# The Use Of Computers In Literature Searching And Related Reference Activities In Libraries

Edited by
F. WILFRID LANCASTER

University of Illinois
Graduate School of Library Science
Urbana-Champaign, Illinois

LC Card Number: 76-1790
ISBN: 0-87845-043-2
U.S. ISSN: 0069-4789

TABLE OF CONTENTS

# INTRODUCTION

The Twelfth Annual Clinic on Library Applications of Data Processing was held at the Ramada Inn, Champaign, Illinois, April 27-30, 1975. The subject of this particular clinic was the application of computers to information retrieval and other aspects of reference service in libraries. An attempt was made, through a combination of formal papers and demonstrations, to cover a wide range of applications of machine-readable data bases in support of the reference activities of libraries. The papers contained in this volume represent a number of different viewpoints and different types of user. Public, academic and industrial libraries are all represented. Some of the papers are presented by consumers, while others are given by retailers of information service.

The paper by Dowlin represents an in-house application of a minicomputer in an information retrieval application, while the paper by Waltz goes beyond citation retrieval and investigates the possibilities of machine systems that might actually answer factual questions posed by library users. Lyon discusses the use of on-line systems in computer-aided instruction in libraries, and the contribution of Schmidmaier reviews the use of machine-readable data bases in Australia. The opening paper and the closing paper are general reviews, the former dealing more with our achievements and the latter with the failures or limitations in the provision of information services.

F. W. LANCASTER
*Editor*

MARTHA E. WILLIAMS
Director
Information Retrieval Research Laboratory
University of Illinois
Urbana-Champaign

# Machine-Readable Data Bases In Libraries: Criteria For Selection And Use

Traditionally, libraries have been the source of stored information—the collective memory of a community or a civilization. In recent years, with the proliferation of publications and governmental involvement in research, machine-readable data bases have evolved as separate entities which store in indexed and abstracted form much of the current information found in libraries. As the quantity of recorded information increases, libraries are, and will be increasingly, forced to rely on these machine-readable data bases to search the accumulated knowledge if they are to retrieve it efficiently—or at all. Some libraries are now providing data base search services to their patrons; many more are considering it.

## What is a Data Base?

Data bases are organized collections of information in machine-readable form and exist in almost all of the major fields of science and technology as well as in the social sciences. The collected information may be of several types: bibliographic or bibliographic related, natural-language text, numerical, or representational. An example of a bibliographic data base is the MARC II data base of the Library of Congress, or the Chemical Abstracts Service's (CAS) CA CONDENSATES tapes. CASIA (Chemical Abstracts Subject Index

Alerts) tapes, which contain subject index terms and postings that consist of Chemical Abstracts citation numbers, is an example of a bibliographic-related data base because the citation number refers the user to other tapes or hard-copy sources that contain the full bibliographic citation. A natural-language text data base would be the text portion of the New York Times Information Bank, which contains not the full text of newspaper articles, but textual summaries or abstracts of the articles. System 50 for State Statutes of Aspen Systems Corporation, an example of a full-text data base, contains more than 200 million words of statute law. A familiar example of a numeric data base is the current U.S. census tapes produced by the Bureau of the Census. A data base that contains not alphameric data but graphic or pictorial representations, such as the CAS Registry Structure data base which contains chemical structures, is referred to as a representational data base.

## Who Produces Data Bases?

Data bases are produced, or generated, both by governmental sources and by organizations in the private sector. Included in the private sector are profit-making as well as not-for-profit organizations such as professional societies. Although the government is responsible for the generation of numerous data bases, in many cases the actual production work is carried out under contract by either not-for-profit or commercial organizations.

Many of the largest and most heavily used data bases were produced by the federal government, including: the MEDLARS (Medical Literature Analysis and Retrieval System) tapes produced by the National Library of Medicine; the MARC II (Machine-Readable Cataloging) tapes produced by the Library of Congress; the ERIC (Educational Resources Information Center) tapes of the National Institute of Education; the DDC (Defense Documentation Center) tapes of the Department of Defense's Defense Documentation Center; GRA (Government Research Announcements) tapes of the National Technical Information Service (NTIS); and STAR (Scientific and Technical Aerospace Reports) tapes produced by the National Aeronautics and Space Administration. The fact that government-generated data bases are heavily used is a function not only of their usefulness but also of the fact that their production and use are subsidized by the government.

