Automated Access to Archival Information: Assessing Systems

DAVID BEARMAN

MANY ARCHIVISTS WILL SOON BE CONFRONTED with deciding whether to introduce automated information retrieval systems into their institutions and, if so, with evaluating both their needs and the capabilities of a variety of systems. Whether the pressure for adopting automation arises from the central administration of the organization which the archives serves, from an equal branch, such as the library or records management staff, or from within the archives itself, most archivists will initially resist. They will be concerned that their unfamiliarity with automation will lead to costly and embarrassing mistakes, and they will feel that automation to meet the need of one constituency will detract from the overall effectiveness of an archival program whose scope is not fully appreciated by their clients. Skepticism about automation is healthy. Automated systems are not desirable just because they have loyal lobbyists inside or beyond the institution. However, blind opposition to automation will probably be counterproductive regardless of who the constituency calling for adoption is. Thus, learning to assess institutional needs, especially needs which might be met by automated information retrieval systems, is rapidly becoming an essential archival skill.

This essay is intended as a guide only. The identification of needs specific to each archives, and the evaluation of automated systems to meet these needs, can be performed only by the staff of each archives. This paper poses questions, but rarely provides solutions. Indeed, the process of assessing institutional needs and automated systems capabilities is necessarily continuous; any decisions reached now will have to be periodically reconsidered as conditions change in both the archives and the systems options.

This discussion is based upon the author's experience with an archival survey project, the Survey of Sources for the History of Biochemistry and Molecular Biology, which began in 1975 and will be concluded in the summer of 1979.[1] Our

[1] For a detailed description of the project and its aims, see John T. Edsall and David Bearman, "Survey of Sources for the History of Biochemistry and Molecular Biology," Federation of Allied Societies for Experimental Biology, *Federation Proceedings* 36 (1977): 2069–73.

experience illustrates the steps involved in assessing needs and systems capabilities as well as some of the pitfalls to avoid in a choice which can either limit or open so many options.

Would Automation Serve Our Needs?

Finding aids are the backbone of archivists' current information retrieval system. Typically, manual aids are separately compiled to serve a single purpose: to provide intellectual control to guide users to appropriate materials, or to maintain administrative control of various sorts, or to report on holdings in a variety of annual reports and repository guides which have both intellectual and administrative uses. An automated information retrieval system is composed of a *data base* containing all the information used by the archives in its administrative and service functions; and *software,* or instructions to the computer to retrieve, sort, and print some or all of the data. Such a system can make possible numerous applications with little additional effort on the part of the staff. These capabilities make the idea of automation extremely attractive. A single data base can provide annual accessions lists; cumulative guides; reports for tax or insurance purposes; information on the location and physical condition of records promised to the archives, along with the rate at which they are growing and reminders of scheduled transfer dates; and answers to specific reference questions posed by users seeking to locate specific materials whether by type of material, by names, or by subjects. However, applications such as these must be considered at the time the data base is designed and the software is written. Unlike manual methods, where we are always free to change the format and even the information to be reported as we go along, automated systems can be modified to accommodate new features only at great cost.

Nonetheless, the capabilities of automated systems—including the ability of some systems to translate photocomposition instructions in order to produce printed guides, to serve management purposes ranging from writing of paychecks or staff directories to computing the monthly budget, or to be used as automatic memory typewriters in order to write "personalized" form letters to users or patrons—make automation seem attractive. If a single system were currently available to perform even the limited set of tasks suggested here, and was adapted to an archival application, this discussion would be unnecessary. However, while automated information retrieval and data processing systems can do these and many other things for us, no single system currently available does all of them well enough to win unanimous acceptance from archivists.

