DOCUMENT RESUME

ED 048 905 LI 002 714

AUTHOR Williams, Martha E.
TITLE Cooperative Data Management for Information Centers.
INSTITUTION Illinois Inst. of Tech., Chicago. Research Inst.
PUB DATE 71
NOTE 13p.; Paper presented at joint meeting of Association of Scientific Information Processing Centers and National Federation of Science Abstracting and Indexing Societies, Washington, D.C., February 24, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Abstracting, Data Bases, Data Processing, Indexing, *Information Centers, *Information Dissemination, *Information Processing, *Information Retrieval, *Information Storage, Management, Problems, Standards
IDENTIFIERS ASIDIC, Association of Information Dissemination Centers, *Machine Readable Bibliographic Data Bases

ABSTRACT

The Association of Information Dissemination Centers (ASIDIC) formed the Cooperative Data Management Committee to address the problems of information center operators and data base suppliers. The number of operating centers in the U.S. is limited and their future expansion in numbers and in type of services, will depend on the education of users. Users must be trained in the new information sources and techniques and in the necessity of paying for information. Uniform data bases are essential and standards must be set for suppliers to reduce processing and conversion costs, and insure greater utilization of available data bases. The sharing of resources through a network, the repackaging of data, creation of merged data bases, creation of retrospective files, creation of personal files from data bases and distribution on nonstandard media are considerations being entertained by centers which could seriously affect data suppliers. (AB)

COOPERATIVE DATA MANAGEMENT FOR INFORMATION CENTERS

Martha E. Williams*

presented at joint meeting

Association of Scientific Information Processing Centers
National Federation of Science Abstracting and Indexing Societies

Washington, D. C.
February 24, 1971.

* Manager, Computer Search Center IIT Research Institute, 10 West 35th Street, Chicago, Illinois 60616
Chairman, ASIDIC Cooperative Data Management Committee

COOPERATIVE DATA MANAGEMENT FOR INFORMATION CENTERS

Inasmuch as the audience today is comprised of members of the National Federation of Science Abstracting and Indexing Societies and the Association of Scientific Information Dissemination Centers, visitors and new members of ASIDIC who may not be familiar with reasons for the existence of the ASIDIC Cooperative Data Management Committee my presentation will cover some of the problems being addressed by Information Center operators and data base suppliers. The existence of the problems that I will discuss, was the reason for the creation of the Cooperative Data Management Committee and it is through the efforts of the committee and subsequent cooperative actions by centers and suppliers that we hope to solve or alleviate some of the difficulties we now face.

For purposes of this presentation an information center is defined as a center that searches one or more of the machine readable data bases produced by outside suppliers. For the most part, centers and suppliers with which we are concerned are represented by ASIDIC and NFSAIS.

There are, at present, a rather limited number of operating information centers in the United States, each processing a handful of machine readable data bases and providing information services to a limited number of users in their own institutions or in outside academic and industrial institutions. If events proceed as they have in the past, we can probably expect to see the continued proliferation of published information and hence of data bases, as well as a continued growth of centers and their services--and hopefully of information users. How will these developments influence the future of information

science and its practitioners? Will the number of information centers continue to grow in an uncontrolled way or will their growth and development be organized at a higher level? What will the relationships between centers be? Will they exist as totally independent competitors, or will they cooperate in some kind of network? Will many centers offer identical coverage and services, or will each center develop its own unique area of specialization? How many data bases will there be 5 years from now? How varied will their content, format and structure be and how will they be searched?

These are rather broad questions and we can only guess at their answers--unless we do something to influence the outcome of the processes of growth and change which are now at work shaping our information centers.

Relevant developments will focus on the suppliers who provide the data bases to be searched and the centers which perform the searching, and developments will affect the consumers who use and pay for the output of centers.

One problem facing centers and suppliers is that of education. Many potential users may be reluctant to give up the inefficient, time-consuming, but agreeable and effective practices of browsing through journals and depending on the grapevine to keep informed and up-to-date. While these practices are good and should not be abandoned they are not adequate to satisfy all of the information needs of scientists and engineers. There are new sources, techniques and services and the potential users must become aware of them if they are to use them.

Education of users is crucial. As increasing numbers of abstracting journals, primary journals and data compilations are prepared by computerized typ setting, more machine searchable files come into existence as a by-product of this effort. Meanwhile, the increasing cost of manual handling of information, together with the proliferation of information and publications, would indicate that more searching will either be relegated to machines or not be done at all. To those who prefer the former alternative, it is obvious that users must be educated to the fact that machine readable data bases will be the principal search tools of the future. If students and practitioners do not become convinced of the advantages of the new sources and techniques, they will not demand them of their employers.

Unfortunately, the average scientist has never used one of the modern data bases. Therefore, significant progress in educating undergraduates, graduate students, and practicing scientists in the uses, availability, problems and benefits of machine readable information and data sources, as well as in the techniques for searching the sources, is needed to prepare the market place for the new products.