Of the large scientific, technical, and discipline-oriented data bases, many have been produced by professional and technical societies in the not-for-profit part of the private sector. Some of these are: the SPIN (Searchable Physics Information Notices) tapes of the American Institute of Physics; *BA Previews* (Biological Abstracts Previews) of BioSciences Information Service; CA CONDENSATES of Chemical Abstracts Service; PATELL (Psychological Abstracts Tape Edition-Leased or Licensing) of the American

Psychological Association; COMPENDEX (Computerized Engineering Index) of Engineering Index, Inc.; and METADEX (Metals Abstracts Index) of the American Society for Metals. These data bases are produced within the private sector; however, many of them have received research and development funds from the government to help them get started or conduct research associated with systems or products.

The number of profit-making organizations producing data bases is small, but some of these data bases are very important; for example, the Institute for Scientific Information publishes the Science Citation Index (SCI) tapes and the Social Science Citation Index (SSCI) tapes; Excerpta Medica is produced by the Excerpta Medica Foundation; the F & S Index of Corporations and Industries is produced by Predicasts, Inc.; and the New York Times Information Bank is produced by the *New York Times*.

A few data bases were generated specifically for the purpose of information retrieval, but because the cost of data input is high and could seldom be justified for the purposes of retrieval alone, many more were created as by-products of other activities. Some were created because machine-readable data were needed as a component of a computerized-process control or production system for publishing primary journals, indexes or abstracting journals. Others were created as a result of the fact that computerized typesetting was used to produce a hard-copy publication. Computers have proven to be economic and effective tools for producing primary and secondary publications. Consequently, every time a publisher uses computerized photocomposition, a potentially machine-searchable file exists. The machine-readable file, once created, can be automatically reorganized, merged with other machine-readable files, reformatted, and repackaged to meet the demands of various markets. It has become obvious that machine-readable files are considerably more flexible and can serve many more functions than can hard-copy records.

## What Kinds of Data Base Services Exist?

Data base services differ in types of service offered and can be classified as either batch mode or on-line, depending on the method chosen to process the information. The different methods of processing are related to the types of service and determine the type of file structure that is best suited to the particular purpose. The basic types of file structure for information retrieval purposes are: (1) inverted, e.g., the alphabetical grouping of terms with postings; and (2) serial, or sequential, in which each record or citation is examined in turn.

An on-line system is one in which the user is in direct communication through a terminal with the central processing unit of the computer. An

on-line interactive system is one in which there is literally an interactive two-way communication between the user and the machine, and response by the machine is immediate. On-line searches of bibliographic data bases are usually run against inverted dictionary-type files. On the other hand, a batch processing system is one in which multiple jobs or search questions are "batched" together and run at one time. The search questions can be entered via a terminal, cards or tape; however entered, they are saved until the time of the batch run. Searches against a serially or sequentially arranged file are usually run in the batch mode so that the basic cost of spinning the tape once can be spread over several search questions rather than requiring one question to bear the total cost. There is, of course, some incremental cost for processing the additional questions.

Retrospective and current awareness searches differ with respect to the currentness of the files against which they are processed, and with respect to the number of times the question is run against the files. A retrospective search, or question, is one which is run against older, historical or past files, whereas a current-awareness search is run against only the most recent file. A retrospective question is usually run once against the entire collection of data base issues or volumes, while a current-awareness profile is run many times—each time against the most current issue of the data base. Computerized current-awareness systems are usually called SDI (selective dissemination of information) systems. Information is searched for and retrieved from the file in accordance with a profile of the user's search interests. The output or search results are then disseminated to the user. In the case of SDI, once a profile of the user's interests has been developed and refined, it is run on a regular basis against each new issue of the data bases requested by the user. SDI searches are usually run in the batch mode against sequential files. After an SDI run has been completed on the most current issue, the tape of that issue is added to the retrospective file for its data base. Several of the on-line services now offer SDI in addition to retro-searching. Since they have to process incoming new data base issues as they arrive anyway—in order to add them to the retrospective files—they can conduct the SDI searches at the time of that initial processing. In these cases, search output can either be disseminated to the user through the mail or stored for later retrieval through his terminal. Retrospective questions can be run in either the batch or on-line modes depending on the system on which the search is processed. In most cases the file that is searched is in inverted form for fast searching.

SDI and retrospective searches differ in purpose. The purpose of a retro-search may be to provide the user with: (1) a few relevant references to become acquainted with a topic; (2) a thorough coverage of the literature on a particular subject; or (3) one or more references that contain the answer to a specific question. These searches are conducted on demand and always in "past" or retrospective files. The completeness of the search question

processed against the file varies considerably with the user's purpose. In contrast, SDI searches are conducted in order to keep the user up to date with the published literature in his field. The user profile is usually designed to be as complete as possible and to achieve high recall. The same profile is used over and over against new issues of the data base. The profile is modified over the course of a year if changes in user interests or data base output indicate the need. Since SDI and retrospective searches of data bases differ in purpose, comparisons of the two with respect to performance and cost make little sense.