Therefore, determining whether particular automated retrieval systems have capabilities needed by a particular archives requires a careful assessment of the information processing costs and capabilities of the current, manual system of the archives. This assessment must consider what functions now served by a variety of manual aids could be served by an automated system, and what further functions not now met at all such a system could meet. This essay does not attempt to detail institutional cost assessment techniques, since a vast literature on management and management of information services can guide the archivist in determining the costs in staff time and material involved in filling information re-

quests of clients and staff.[2] This essay is limited to posing questions which will assist in evaluating needs, assuming that costs are already known. We can begin by examining some of the current functions of the archives:

(1) What staff activity is involved in recording data in the processing of new collections? How are these data made available for each of the administrative and intellectual control purposes for which it will eventually be used? How much time is involved in collecting, editing, and committing to paper each of the aids, lists, guides, and reports currently prepared by the archives?
(2) What duplicate data, such as collection name, call number, size, dates, and the like, are collected for each of the applications above? What methods are used to verify data such as personal names, institutional subdivisions, or publications titles appearing in more than one collection?
(3) Who uses each of the aids currently produced and what further processing do users do to get the data into a form useful to their purposes (i.e., how does the administrative vice-president summarize the annual report for the trustees, for the insurance agents, for donors, etc.)? How are specific reference queries handled with the current system? Are portions of aids photocopied or retyped in letters to users? What level of staff expertise is required to produce each answer to each kind of user?
(4) How are finding aids currently disseminated? Does the form in which they are disseminated result in user requests identifying exactly what information is needed? How much work is involved in translating user requests from your aids?

What Should We Know to Assess Automated Systems?

Before examining specific retrieval systems many archivists may be repelled by two powerful misconceptions. It is appropriate to stop here to dispel these imaginary barriers.

Frequently archivists assume that they must understand computers and possess programming skills in order to implement an automated system. This is unequivocally false. Archivists need to know the needs of their institution only; these needs can be translated into systems terms without attention to details of programming. Just as the commuter need not understand how the dispatcher directs buses and subways in an urban tranportation network in order to use the system, the archivist need never be a programmer. Both the archivist and the commuter, however, need to understand the system design and where they wish to go, or they will be hopelessly lost.

The second stumbling block is equally chimerical. We are often told that computer hardware is becoming obsolete faster than even the experts can keep up with. Will this leave us with a tool which is inefficient and cannot be repaired? Absolutely not. First, it is extremely unlikely that many archives will purchase computers, even mini-computers. We need not purchase the input devices either, any more than we are forced to buy typewriters or other office equipment. All these items may be leased and the same cost calculations which are involved in other office equipment should be the guide to whether to buy or rent. But more

[2] The annual bibliography in *The American Archivist* can here be usefully supplemented with articles on management and cost-benefit analysis from *Library and Information Sciences Abstracts* (LISA), and *Information Science Abstracts* or the *Bulletin Signaletique.*

important, hardware is only ancillary to the information retrieval systems we are considering. The systems consist of software that will not be made immediately obsolete by the new technology.

With the imaginary obstacles behind us, our attention can be focused on the real hurdles in our path. We must establish broad criteria for automated systems in order to identify those systems that might reasonably be expected to meet our needs. To get to these broad criteria we must evaluate the context, functions, and goals of the archives and consider the constraints under which it operates. For purposes of illustration, I will refer here and later to the specific needs identified by the Survey of Sources for the History of Biochemistry and Molecular Biology. Outlining our scope and purposes as a prelude to considering automation, we arrived at the following self-description:

CONTEXT	Implications
Three year project	Reporting and reference work should interfere as little as possible with data gathering. The staff not involved in the data collection or indexing must be able to continue updates and reference work.
International in scope	The project relies to a great extent on collection descriptions written by others. Descriptions will vary widely in depth and quality and in subject terms employed.
FUNCTIONS	
Clearinghouse function	Information should be preserved in terms as close to the original as possible while being accessible to search on as many terms and variables as possible.
Scholarly reference service	Intellectual, social, and other historians as well as social scientists and philosophers will approach materials with different kinds of questions.
Advising donors and private owners	Access to information about undeposited papers and owners' intentions as well as about potential repositories and their interests is critical.
Service to archivists	The system should inform an archivist what materials relating to his collections are located in other repositories or in private hands.
GOALS	
Intellectual access to collections	The data base should be searchable with *Boolean* capabilities.