Since the centers want to sell the products of their processing efforts, it devolves on them to undertake a significant share of the education effort. Naturally, every center and supplier salesman is deeply involved in this educational effort, but the entire burden of education cannot rest on the marketing staffs of centers and suppliers. Education on a more formal basis must pave the way for marketing.

Since this task fits in well with the function of established educational institutions, where new concepts and techniques are received by the most malleable (not to mention captive) audience, it is fitting that a number of centers have been developed within or in affiliation with universities and these centers have taken steps to provide user education. For example, IIT Research Institute (IITRI). in cooperation with Illinois Institute of Technology, provides graduate and undergraduate courses in modern techniques in chemical information. IITRI also gives seminars, workshops, and short training sessions for industry, colleges and universities. Through its information center the University of Pittsburgh has given short courses and has done work in the area of on-line tutorials for profile development. And, the University of Georgia and IITRI have given workshops in profile preparation. These activities may not provide returns immediately, but they help prepare the climate for the future.

The education of potential center users involves more than training in the use of new information sources and techniques. Users must be made to realize that they have to pay for information. People accustomed to free public libraries and free use of information do not generally associate a dollar cost with information. Such an attitude is no longer realistic. And once significant numbers of people are paying for information, it is very likely that they will become more discriminating and critical, both of the data bases and the services that are provided. A more educated

and sophisticated user clientele will aid in the winnowing out of less desirable or less qualified sources and services.

Another area of concern to centers and suppliers is that of standards. All operating centers have been made painfully aware of the need for uniform data base standards: standards for the type, designation and completeness of data elements contained on the tapes; standards for formats; and standards for machine code representation, etc. Currently, the lack of such regulation is apparent within individual data bases, between multiple data bases prepared by the same supplier, and between data bases prepared by different suppliers. The present lack of uniform standards imposes on centers the added cost and burden of: preprocessing tapes; altering search strategies; training operating staff in the conventions used by different sources; and familiarizing users and potential clients with the varying contents, benefits and limitations of various sources. Beside seriously affecting the use of individual data bases, these factors present obstacles to the creation and use of merged data bases.

Compounding the problems caused by the paucity of standards is the use of confusing conventions within data base services. For example, one service includes patent assignee company names as index terms rather than identifying them with data tags of their own. This causes problems for the users who wish to use search terms that happen to coincide with terms that occur in company names. For example, the term CO for carbon monoxide

would retrieve every occurrence of the term CO which is the abbreviation for company. And, the term oil would retrieve every occurrence of that term in a company name such as Standard Oil, Pure Oil, etc.

Another instance of a problematic convention used by an individual data base supplier is the use of imbedded blanks within single word terms. In this case blanks are inserted into words to allow sorting on certain word fragments. Since there is no intrinsic distinction between these imbedded blanks and word-separating blanks the search strategy for this data base must be adapted to allow for the possible presence of blanks in search terms. Needless to say, this causes considerable added effort on the part of centers when searching these tapes.

Because of the impreciseness and lack of structure of the English language, we will always have to live with certain ambiguities and problems associated with vocabulary. However, many of the data base problems encountered by centers are associated with lack of standards, lack of uniformity, use of individualistic conventions, lack of vocabulary control, variations in nomenclature, and other factors. Some of these problems are inevitable when one is searching titles because titles are generated by authors and involve no vocabulary control. But some of the problems are within the control of data base suppliers.

Centers are not unaware of the reasons why suppliers have not yet implemented all the standards we want, nor worked

together to provide compatible data bases. Their production problems are numerous, their motivations are many and varied, and, they face the problems of making their new data bases compatible with their older files. Methods of achieving internal compatibility with their own historical files and external compatibility with others files are not immediately obvious. Additional processing must take place.

Recommendations for standards have been drafted by the ASIDIC Standards Committee with representation by both centers and suppliers and they are communicating with other standards committees. On the other hand, in cases where standards have not been implemented and perhaps will not be implemented, many of us face the problem of preprocessing tapes before they can be searched. Currently, this preprocessing takes place at many individual centers and the cost is duplicated several times over. Alternatives to the preprocessing by centers are that it be done either by the supplier or by an intermediate preprocessing center. The preprocessing problem was discussed at the first meeting of the Cooperative Data Management Committee and we decided to look into the costs that would be associated with an intermediate preprocessing center. Dr. J. L. Carmon developed a cost estimate for a preprocessing facility and it was obvious from the figures that a separate facility could not be supported for this purpose. The Data Management Committee will continue to study the problem.

In a recent survey Ken Carroll of the American Institute of Physics identified some 50 machine readable data bases. Perhaps

5 - 10 of these can be considered major popular data bases and the balance are in lesser demand because of their limited size, coverage, their high processing cost, or because of the limited market of highly specialized users to which they appeal.

