MARTIN DILLON

The Impact of Automation On the Content of Libraries And Information Centers

Information needs are growing more rapidly than the abilities of research libraries and information centers to meet them. Two reasons and their influence on information systems are discussed: a shift in scientific endeavor from basic science to applied, leading to the emergence of programmatic research; and the technology of science itself.

INTRODUCTION

IT HAS BEEN GENERALLY AGREED IN RECENT YEARS that information needs are growing more rapidly than the abilities of research libraries and information centers to meet them. Most often, the reason given for this phenomenon has been the growth in the amount of published literature. Two additional reasons, which are little noted though perhaps more important, will be discussed here, along with their consequences for information centers: a shift in scientific endeavor from basic science to applied, leading to the emergence of programmatic research; and the technology of science itself. These sources of difficulty will be discussed briefly before considering their influence on information systems.

BASIC TO APPLIED TO PROGRAMMATIC RESEARCH

Scientific research is traditionally divided into basic and applied, where the former is described as an activity directed toward "a fuller knowledge or understanding of the subject under study, rather than a practical application thereof," and the latter is "directed toward practical application of knowledge."[1] No great insight is necessary to see that stock in knowledge for its own sake has taken a tumble in recent years.[2] A more business-oriented federal government, a greater consciousness of the ill effects of socially undisciplined research (the Vietnam war), and a heightened awareness of short-term social needs compared with long-term benefits from basic research—all of these contribute to the disenchantment with basic research and a consequent shift in emphasis toward practical application in the ways we use knowledge and generate new knowledge.

In particular, a new form of applied science is emerging: programmatic research, the marshalling of technology and men to the achievement of some change in the world. As it is used here a program specifies a sequence of actions organized and directed toward solving a specific problem, or system of related problems in our physical environment, as contrasted with efforts which attempt

Mr. Dillon is in the School of Library Science and the Department of Computer Science at the University of North Carolina, Chapel Hill.

to discover, describe, or add to knowledge. Urban redevelopment, environmental protection, population control—the pursuit of goals such as these circumscribe action-oriented disciplines with relatively specific, even short-lived goals. As organizations of technological and scientific effort, these efforts contrast with knowledge-oriented academic disciplines where goals are diffuse.

Since it is possible that these new disciplines are transitory, that as the problems which prodded them into existence disappear, they will also, it may be asked why we should consider developing approaches and techniques to cope with their special needs. There are two reasons. Such problem-focused disciplines are important enough on our current horizon, both in size and import, to elicit special attention from information scientists. Moreȯver, even if today's distinct forms disappear, the generic activity is likely to persist: other special forms will arise to take their place, organized around other large problems but with similar information needs.

How do the information needs of such activities differ from normal science? This question is hard to answer, and a complete answer will not be attempted here. On the one hand, the information needs—both in sheer bulk and in management tools—are usually greater: normal science can often advance fragmentarily, through the efforts of individual scientists and without integrative paradigms. Spaceships or antiballistic missile systems cannot, generally, be built that way. Usually the new sciences are conducted by large multidisciplinary teams organized in an explicit way. There are delineated lines of authority and responsibility, with subgoals and divisions of labor clearly specified. Increased organization increases the burden of communication. The increase and demand for communication often lead to informal channels, and an overall drop in system efficiency. So far, little is understood about how best to organize information for such efforts.

The Technology of Science

A second source of trouble for libraries and information centers, contributing to the gap mentioned above, and one which in the long run may be more significant, comes from the product of science: technological development. The advent of the computer and the subsequent growth in its use is rapidly redesigning most human efforts, and science is no exception. The degree of complexity which can be meaningfully managed either for practical ends or in basic science has grown enormously in the last decades.

One can appreciate the extent of the change by considering application areas where the techniques of operations research are suitable tools of analysis. Changes of degree—in terms of the size problem which can be meaningfully tackled, and amounts of data which can be represented or analyzed—are so great as to produce changes in kind. Linear programming models can employ thousands of constraints to model problems where fifty were excessive before computers were available to solve the resulting system of equations.[3] Simulation models accounting for thousands of relationships likewise are beginning to be commonplace.[4] In statistical modeling, to mention only one change, factor analysis of hundreds of variables is possible where twenty was arduous labor for the analyst using a hand calculator.

In similar ways, the computer is revolutionizing the technology of science, necessarily leading to changes in the conduct and conceptual structure of science. The paramount change to date has been in the role of data: the amount which can be easily manipulated has increased by orders of magnitude. Though often altering little more than the bookkeeping, this radically enlarged empirical basis will no doubt soon lead to qualitative changes beyond bookkeeping and affect the very nature of science.