Administrative control	The system must generate guides, link holdings of various collections, produce authority files of names and institutions indexed and thesaurus of subject terms, and tabulate statistics.
Demonstrative project	The system should accommodate data conforming to national and international standards. The output should be in forms comfortable to archivists, and user interaction should be as simple as possible.
CONSTRAINTS	
Access to hardware	Input devices and computer will not be purchased. The system should use computer facilities available through the university computing center.
Access to consultants	After system is set up, it must be totally maintained by archival staff.
Costs	With fixed budget for the overall project, any systems costs must produce equivalent savings in time and staff.

This definition, obviously, is specific to one institution. However, the assessment of context, functions, goals, and contraints can produce a list of needs or implications, which will be essential in appraising any potential information retrieval system. The crucial characteristics of our needs proved to be the necessity of accommodating information derived from descriptions containing widely varied terminology and differing in their scope and depth. Another significant criterion proved to be the need to have access to all of the data elements in each finding aid (dates, subjects, names of persons and institutions, quantities of material, and the like) in order to provide access to users from a variety of disciplines who would want to search the data base in different ways. Finally, our desire to report to archivists in a familiar form imposed additional criteria which became important in the selection of a system. Other institutions will find that different aspects of their special needs have limiting implications which will help them in similar fashion to rule out certain systems for their purposes. As more institutions adopt automated retrieval systems, one increasingly important criterion limiting the selection of a system is the desire of archivists to cooperate with each other, to exchange information with other systems, and to report holdings to state and national agencies.

How Do Systems Differ?

Information retrieval by computer necessitates the prior input of data, in a machine-readable form, to the computer. The machine-readable magnetic tape, which can be read and stored, need only be input once, regardless of how many

different applications are desired. Typing data onto a keyboard once, rather than numerous times, is one of the principal advantages of the information retrieval system. The newness of the input equipment and the inflexibility of the *format,* or instructions to the computer which identify each data element, are temporary disadvantages. However, some input devices such as memory or wordprocessing typewriters have so many advantages that they should be considered additional benefits of adopting an information retrieval system. The way in which the data is input to the computer, or the input options, may be considered the first systems attribute. A few input devices, such as paper tape, should probably be ruled out in any new system; others, such as card input, are limiting and cumbersome but may be suitable to some institutions, like our own, which cannot or will not buy or lease other input devices and which process data at a remote facility where keypunch services are available.

A second characteristic of a retrieval system that should be considered is how it can be accessed. Computers can communicate with input-output (I/O) devices to which they are directly attached (off-line) or connected by phone hookup (*on-line*). In on-line hookups the user pays the computer facility both for the elapsed "connect time" and the time used by the computer to process his/her request. The I/O devices are paid for separately. In off-line use, the computer center charges for time used by the computer and for the I/O devices. Systems designed to take advantage of on-line capabilitites are designed so as to interact with the user. Each instruction elicits a response which helps the user determine how best to formulate the query. On-line, interactive systems are more expensive to operate, but generally give greater precision and better recall than off-line systems. Off-line systems are best searched in *batch* mode, in which a number of questions can be processed at the same time. While these inquiries can be run late at night on cheaper computer time, they cannot be changed in response to the kinds of information which an interactive system can provide (such as the number of postings for each term, or related terms, or other forms of the word searched). This information must be sought by the user from thesauri and indexes compiled in prior computer runs but examined manually before articulating the search query.

A third significant feature of an information retrieval system is the *file* structure, or the way the data are stored in the data base. Each data item must be uniquely identified in the computer. This is done by giving each an *"address"* within the system. In some systems data is stored randomly and accessed directly by its address. In other data bases, the data is stored in files and arranged alphabetically, numerically, or in another logical order; these are called sequential files. They may be indexed files reserving a segment at the beginning of each file for an index to its contents, or not. File structure influences both costs and searching capabilities. Searching for data in an indexed sequential file is less expensive than searching a sequential file without an index. Random storage of data makes good use of large computers for interactive searches, but may be a very expensive way to store data if the most common application is to be a listing of the data arranged by one variable (e.g., a name index). File structure considerations can be very complex,[3] but awareness of the alternatives and their possible implications is sim-

[3] An excellent review of literature on assessing file structure is Chung-Shu Yang, "Design and Evaluation of File Structures," *Annual Review of Information Sciences* 13 (1978): 125–43.

ple and can be extremely valuable even without further technical help.