There is another type of service which libraries should be aware of—"private library" service. In this service, the user can have output from any machine search stored for him on a separate disc file along with his own judgments about citations he has received. This feature is now available from several organizations that process data bases. At their own discretion, users may discard unwanted references, add new material, or even augment the file with additional indexing terms for the references already selected. This type of service can be provided on a personal basis or on a company basis. It would be possible, in this way, for a library or a company to generate its own machine-readable files without having to develop its own data base or search strategy.

## What Effect Does Data Base Service Have on Library Operations?

Data base searching can have a direct impact on libraries in several ways: (1) it can affect the acquisition policy of the library—either increasing or decreasing acquisitions by pointing out the nonuse of some journals and/or the need for other journals; (2) it can affect the interlibrary loan traffic of the library as either a borrowing organization or as a lending organization, depending on the correspondence between the library's serials and monograph collections and the retrieved citations from data base searches; (3) the library can expand or deepen its services by offering personalized data base search services, for both individuals and organizations, from data bases it processes; (4) the library can function as an intermediary, preparing search questions and processing them via an on-line service, or through another center; or (5) the library can function as a referral center, directing its customers to the appropriate data bases and service centers.

If a library is considering providing data base services to its patrons, it must understand before making a selection what types of services are available, and it must know how to evaluate both data bases and processing centers.

## How to Evaluate a Data Base

The potential user of data base services will have to evaluate not only the searching methods available but the content of the data base itself. The

subject coverage of data bases may be discipline oriented, mission oriented, problem oriented or multidisciplinary. In evaluating them, a library must first know how their coverage matches the objectives and the breadth of its own collection. Does the data base cover material such as government reports, journal articles, patents, monographs, theses, reprints and news items? If it does, how complete is this coverage? That is, if it claims to cover a particular journal, will it be covered in its entirety or only for selected issues or articles? This information can be hard to find, although many data base producers provide lists of the journals and other items indexed.

Another important consideration is the time lapse between the item's appearance in the primary source, in the secondary source (or index), and finally in the data base. (In some cases a citation appears on a tape before it is produced in a hard-copy secondary source because the hard-copy publication is produced from the tape.)

In addition, one should question the indexing and coding practices. Does the data base include free-language keywords on the tape? Does it include a controlled thesaurus or hierarchical vocabulary terms? Are titles given exactly as the author provided them, or are they augmented titles as in the case of BIOSIS in which additional terminology is added to the author's title? Does it include other kinds of codes to indicate subject matter or any other criteria about the item itself? Are abstracts and extracts available on tape for search and display, or will the library have to go back to the hard copy to obtain them?

The size and growth rate of the data base will indicate something about the number of citations available from one year's accumulation of that file. It is important to know how the tape version corresponds with the hard-copy version. In some cases there is a one-to-one correspondence; that is, for each abstract or reference contained on the hard copy there is a tape representation. In many cases, the data base itself is a subset of the hard-copy version; or the reverse may be true, i.e., the data base may contain more citations than the hard copy. In other cases, such as the MARC tapes from the Library of Congress, there is no corresponding hard-copy publication (except the collection of LC cards). If there is a corresponding hard copy, a library can occasionally do both computer and manual searches of an issue as a cross-check to be sure it is using the right terminology and really getting what is wanted.

## Important Concerns for Library Internal Data Base Processing

As a processor, the library must investigate the consistency and quality control exercised by the data base supplier, and must also be aware of the frequency with which changes are made in the data base and in the provisions for notification to the processor of data base changes. If the supplier indicates forenames of authors by first initials and later decides to use full first names

instead, this of course affects processing. Addition of new data elements affects processing time as well, and possibly requires a change in the search program. Adherence to a delivery schedule is another concern of the processor. If the data base supplier sends his tapes late, then the library will be delayed in providing output to its clients.

There is another consideration in looking at data bases: If the library plans to use more than one data base, is there overlap of subject coverage between them? There are costs associated with intellectual processing (indexing and abstracting) and manual inputting of citations. If the same citation is handled more than once, this can represent wasted time, effort and money. The processor also wastes money by having to search for the same material on more than one set of tapes. There are a few processing centers that merge several data bases to create one common data base. This is being done at Ohio State University and for the pollution data base (Pollution Information Program–PIP) at the National Science Library in Canada, but in general most centers search each data base as an individual entity.

