-line system in libraries has doubled the use of the citation information in the system. Use is one measure of the value of a service, and one can conclude that the on-line service is twice as useful as the standard MEDLARS service.

The evidence is clear that the on-line system at SUNY increases use, but how much of the total demand does it satisfy? In order to evaluate this in at least a rudimentary manner, the SUNY data were examined from another point of view. The access terminals for the SUNY system are concentrated in four cities; these cities have provided the major use of the system. Use in these cities for a three-week period during April to May 1969 is shown in Table 1.

Syracuse, as the home of the system, has significantly higher use per physician than any other city. If one drops this city, the remaining cities show a use rate of 2.3 searches per physician per year.

Several objections can, of course, be raised to this somewhat simplistic estimate. Physicians are probably not the major user of the MEDLARS service. However if total population had been used for comparison and extrapolation purposes, the results are not materially different. The kind of searches actually done may also be different; certainly more exploratory searches are done on an on-line system, which would account for some of the increased usage. If all the physicians in the state had had the access to the same service the total use would have been more than three times that actually experienced which was .65 searches per physician in New York State. In particular, it would appear that the physicians in the New York metropolitan area are under-represented because of the availability of only a single terminal in the Parkinson Information Center. If one adjusts this number of searches per physician downward to reflect the coverage of the data base (i.e., the fact that up to seven searches are required to cover two and one-half years of literature), it would still appear that, given ready access, there may be a national demand for 100,000 to 300,000 searches per year, eight to twenty-four times the number now processed by the MEDLARS system.

The real existence of such a demand is evidenced by the potential emergence of local commercial services such as the one in Seattle, Washington, which proposes to provide searches on a data base of seventy journals, answering queries by phone at $5 a search. This fledgeling service generated enthusiasm in the Washington State Medical Association and in the Washington/Alaska Regional Medical Program.

Data Base

The National Library of Medicine has maintained a bibliographic access system to medical literature for nearly a century. Bibliographic data have been computerized since January 1964. A cost-analysis study done in 1966 of the cost of maintaining the data base indicated that the annual cost for the basic file itself ran a little over $400,000 a year. Considering the increasing cost of maintaining access to the literature and the length of time that the data base of NLM has been maintained and augmented, it can be estimated that the value of the current capital investment in the MEDLARS data base approximates $3 million.

TABLE 1
USE˙OF SUNY SYSTEM

City	Number of Terminals	Number of Requests	Number of Non-Federal Physicians
Albany.......	3	127	1336
Buffalo.......	2	462	2282
Rochester.....	2	186	1631
Syracuse......	4	541	1099
Total.......	11	1316	6348

This data base now consists of bibliographic information on over one million medical journal articles. This data base is too large for existing remote-access computer systems. Two analyses indicate what might be a reasonable subset for most useful on-line availability. First, a review of the sample volume of the *Abridged Index Medicus,* containing literature from a select group of 100 journals in clinical medicine, indicates that this group of 100 journals represents 10 percent of the total file; 100,000 citations is a manageable data base. Second, a review of the literature indicates that as little as a 20 percent increase in this list would cover almost all the entries on lists maintained elsewhere of most significant journals and allow coverage of other disciplines in health care. Further, most of these journals are among those so readily available in larger medical libraries that NLM refuses to fill interlibrary loan requests for them, but when NLM did fill loan requests this group of journals would have covered about 30 percent of all such requests (2).

Thus, a smaller, more manageable, data base of great usefulness can be selected from the existing system for input to an on-line service. The relatively simple computer programs to convert such an extracted MEDLARS file to a proper, decoded IBM-compatible tape have been written and tested.

Computer Technology of Bibliographic Search

The computer system technology required for effective development of remote access bibliographic system is of recent development. Remote-access, time-sharing computer systems have been under development since 1963 when both the Massachusetts Institute of Technology's Project MAC (MIT/MAC) and System Development Corporation (SDC) project were initiated. These systems, however, were primarily designed for remote scientific processing, providing effective computer support of short mathematical calculations. The problem of providing remote access to large data bases is substantially different from that of merely providing remote mathematical support. Both the MIT and SDC systems have been used in attempts to provide remote access to small data bases. The Technical Information Program (TIP), one part of the MAC project, provides such access to the literature in physics through citation indexing, but is relatively slow. The System Development Corporation has developed a special *O*n-Line *R*etrieval *B*ibliographic *I*nformation and *T*ime-Shared (ORBIT) System for handling the bibliographic literature problem which has proved more efficient than the TIP System of MIT. The SDC system was used to provide a demonstration capability starting in 1966 with literature on foreign technology. It met with substantial success, handling data bases of the order of 200,000 citations. The National Library of Medicine began experimentation with this same system in 1967 with a small data base of neurology journals and monographs, which was used to introduce the Library staff to remote bibliographic access systems. The System Development Corporation has improved the ORBIT system under contract to the Lister Hill Center to make it substantially more flexible and to bring it to an operational capability on its IBM 360/67 computer.