A final major characteristic to consider of information retrieval systems is the output options it provides. Output limitations may be either format restrictions or media limits. In some systems the format of the data on a page is virtually free of restriction, and changes in format from one application to the next require only simple new instructions easily learned by the non-programming user. In others the formats are virtually unalterable. If a more restrictive system is being considered, all desired applications should be carefully reviewed; if possible they should be tested. While this discussion focuses on the printed page, output can be in various media, such as a display on a cathode ray screen, *electronic photocomposition* instructions, or microform. If available, such options may save the archives many intermediate steps in preparing reports or guides, and may be important variables to consider in adopting any system.

How Do We Relate Needs and Systems Design?

Identifying systems which might be appropriate for our archival applications is as difficult a task as assessing these systems once they are located. A few systems have been discussed in the archival literature—SPINDEX, NARS A-1, SELGEM, PARADIGM—but archivists are unlikely to be acquainted with general data base management systems or flexible information retrieval packages. Existing directories are of some help,[4] as is the professional literature of the information science community; but in these we find so many apparent possibilities that some quick method of eliminating a large number is essential. Our method was systematically to ask potential programming consultants, whom we were interviewing to serve as the intermediaries between us and the system during a testing phase, to go over our self-definition of needs and to assist us in converting these into terms clearer to them as programmers and systems analysts. As part of the interview process this helped us to evaluate the skills of the potential consultants in relating archival needs to systems capabilities. It also made us aware of the weight we assigned to each of our criteria and the problems we could expect from systems. Our method was important in serving to identify systems which might be applicable to our purposes, since we asked the interviewees to name systems they knew which might serve. By asking subsequent candidates to evaluate the choices of prior candidates, and recording their objections and further recommendations, we were able to draw up a short list of software packages not previously applied to archival situations, at least to our knowledge, but seemingly suitable to our needs. Then we wrote for *documentation* of each of these systems.

The interview process had begun the task of translating our general system desiderata into specific criteria directly related to features of software which we would encounter once we began to examine concrete documentation of the systems we would assess. It led also to the employment of a programmer who understood our needs and who helped us translate them further into systems terms. This radically changed the appearance of the implications listed earlier, but not the intent. By the conclusion of the process our list appeared as follows:

[4] Martha E. Williams and Sandra A. Rouse, compilers and editors, *Computer-readable Bibliographic Data Bases: A Directory and Data Source Book* (Washington, D.C.: ASIS, 1976) is a looseleaf service, updated annually, which lists many systems. Other guides and directories appear regularly.

HARDWARE:
—System uses university computer on which we had favorable rates, card input, line-printer and Computer Output Microform output.

SYSTEM STRUCTURE:
—More than one distinctly structured data base can be constructed.
—Data bases can be linked for search or index purposes, or used separately.
—System does not contain restrictive limits on length or number of *data fields.*
—It functions with missing data in any variable.
—It does not contain restrictive limits on lengths or number of *records.*
—Records may be subdivided as desired.
—Any data element may be searched.
—Indexes can be generated on any term.
—System can accommodate multi-level index terms.
—Data bases can be searched by Boolean operators.

SOFTWARE:
—System is fully documented.
—Software can be easily learned, instructions are simple and flexibility is built in.
—Purchasing software is cheaper than developing our own.

Equipped with such a list of criteria, we were prepared to make a decision about specific automated information retrieval systems. By this time three major evaluation efforts had been completed: we had examined finding aids and other services of the archives and established the personnel and materiel costs of each phase of the archives' information processing activities; we had arrived at a list of implications of the context, functions, goals, and constraints under which the archives operates; and we had searched both for systems and for a consultant who could help us set up the systems.