From among these data bases, each center must choose the data bases it will use for the Selective Dissemination of Information (SDI) and/or retrospective searches it will offer. And the data bases chosen will probably be those that are most attractive in term of marketability.

Data bases such as CAS Condensates, BA Previews. COMPENDEX, INSPEC, ASCA and others may be in wide enough demand to warrant their use by many centers. However, there is a saturation point.

While relatively few centers now provide computerized searching services, there is every indication that the number will increase significantly within the next 5 years, yet the number of popularly demanded data bases is not likely to change drastically. Thus, the growth in the number of processing centers, and the motivation for selection of data bases, may lead to problems such as the demise of some worthwhile data bases of limited popularity, and the growth of too many centers with identical data base coverage. Theoretically, when and if this point is reached, quality will tell and the fittest will survive. But there is also the very real possibility that a limited group of users could be distributed rather evenly across the entirety of centers, so that none of the centers could be sure of a large enough portion of the market to insure economic survival.

As more and more data sources become computerized, information centers will face the responsibility of being in a position to contribute significantly towards the success or failure of supplier data bases simply by either using or not using the data bases. There is little risk of information centers endangering the secure position of the most popular data bases, but there are also some very valuable, high quality, specialized data bases that might not appear to be as readily marketable as others. Their apparent marketing limitations may be due to any one or combination of the following factors: limited utility; high lease price; high processing cost; the sporadic nature of search questions; limited audience; non-availability of search programs for specific hardware configurations; need for highly specialized personnel for processing requests, etc. Data bases such as the Chemical Abstracts Service Registry and Substructure Search System, some of the analytical data collections, and data tapes resulting as a by-product from some publishing efforts might fall into this category. It would be economically unfeasible for very many centers to make the necessary investment in manpower, training, machine time, and associated costs to process and provide search services from such data bases. Yet, as members of the information community, centers have a responsibility to the users to bring into use those collections that are truly useful.

This problem which deals with the distribution of data bases among centers has been addressed by ASIDIC. In order to find out

what data bases are being used or are planned for use by the various centers, Marilyn Brown has sent questionnaires to all institutions on the ASIDIC mailing list. To date 25 centers have responded and the initial indications bear out what one might suspect, i.e., a small number of data bases -- the major ones--are used by a significant number of centers, and a large number of data bases are used by a very small number of centers. Specifically, 31 data bases are currently being offered by 25 centers. Five data bases of the 31 are used by 5 or more centers and the remaining 26 data bases are used by 3 or fewer centers--19 of the 26 are used by only one center. These data are only partial data and we expect to get responses from additional centers; however, I suspect these data are representative.

One way of ensuring the availability of the data bases that have limited appeal, and, of distributing the cost of processing, searching and maintaining them, would be through a cooperative network of centers. This would reduce the financial burden of individual centers and might also allow each center more opportunity to develop greater specialization in terms of providing specialized coverage, specialized output or products, and perhaps the creation of new subset data bases comprised of selected portions from many data bases.

A network of centers would also make retrospective searching more economical by reducing the need for duplicative efforts.

The creation, maintenance and searching of large retrospective files of individual or merged data bases will be expensive, and unnecessary duplication of such efforts in several locations would be wasteful. Within a cooperative network of centers it should be possible to share resources and efforts. But, this naturally raises other problems--both legal and financial.

The current lease agreements of data base suppliers are very specific with respect to: how, where and by whom a data base can be used; how the output can be disseminated and in what form; copying of data bases; royalty and use charges, etc. The restrictions are imposed in order to ensure the economic viability of the supplier. Suppliers want to be able to recover their costs and be adequately compensated for use of their material and centers on the other hand want to use the data bases in a manner that will permit them to provide the services and products their users want.

It becomes apparent that the sharing of resources through a network, the repackaging of data, creation of merged data bases, creation of retrospective files, creation of personal files from data bases and distribution of data on nonstandard media--for example, computer output on microform--are all considerations that are being entertained by centers and they are also considerations that could seriously affect the suppliers. Centers would like the freedom to create new and innovative products and services using data bases as their raw material and they want to do these things in a manner that will supply the equitable remuneration

that suppliers justly deserve. We have no desire, for example, to provide a retrospective search service that would decrease a suppliers sales of back issues of journals without providing adequate compensation.

The solution to these and many other problems are under consideration by the Cooperative Data Management Committee. The Committee met yesterday and a specific recommendation coming from that meeting was that a mechanism be established to ensure the center-supplier interface that is necessary to air their mutual problems. Following the committee meeting, we contacted the major suppliers who are represented here at this NFSAIS-ASIDIC meeting and found that they were in agreement with the recommendation. Once all the findings of the center survey are in, arrangements for informal workshops will be arranged via ASIDIC for suppliers and centers. In the meantime we have arranged informal meetings following this session.