As a result of planning begun in 1965, the State University of New York Biomedical Communication Network at Syracuse adapted an existing IBM program system, the Document Processing Package, to on-line retrieval of MEDLARS information which, as noted above, became available as an operational capability in December 1968. Thus, at least two systems have moved toward providing the kind of remote access to large data bases which is necessary to provide effective remote bibliographic access to the medical literature.

ACCESS

The audience of potential users for such systems is still quite limited. This limitation is a result of both the kinds of equipment used to access the systems and the policies of the common carriers. The State University of New York system uses special terminal equipment which is connected by leased telephone lines to the computer; it cannot be expanded economically to a national audience without a major modification of the way in which access to the system is provided.

An alternative means of providing access through commonly used terminals, i.e., teletypes, using the dial-up telephone network has been used by the SDC system and by the MIT/MAC system. These alternative means of access also pose a substantial problem because teletypes must be leased along with the appropriate "data set" to provide access to the telephone network and the computer at a cost of approximately $50–$60 a month. Thus, widespread use of such systems presupposes a substantial investment in equipment at using locations.

The basis for this expensive course of action has never been clear. There does exist a major network, the TWX network, which interconnects teletypes across the United States through its own system. Substantial economies in cost and development time could be achieved if this network could be used to provide access to a bibliographic system. Computers have been placed on the TWX network; both SDC and MAC were for a time connected to the TWX network. However, when they were so connected, a special feature was added called "inverted frequency" which made it impossible for any regular teletype in the network to call the computer. This special feature was justified on the basis of avoiding unnecessary and mistaken calls to the computer system. All users thus had to lease new equipment or special devices to put on their existing teletypes.

The medical library community is well versed in the use of the existing TWX equipment. For example, NLM now receives over 1,200 requests for interlibrary loans a month over the TWX network. These come from several hundred different libraries across the nation. A review of the directory of the TWX network indicates that there are 120 hospitals on the network, 150 pharmaceutical firms, 125 schools including many medical schools, and a sprinkling of clinics and physicians. On the basis of this review, it seemed clear that if a computer system could be tied to the normal TWX network, an immediate user community could have access to this medical data base without additional investment for terminal equipment. One phone company has indicated a willingness to connect a computer to the TWX network.

Alternatives

Three major alternative systems were examined. These alternatives were: (1) to contract with the State University of New York Biomedical Communication Network to provide TWX access to their computer system; (2) to move either the SDC or the SUNY system to the National Library of Medicine and provide service from NLM; or (3) to provide service on an experimental basis through the time-sharing system at the System Development Corporation. The first of these alternatives, the use of the State University of New York system, was examined extensively. It is still a possible alternative provided that substantial improvement in performance could be achieved through improved hardware. However, our analysis indicated that with existing hardware the system could not support a national experiment of the kind proposed. This conclusion is based on the analysis of the performance characteristics of the SUNY system contained in Appendix 1. The results of the analysis were that the existing computer hardware and software were not fast enough to support any appreciable increase in the use and that use of this system might prejudice users and the service against it because of unavoidable service delays.

The second alternative considered was that of using the IBM 360/50 now installed at the National Library of Medicine. Cost estimates were developed in a preliminary way of the effort required to move either the SUNY system or the SDC system to the NLM 360/50. SDC indicated a total cost of over $100,000 to make their system operable on NLM equipment, and indicated, in addition, substantial additional cost for hardware that would have to be rented to allow such a transfer. Transferring the SUNY system also appeared difficult and expensive in terms of Library manpower which could not be diverted from other priority tasks. Finally, it appeared that the effort to provide such a timely service during normal duty hours would seriously disrupt the development of MEDLARS II and other operating programs on the 360/50. Accordingly, it was judged impracticable to use the NLM computer to provide this service at this time.

The third alternative of providing an experimental service through the System Development Corporation was also investigated and judged to be feasible.

Experimental Service

The Library has now contracted with the System Development Corporation to provide an experimental service in bibliographic access to a limited group of users. It will be possible for either Teletypes on the telephone network with standard 103A2 data sets or Model 33 or 35 Teletypes in the TWX network to access the computer system in Santa Monica. Five special lines will be installed by the General Telephone Company to connect the IBM 360/67 to the TWX network. SDC will provide the computer time and an additional disc storage unit to allow the provision of bibliographic access to about 150,000 citations out of the MEDLARS data base. Thus, this available storage space will be adequate to provide access to a consensus list of medical journals over the five-year period covered by MEDLARS.