In the Survey of Sources appraisal process we discovered at this stage that at least three important criteria had been overlooked in our considerations because we had unconsciously made decisions which precluded some options. We had assumed, since the computer center was remote from us and we had decided not to purchase terminals, that we could ignore on-line, interactive capabilities. Later we were persuaded that researchers would be more likely to use an interactive system, even if we did not, and we reconsidered our criteria. Secondly, we had assumed that the index terms which the system would search would be supplied by us, not generated from the text of the record by the computer. Full text searching had not been thought about, but when we did consider it we chose to continue to supply index terms, in part because of the varying quality and multiple languages of the descriptions which served as the data for our system. Finally, we had assumed that the *line-printer* output, perhaps on microfiche but usually on paper, would be the final product of our searches. As a result we made no provision for printing instructions or even for distinguishing between capital and small letters. At the time, none of the systems we examined had electronic photocomposition options; but now that some do, we recognize both a hidden assumption in our assessment and an unfortunate consequence of using punch-card input blind to capital and small letters. No doubt a different selective blindness faces us all; one way to avoid pitfalls, however, is to consider as wide a variety of systems as possible.

Which Criteria Are Crucial? Which Secondary?

The most basic rule in assessing the viability of any system should be that it is adequately documented. Any system without adequate documentation should be avoided, no matter what claims are made for it or how well it has been seen to work in practice. Lack of documentation reflects a lack of concern for potential users and will almost certainly lead to complications which will become unnecessarily expensive to solve. In 1976, when we began evaluating systems, two of the six we finally examined with care appeared to be fully documented but had been challenged by users who claimed that documentation was inadequate. At the time we considered it, SPINDEX was one of these. Now it is fully documented and supported and has a user lobby (the SPINDEX User Network) which can help to develop new software and advise on hidden capabilities of the system. Because of its flexibility, the cooperative advantages of its use, and the electronic photocomposition features of the latest version of the software (SPINDEX III), SPINDEX is becoming a leading contender for the all-purpose archival information system. SPINDEX would not have met our 1976 criteria or our requirements today in any case, since it lacks the (to us essential) capability of being searchable. The SPINDEX system is oriented to the printing of guides rather than to information retrieval in response to specific user queries.

Two other systems we examined were dismissed, notwithstanding their powerful indexing and searching capabilities, because they were limited in their ability to sort alphabetic fields of *variable lengths,* which were the principal characteristics of our data and which will almost certainly be present in other archival descriptions as well. Of course it is possible to adapt information about archival holdings to fit into systems designed for business applications—such as one we evaluated, the MARK IV system—or to the fixed formats of bibliographic descriptions used successfully in numerous library applications. Such systems are readily available, have inexpensive search capabilities, and may even be in use in other departments of the parent institution which the archives serves. There are pressures to adopt such systems, especially from within the organization and from programmers who are more familiar with these applications. However, our finding aids do not fit well into these procrustean beds.

Unlike inadequate documentation, these limitations present us with a choice. We may opt to use more complicated, expensive, and flexible systems to accommodate data of the sort we are most used to; or we can change our methods of description and our output requirements to use fields of more uniform, limited, and definable lengths and benefit from more efficient automated systems. Such systems could completely replace the layers of intellectual control we now maintain, from inventories and descriptions of individual collections to repository level descriptions and national catalogs; but they would not be used primarily to print guides such as those we have prepared manually for years, but rather would comprise a data bank, linked in a national network, searched interactively, and available to archivists and the end user. In the long run I suspect that the decision to retain traditional formats rather than adapt ourselves to the possibilities of large scale information systems may prove unwise; but for a project of three years duration, in which the end products were to be familiar to archivists, such long-term considerations had to be dismissed in favor of more immediate practical concerns.

And the business and library oriented systems we examined did not meet the previously formulated criteria.

Finally, two other software packages we examined were too closely designed around specific applications to suit our needs. They contained field length and record length limits which would have defeated our effort to retain as much of the information as we could in the same terms in which it was supplied to us. Both systems were also judged inadequate to our needs because of a more general failing relevant to most archival automation efforts. Each permitted the creation of records to which a number of index terms could be supplied, and each could be tinkered with to allow the creation of sub-records. Thus it was possible to assign index terms to a record called the John Dos Passos papers, or to a sub-record called correspondence 1924–25, but neither of these systems allowed us to assign index terms to a number of different levels within the same collection and to be referred back only to the part of the record to which the term referred. Each sorted the entire record on all the terms which were supplied, and none could link records in such a way as to permit only the collection to be searched, or only some other level such as series or box, at the same time. These systems, of which the PARADIGM system is the only one still in use, seemed to be essentially administrative control tools, and not intellectual access aids.