Each of the above points has serious implications for the organization and operation of libraries and information centers. The discussion which follows attempts to highlight the more serious, organized around three central artifacts of science: (1) *documentation:* the detailed public exposition of research results, whether basic or applied; (2) *theories and models:* the construction of formal representations of research results which synthesize, integrate, or explain them; (3) *data:* the organized groups of symbols or numbers which are the results of scientific observations of the world and serve as the empirical roots of theory.

Though it will not be argued here, it is likely that all these artifacts are poorly understood—both in what they contain and in how they do or should function. Automatic information retrieval systems have had very limited success in explicating documents and their use; philosophies of science have made little progress since Newton; and both macro and micro physics have their trouble with data (in giving exact meaning to very small or very large measurements). These fundamental difficulties at the root of scientific activity are a major source of uncertainty in designing information systems which function as faithful adjuncts to scientific research. The implicit emphasis here is not on science itself, but on systematic program planning where the same uncertainties are ameliorated somewhat by explicit program goals. The systems design can be evaluated for such programs in ways unavailable in basic science.

Documentation as Systems Description: The Organization Gap

In research undertaken to support some facet of program planning, the achievement of program goals is paramount and serves as a focus to the research. Such an obvious observation seems barely worth making were it not that the organization of its documentation rarely clarifies the role of such research in the overall planning effort. Access to research documentation is usually determined by channels and methods within the discipline in which the research was carried out, not through program-determined organizations of knowledge. Very rarely are attempts made to show relationships among interdisciplinary materials or to model document collections on program needs.

Some examples will help here. In the field of population activities, a fairly single-minded short-term goal can be cited: control of population. Related research and its documentation is growing at a remarkable rate, but as yet no substantial subject indexing system is available, much less one specifically organized to aid in the development of population programs. Moreover, the research effort itself, which is presumably geared to aid specific aspects of program design or implementation, cannot by any reasonable method be connected to programming concepts. And the related research is being carried out all over the globe and by workers in many disciplines. This research area, due to its easily expressed global goals, is more organized than some which might be cited. Urban redevelopment has barely started on such efforts and as a consequence information organization is solely along specialty divisions, or *ad hoc* constructions of local libraries.[5]

Such defects in documentation control are not to be blamed necessarily on the libraries. At least part of the responsibility lies with current problem-solving techniques as they are embodied in science: they are fragmentary, rarely explicit, and probably incoherent. Even the health sciences, presumably guided by expressible goals, pursue them unsystematically, in an order and with an emphasis determined by the puffs of perfidious politics.[6]

As a consequence, organizations of knowledge specific to a discipline are inferred from existing literature; documentation of research undertaken to support programming sporadically borrows from these existing structures and attempts, in a makeshift manner at best, to relate them to program goals. For the multidisciplinary research teams which participate in program planning on a large scale, communication through a centralized document collection is as essential as the rarity of such systems in practice. The nature of the tasks currently under attack (waging war, exploring space, remaking cities) usually involves natural, social, and engineering scientists, where differences of viewpoint are almost cultural.

A slight alleviation of the organization gap is available through automated systems which allow some degree of user organization through formulation of specific queries. Such systems in conjunction with vocabularies developed in cognizance of these problems go some way toward facilitating use of interdisciplinary document collections. When two or more disciplines have a common object for analysis or similar goals, their vocabularies usually express this intersection, though not systematically. Exactly how this occurs has not been investigated as yet, but since the practitioners of each discipline share a common world and a common natural language, it is easy to see why. Since the overlap is only rough, the individual scientist must interpret the details according to his own light and in conjunction with his own goals. Though such techniques are still primitive at best, their principles of operation shed some light on the defects of organization in science and its documentation.

Computer Programs as Theory

A further problem, and one whose implications and consequences have been even less well perceived and attacked, is the growing tendency of computer programs to embody the essential properties of theory. The reasons for this development are fairly straightforward: programming languages possess many of the communicative advantages of formal languages used in mathematics: they say what they mean in a sense so literal that to translate their logic into other languages is not attractive. Second, the computer program is always a strictly formal entity: it is well defined and its parts have a clear meaning (to the initiated). Third, the program itself is available (the complete theory) and can always be used by someone wishing to explore the consequences of a particular formulation, or to trace the effects of specific assumptions.

This development is particularly noticeable in applied fields where large models are the vehicle for exploring system interactions. The model can either have an explicit mathematical formulation, as do linear programming models, or, there may be no alternative representation, as is so often the case with simulation models ("A simulation model is a theory describing the structure and interrelationships of a system.")[7] Both techniques are of increasing importance in program planning due to the complexity of the problems which must be solved. In either case, the computer model serves as a theory for the system which is being modeled.