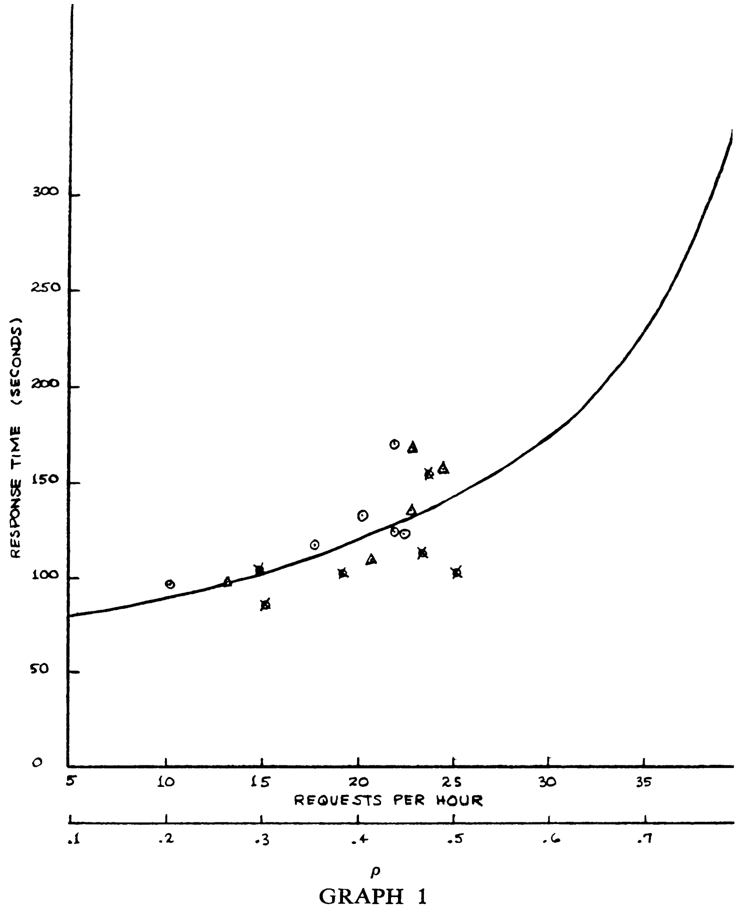
The contract provides for the conversion of a MEDLARS-produced tape on this data base to the SDC computer system and the provision of five months of service starting in April 1970. This service would be provided each day from 11:30 A.M. to 3:30 P.M. EST/EDT. The cost of calling the system would be borne by the user, and TWX charges vary from $.20 to $.60 per minute depending on whether the station is calling from across the continent or from 50 miles away. The cost over the telephone network would be somewhat less, varying from the cost of a local call to $.45 per minute. Based on previous experience, the average search might take an average of fifteen minutes. In addition, SDC will provide on-site training at a designated set of locations to users of the system, and will run a spot survey on organizations which will use the system without training. The SDC system is a system which can be used without training, and instructions are provided from the terminal giving a minimum amount of information about how to call in to the system. In addition, the language interface, the way the user phrases his request, is variable under this revised ORBIT system and can be modified during the service period as a result of the experience gained in the use of the system.

APPENDIX 1

Tables have been developed by SUNY which show number of requests from each terminal for each hour of operation and the average response time for each terminal. Only "good" searches are included in these tables, and good searches are defined to be those which take more than twenty seconds. (A response in less than twenty seconds is usually a message rejecting the search formulation.)

The software system for the SUNY system essentially consists of two parts: the first handles the console interactions and prepares a search for processing; the second part of the system takes searches one after another and edits them, then runs them against the data base, prepares the output and turns control back to the first part, which prints the response out at the console. For the purposes of this analysis it is assumed that the first part of the software, the teleprocessor, takes a small part of the time, and that the serial processing of the searches dominates the reactions of the system. If this is a reasonable approximation to the way the system behaves, then its performance should be describable in terms of a single-server queue.

The Tables show a computation fitting this single-server queue model to the SUNY data for eighteen hourly or half-hourly periods on which useful statistics are available since February 20, 1969. The estimates of the mean response time range from 60.4 sec. to 83.7 sec. with a mean of 71.9. The theoretical standard deviation of this sample should have been between 4 and 5; the actual standard deviation is