The Survey of Sources System

Because the systems we examined proved inadequate to our needs, we faced the options of scaling down our requirements, expanding our search for systems, or developing our own software. If other archivists with needs such as ours were to begin appraising systems today, they might find one which could satisfy their criteria; but at the time the prospect was dim that we would find an already developed system by expanding the search. Furthermore, because we could justify the demonstration of systems design as a reasonable extension of our overall demonstration purposes, we decided to develop our own system.

That system is now completed. It consists of three data bases linked together for certain indexing and searching purposes. The first contains biographical data on scientists. Unlike the other files, it is not available interactively because further development was suspended when it proved more expensive to research and prepare biographical data than seemed justifiable for an essentially archival project. The second data base is a straightforward bibliographic file containing fully indexed secondary source citations relevant to the history of biochemistry and molecular biology. The third is an archival information retrieval system.

Data are input to the archival data base by punch cards. It can be searched either interactively, on-line or off-line, usually in batch. Output is available at a terminal, remote printer, or on computer output microfiche. The software that was developed can be used by operators without special training. Data input is in a format close to that of the English language sentence in order to capture almost any text in terms close to the original. Data output in response to reference queries is in pre-established formats close to English, or it may be tabular for reports and guides.

Any field may be searched in any Boolean combination with any other field. Indexes may be generated to provide the context in which each term appears and

specify the level of the collection which the term describes. A special feature of the system is that it employs index terms containing numerous facets; so the phrase "correspondence of John Doe with his colleague Harry Jones, from 1919 to 1923, concerning X university, department of biochemistry, tenure reviews" could be incorporated within a single term. The term could also contain further information about the quantity of correspondence and the location of the material within the collection, and could always be linked with information about restrictions which might apply to the use of the material. Such terms are disaggregated and then rearranged in indexes organized by personal names, corporate names, subjects, dates, and kinds of materials.

The system we developed for our purposes has other applications as well, but it is limited in ways important to anyone who might consider its use. To reduce machine costs, personal and corporate names are actually searched on eight or twelve-character mnemonic codes rather than on the full name, even though the full name appears in output. As a result, authority files must be maintained carefully, and often must be consulted in order to formulate a search. Output formats are set to our specifications, and variation of them requires reprogramming. Furthermore, while searching the system has been made extremely inexpensive by the generation of *index* sequential files for every facet of the index terms described above, the derivation of these files from the master data base, and hence all updating, is relatively expensive. We have purposely chosen to bear the costs of generating simple-to-search files, instead of passing these costs on to the users.

The Survey of Sources will be finished in June 1979, after which the software we developed will be available to anyone for the cost of a copy (as, incidentally, will be the data bases themselves). However, while the system is fully documented it will not be supported, and any further developments to the system will be the responsibility of users. Indeed, it is not our intention to encourage the adoption of this system, and its description in this essay is only the end result of tracing the history of our efforts to assess automated information retrieval systems for an archival application.

Conclusion

The automation of archival information may be considered in a number of broader contexts. It may be part of a strategy for expanding institutional services, cooperating with other archives, reducing the demands of routine tasks on staff time, or integrating the archives with other departments of the institution of which it is a part. Whatever the larger purposes are, automation will succeed only to the extent that it takes into account the specific needs and goals of the archives. When systems that meet these needs are successfully adopted they will almost certainly change traditional archival practices. To the extent that they force us to standardize within and between institutions and to pay closer attention to the problems of establishing intellectual control through subject access, they will result in an improvement of our methods. If the systems we adopt are chosen with attention to our needs, we may discover how many of our practices have developed because of the constraints of manual processing, and we will be able to alter these and retain only those descriptive techniques which truly provide more accurate and rigorous control. For the users of archives, archival automation holds

many promises that are obvious benefits of better control; it offers archivists not only more satisfied clients but also an opportunity to add significant new skills to their armamentarium, and, indirectly, to augment the respect in which the profession is held.