A good example of this tendency is available in Jay W. Forrester's *Urban Dynamics* where a simulation model of urban areas is described.[8] Indeed, Appendix A is entitled, "The Model–A Theory of Urban Interactions," and the language used to express system functions and equations is exactly the language accepted by a computer system for constructing and executing simulation models. The dependence on the computer formulation is understandable when one considers that the model contains some 150 equations which interact to describe the urban areas in complex ways, and involving hundreds of parameters and variables. Equally interesting is the "world model" described in *World Dynamics* and developed in *The Limits to Growth*.[9] The technical problems are similar: great complexity and detail; the solution the same: formulate the theory in terms of a computer simulation model.

How these developments affect the underlying assumptions and formal apparatus of science cannot be determined as yet, but sure to be altered is the shape and function of theory. Part of the reason for this is the speed with which developments are occurring. There was a time when today's theory was tomorrow's computer program. Increasingly, today's program is today's theory. For program planning especially, computer models replace theories, both as organizations of knowledge and predictors or determiners of the future. It is likely that we are at the periphery of such use, and that the future will see more and more of it.

Whether this development is to be lauded or regretted may be debatable. Second class theories or not, computer models cannot easily be dismissed: their numbers are growing. In this context, the point to be made is that the computer program represents something essential about the theory, and the theory is often approached and understood by researchers through its computer representation. Libraries and information centers which support research, if they are to satisfy the information needs of their users, must get into the business of providing access to such programs. Exactly how this need should be met remains a mystery, though some ideas follow.

The Changing Role of Data

It is obvious that as computer programs like those cited become the standard means of communicating results of research, the role of data in libraries and information centers will grow apace.

In many ways, scientific documentation is primarily processed data. The contents of research articles are often formed from samples of the data and fragmentary evidence in support of the author's conclusions; when the conclusions are questioned or, more often, when different questions need to be asked, the data is more valuable than the documentation. As a form of knowledge, the article becomes less attractive as the ways of processing data increase. As the variety of analytic techniques increases, the likelihood that an analyst will be satisfied with this or that particular analysis decreases. We now have automated procedures for data analysis: everyone becomes his own analyst and can perform his own analysis tailored to his own needs.

The same point can be made in relation to the simulation models cited: they are ways of processing data; they are a means for digesting pasts. Both developments increase the importance of raw data. Libraries of data are becoming commonplace and, as a national asset, it can be argued that the Bureau of Standards with its data collections is more valuable than the Library of Congress. Certainly for science and technology this is true. As evidence of this changing role of raw data, two important examples can be cited, each of which incorporates extensive data bases and a means for selectively processing them. At the Bureau of Labor Statistics, U.S. Department of Labor, a system has been developed to provide analyses of a growing body of U.S. economic data (described in "The Computer and Economic Analysis at the Bureau of Labor Statistics").[10] At the Bureau of the Census, a more elaborate system has been under development to handle demographic data, primarily the 1970 census data.[11]

Do these developments have implications for research libraries? It is after all an accident of technology that books and journals are the vehicles for storing and communicating research results. As data becomes a more dynamic part of an information system, continually reanalyzed from varying points of view, representations of the data in printed form are reduced in value, and the corresponding computer representation has increased value. If libraries and information centers are to continue as vital adjuncts to science, they must accommodate themselves to this shift from product to process.

Integrated Information Systems

From the foregoing it would appear that scientific research efforts would be best supported by an information system of three major components: (1) A documentation component, which included interactive text editing facilities as well as retrieval capabilities. Question-answering systems as they are currently understood would derive from this component, insofar as they are based on natural languages such as English. (2) The second component would include data bases, especially those which contributed to the technical papers in the document section. They would normally contain far more, even data which had not been documented, though unanalyzed data would require sufficient definition to be used by the community. (3) The third component would be the techniques for analysis, especially those which were actually used in the reported literature. The term "techniques" as used here is merely a euphemism for computer program or technique available through one.

It is likely that the long-term solution to these problems will be through integrated systems such as Project INTREX where data, analytical procedures and documentation each will have a place.[12] In such systems, users will communicate with one another and to their programs and data through terminals in an interactive environment. With common file definitions and the facilities for working with them, the frame exists for an on-line community of scholars or research specialists: instant publication.

Unfortunately, such systems are far in the future. What can be done in the interim? For information centers already employing automated components the hardware and software technology exists to eliminate or minimize some of the deficiencies noted above. Each of the three facets of scientific research—documentation, programs, and data—has been dealt with separately and fragmentarily by different approaches and procedures. Though no system exists which incorporates all three suitably, they can be had individually with less effort than might be supposed.

The first major task is to tie the production of research documentation more closely to the automated system. This can best be done within the constraints of current technology, through available text editing systems.[13] Text-editing systems allow alterations and corrections to be made to manuscript material through its computer representation. The advantage is that only a small portion usually need be changed, the fitting of the text to pages, including altered pagination, spacing, paragraphing, etc. being done automatically in a subsequent reprinting of the text. Since no new errors are introduced, such systems usually reduce the overall labor and improve the product at the same time. In theory at least, text-editing systems can interface directly with computer driven printers, removing the need for additional proofreading (especially valuable for texts heavy with formulas). Why they are not in more widespread use is a mystery, but one which will not persist for long: their advantages will soon make them commonplace.