GRAPH 1
SUNY experience and fitted single server queue curve

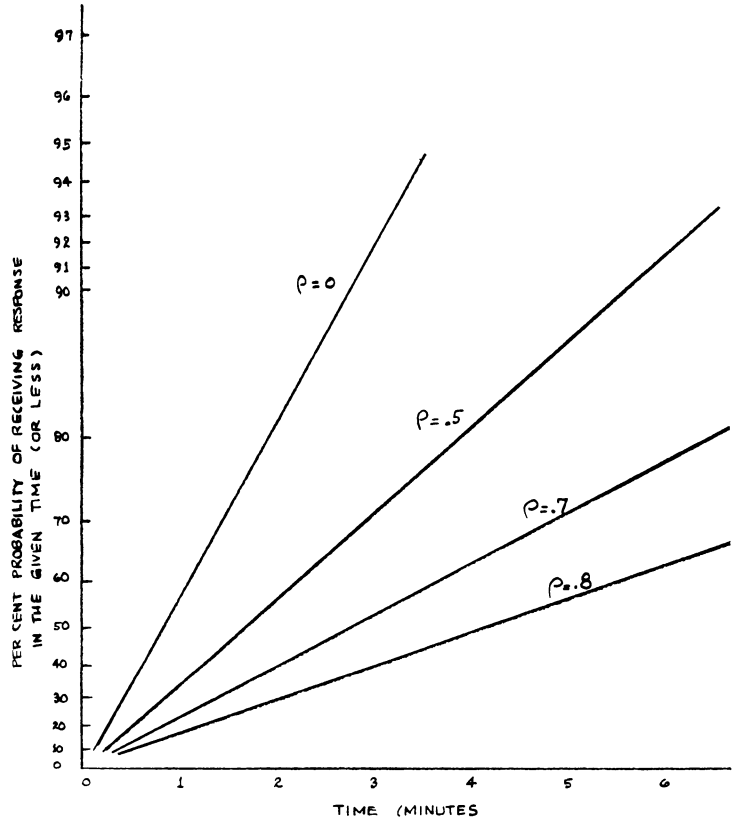
6.3, which is somewhat larger. The model, however, does not provide for the "down time" of the computer which has been a frequent occurrence and which could be expected to increase the variability of performance.

The significance of this model is shown in Graph 1. On this graph the observed values of the request rate and response time are plotted and a curve showing the mean value of response time for various request rates given a mean processing time of 71.9 sec. Also indicated on the curve are the server utilization values, ρ. The curve illustrates the rapid increase of response time as use increases; for example, at $\rho = .5$, the average response time is above 144 sec.; but at $\rho = .7$, it is 240 sec.

The meaning of these averages is shown in Graph 2. The graph shows the curves for the fraction of requests satisfied within given times at three use rates. The three rates are 50 percent, 70 percent, and 80 percent. These curves show that at 40 percent use 87.5 percent of requests are satisfied in five minutes; at 70 percent this drops to 72 percent; and at 80 percent use only 56 percent of requests are satisfied in five minutes or less. This kind of performance is similar for all queueing situations.

Looking at the actual values of response times for the system, it appears as though there were some natural barrier at about $\rho = .5$. It may be that this is the total demand, but it seems more probable that this demand is influenced by the response time and that users become unwilling to use the system when response times get large, thus providing a natural limit on the utilization of the system.

This limit is not the limit on the availability of consoles. NLM experience indicates that total console time runs five times response time. There are ten (out of twenty-one) heavily used terminals. If response time were two minutes, then console time would be ten minutes and six



GRAPH 2
Service time distribution single server queue

requests could be handled in an hour. Ten terminals could handle sixty requests. At 50 percent use this would still be thirty requests an hour substantially more than the actual use. Thus, it must be concluded that in its present configuration SUNY is processor-limited and probably could not support many more users. Improved "up time" will be some help, but probably would not alter this conclusion materially.

IBM has no data on the effect of equipment changes on the speed of the Document Processing System. Which part of the system is the limiting factor is not apparent, but disc access is the most likely problem. This latter would not be helped by upgrading the system to a 360/50.

REFERENCES

1. Cummings, M. M. Modern processing of information in the National Library of Medicine. Bull. Amer. Coll. Physicians 10(7): 326–30, July 1969.
2. Kurth, W. H. Survey of the Interlibrary Loan Operation of the National Library of Medicine. Washington, D. C., U.S. Department of Health, Education, and Welfare, Public Health Service, April 1962.
3. Austin, C. J. Medlars 1963–1967. PHS publication no. 1823. Bethesda, Maryland, U.S. Department of Health, Education, and Welfare, National Library of Medicine, [1968].

AD HOC COMMITTEE TO REVIEW THE GOALS AND STRUCTURE OF THE MEDICAL LIBRARY ASSOCIATION

The following is a list of members and their special interests. The Committee would welcome comments and suggestions from the membership.

Warren Bird	Awareness of MLA officers about the Association's problems.
Harold Bloomquist	MLA Committees.
Alfred Brandon	Central Office.
Estelle Brodman	Continuing Education.
Gwen Cruzat	Democratization of MLA Governance.
Doreen Fraser	MLA as an international organization.
Cecile Kramer	MLA furthering the profession.
Miriam Libbey	Certification and internship curriculum.
Jess Martin	National meetings.
Elliott Morse	Board liaison.
Vern Pings	Composition of MLA membership.
Irwin Pizer	Democratization of MLA governance.
Nancy Zinn	
Erich Meyerhoff	Chairman.