A final–and much more technical–processing concern is that of compatibility between various data bases. The variability among the data bases complicates handling for those who process them. The standard arrangement of data element tags, data content, and directory information for the records is referred to as the format of the record, and the arrangement of the records on a tape or other media is referred to as the file structure or file format.

Unfortunately, file formats and record formats are not standardized, nor are the definitions, contents and representations of the data elements. There are almost as many data base formats as there are data bases, which leads to confusion and, of course, added expense in processing tapes, because it requires the processor of multiple tapes to either develop multiple search programs or to reformat all incoming tapes to one standard format. One important standard has been developed by the American National Standards Institute for interchange or transmittal of bibliographic records: the "American National Standard for Bibliographic Information Interchange on Magnetic Tape." The MARC implementation of this standard has been proposed as a Federal Information Processing Standard and, barring problems, it will go into effect as a federal standard. This standard deals only with the format for records on tape or the generalized structure, not with the contents of the records. It does not define data elements or tags, specify required data elements, or specify data representation beyond that of the required character set.

## Important Concerns for Library External Data Base Processing: How to Evaluate Processing Centers

The processing of data bases is an expensive activity and most libraries will be interested in buying these services from an information center. There

are a number of questions the library will need to ask itself, and ask each center under consideration, when evaluating the relative merits of the many centers available.

Some of these questions are: Does a center have the data base or mix of data bases that will satisfy the needs of the library's clientele? When the appropriate mix of data bases has been found, does the processing center retain all the records from each of the data bases the library is using, or does it strip off certain parts of some of the tapes? It is essential to know whether or not all of the information in a data base is being searched. Does the center employ a standard internal format? Will all data bases be processed the same way? Will all of the output received by the library look (be formatted) the same?

Does the center provide any kind of document backup? Most centers do not because of the cost associated with resource location and acquisition. The Institute for Scientific Information (ISI), however, provides document backup for anything that is in their *Current Contents* through Lockheed's DIALOG and System Development Corporation's ORBIT on-line systems. Anyone searching ISI tapes on these systems can enter a request for a document. The requests are saved by the systems and transmitted back to Philadelphia every night. Documents are mailed out the following day. A similar system is available for the National Technical Information Service (NTIS) tapes. The Ohio State University's Mechanized Information Center (MIC) system also provides document delivery from its own collections.

There are other services which some centers provide: off- or on-site training for library personnel; manuals to assist users in writing profiles and search strategies; free demonstration searches to give an idea of system capability; dictionaries, vocabulary lists or thesauri for controlled vocabulary; term frequency lists for title terms; and free-language keywords. It is important to ascertain whether the revision of search profiles is permitted and whether this imposes an added cost. Some centers supply a newsletter to keep users informed about changes in indexing practices or the addition of new data elements, so that they can modify their search strategies to accommodate such changes. In some instances provision is made for feedback or data base monitoring, to aid in calculating the precision rating for searches. All these things relate to the general cooperativeness of the center staff and their accessibility to the patron.

The data elements provided in the output are of special interest to the client or patron. Which elements are included in the actual printout or display? Are just the title and the author name shown, or also the keywords? Are the terms in the search which caused this particular item to be a "hit" shown? How are the data elements arranged on the output medium? How many citations are there per page? If large computer paper is used, are the

citations printed in two adjacent columns so that the paper can be cut into two file-size portions? Are there options available for sorting the output? For example, can one specify that the output be sorted alphabetically by author's name, numerically by reference number, in descending order according to ranked weight or value, or by date of publication? Will the output from different data bases be displayed in a standard format for easy visual scanning? Will both upper and lower case characters be used? All of these features may be of considerable importance to a library and to its clientele.

## External Processing: Search Output Options

Some centers display retrieved information on cathode ray tubes (CRTs), in which case the user is likely to require that paper copy, or hard copy, also be printed off-line and sent to him. On request, some centers can generate microform output directly from the tape. The output can also be supplied on magnetic tape itself for later in-house use. Most suppliers require, however, that the output be provided on hard copy, both to avoid copyright difficulties and to provide records of the citations retrieved for the purposes of reimbursing the supplier with appropriate royalties.

Assuming that the output is provided on hard copy, there are still many possibilities from which to choose. For instance, some centers can produce output on multilith masters for further reproduction. (This feature could be especially useful to libraries, for example, in the production of SDI bibliographies or bulletins.) The output can be on cards or on computer paper of various sizes. If it is on cards, it will be easily separable into unit records; false hits can then be discarded and pertinent citations interfiled with other material in a constantly updated card file. If output is on computer printout paper, the whole group of citations may have to be retained or the desired references cut out of the pages.

## External Processing: Search Features to Look For

When evaluating the search capability of a particular center, it is important to ask which information items or data elements are routinely searched. Data elements are the basic building blocks of data bases. In the case of bibliographic data bases, some of the commonly searched data elements are: author, title, journal name, volume number, issue number, date of publication, index term, keyword, and publisher name. The data element is the smallest unit or element within an information or data field which contains one or more data elements. Usually multiple data elements and fields make up a bibliographic record. A record is the representation on magnetic tape of the physical book or article, etc., in much the same way that a card in

the library card catalog represents the book on the shelf. Records, in turn, comprise a file. Ordinarily one file makes up a data base, but sometimes a data base contains several files.

A searcher should be permitted to access both individual records within a file, and individual data elements (fields) within individual records. Thus, in formulating a search question or request, the searcher should be able to specify certain search terms or data elements; the computer should search the index portion of each record in the data base to locate search term or data element matches, and then produce a printout or CRT display of the records that contain those matches. On the other hand, if the searcher knows the citation or reference number of certain desired records, he should be able to specify these directly and have the matching records printed or displayed immediately.

It is possible to search specified data elements within an individual record either because the elements are identified by unique codes, or because the position of an element within a record may specify what type of element it is. Often a directory is associated with each record which specifies the elements that are present, their location in the record, and the length (number of alphameric characters) of the data content.

There is a difference between the data elements used in searching and those displayed in output. Searchable elements are often a subset of those displayed. Abstracts, for instance, are seldom searched but often displayed. Abstracts provide sufficient contextual information to the user to aid him in determining whether he has a "hit," i.e., whether or not the terms in his search have succeeded in locating an article he needs. In some cases abstracts themselves may be searched, but this is seldom done because it significantly increases search time and cost, usually with little added benefit. The number of access points, or searchable data elements, greatly influences system ability to achieve high recall and precision in searching. Some centers only permit searches on subject terminology, i.e., words found in the title, keywords, or index terms. Others permit searches on author, company affiliation, LC class number or Dewey Decimal number, report numbers of various types, languages, countries of origin, or other types of data elements. (Obviously, some data elements are specific to certain data bases, e.g., Engineering Index card-a-lert codes are found only in COMPENDEX.) It is important to ask whether, as in the case of MEDLINE, hierarchical terms can be used in searching, and, whether there is a way to distinguish among the data elements—e.g., can an author word be distinguished from a subject term? This last distinction is important in order to avoid retrieving false hits due to homographs.

What kind of logic is permitted in search strategies? Is full Boolean logic permitted (using *and, or*, and *not* operators), or are there restrictions? Some centers provide adjacency logic, i.e., they permit specification of the context in which a term occurs. For example, the searcher can indicate that a term must

occur within one or two words of another term as opposed to being found anywhere in the record. This feature is available in several systems.

Another feature to look for is the availability of truncation, which is the ability to search on a fraction of a term. For example, a user interested in the concept *analysis* can include in his search question the term fraction *analy** (truncating after the y), and thereby retrieve all occurrences of the terms analysis, analytical, analytics, etc. Without this feature, it would be necessary to specify all the variant forms of a word that an author may have used. Title terms, in most data bases, are not generated from a controlled vocabulary but contain natural-language, or freely generated, terms. The searcher must then be able to adjust to the term variability provided. Most centers provide right truncation only; however, left truncation can be extremely useful. For example, if a person was interested in antibiotics and searched under the term fraction **mycin*, he would hit approximately forty different variations of that term, which is probably more than he would have been able to think of easily. The limit to the number of characters that can be truncated must be established, and it must be decided whether truncation is available for term types other than subject words. Another important question about searching is whether or not the system can provide ranges for numeric data. Could one search for items published between 1972 and 1975 only, for example? And finally, in the case of on-line systems, is there the capability for one to review search strategy, or to save search strategies for later use? Can the system answer inquiries about system operation, e.g., explain commands and responses, and are the explanations available at several levels of sophistication? New features are constantly being added to on-line systems; one must keep up to date with them in order to make effective use of the tools provided.

## External Processing: Comparing and Evaluating Processing Center Costs

Cost features, although often considered first, should be secondary considerations after the selection of the appropriate data bases and centers that provide suitable searching and output features. Generally, costs are competitive and do not vary much.

Charges can be established in several ways. Does a center charge an annual fee? Does it charge for profile writing? Is there a charge for the number of terms used in a search, or is some maximum number of terms allowed and then an assessment made when this number is exceeded? Does the center charge on the basis of the number of hits received in the searching, or is a maximum established and a charge levied only after the maximum is reached? Is there an additional charge for postage, user aids, or for the media on which the output is received? If an on-line system is being used, are