In this context such systems are cited for making available machine-readable versions of documentation at their origin, increasing the amount of text so available as well as doing it more speedily. Thus eventual use of such text for retrieval or question-answering is assured.

The second problem—availability of analytical techniques or models in the form of computer programs—requires for its solution better management of research efforts, and more cooperation (legislated or otherwise) among libraries and information systems. For example, we have not yet reached the stage where research designs contain, or in the case of federal funding, are required to contain, explicit means for communicating, in addition to the conclusions of the research, the analytical techniques which were used and the data they were applied to. To get some idea how this approach works out in a book medium, consult Cooley and Lohnes, *Multivariate Data Analysis* where a national survey is referenced throughout as a source of examples.[14] The book itself contains copies of the computer programs necessary to duplicate most of the analyses which were carried out on the data, and the data can be obtained through the authors.

Such an exemplary practice will more and more be copied as its value to the research community becomes more obvious. A concomitant responsibility falls on the information center to make such corroborative or subsidiary tools as computer programs and data available as well as the documentation: they will soon become as essential, if not more essential than the documentation. As analytical techniques become more standardized, it will become easier for information centers to provide them to users. Programming systems which include global file definition and a broad selection of statistical procedures have been commonplace for some time. The Biomedical Package (BMD) series and the *Statistical Package for the Social Sciences* are two of the more generally available examples.[15]

A more critical example, apt since it deals with information retrieval, is in the SMART efforts carried out at Cornell University over the last few years.[16] Part of the burden of the research was the development of specialized files and procedures and their organization into a single system capable of achieving the research goals of the participants. The system itself is available from Cornell and interested researchers are able to carry out duplicate tests, either on the original or their own data, as well as inventing their own experiments.

Conclusion

The final solution, like all final solutions, is far in the future. What can be done immediately? Some action is required on the federal level: the specification of data interchange codes, the standardization of analysis techniques; the requirement that all federally funded research specify fully and in advance the form and ultimate end of any data; standardization of bibliographic records, publication standards, and far more.

On another level what is needed is a more thorough analysis of the relationships among the sciences, and the development of common tools of analysis and common languages. At the same time, more attention needs to be paid to the development of scientific planning methods: too often seat-of-the-pants decisions are based on seat-of-the-pants reasoning. From such studies should emerge better techniques for controlling the access to documentation for the purposes of improved planning, both in the use of scientific resources and in the development of social programs.

Even in the absence of these obviously worthwhile endeavors, information centers must develop ways of managing more than documents; they must develop a means of controlling data and programs as well, and understanding the use to which their users would have them put. As more information centers automate their services, they will more easily be able to extend them to include making available computer programs and data in electronic form.

References

1. These definitions are taken from *National Patterns of Research and Development Resources: 1953–71,* NSF 70-46, p.24–25, and were (in part) used in the questionnaires to gather the data appearing there.
2. Ibid., passim, is available to aid insight. The document is a good summary of research and development expenditures in the U.S. The key figures are: a drop in total R & D spending from 3 percent of the GNP to 2.7 (Chart 3, p.3).
3. William Orchard-Hayes, *Advanced Linear-Programming Computing Techniques* (New York: McGraw-Hill, 1968) presents the state of the art as of 1968.
4. Recently, the capabilities of formulating simulation models for computers has been extended to using the English language. See George E. Heidorn, "The Specification of Decoding and Encoding Processes for Natural Language Man-Machine Communication," paper given at the 10th Annual Meeting of the Association for Computational Linguistics, July 1972.
5. Brenda White, *Sourcebook of Planning Information: A Discussion of Sources of Information for Use in Urban and Regional Planning and In Allied Fields* (London: Bingley, 1971).
6. F. W. Lancaster, *Evaluations of the MEDLARS Demand Search Service* (Bethesda, Md.: National Library of Medicine, Jan. 1968) gives some insight into this problem.
7. Jay W. Forrester, *Urban Dynamics* (Cambridge, Mass.: The M.I.T. Press, 1968), p. 112.
8. Ibid.
9. Jay W. Forrester, *World Dynamics* (Cambridge, Mass.: The M.I.T. Press, 1968); Donella H. Meadows, et al., *The Limits to Growth* (New York: Universe Books, 1972).
10. Rudolph C. Mendelssohm, "The Computer and Economic Analysis at the Bureau of Labor Statistics," *The American Statistician* 22 (April 1968).
11. Abbot L. Ferris, *Research and the 1970 Census* (Oak Ridge, Tenn.: 1971).
12. *Report of a Planning Conference on Information Transfer Experiments* (Cambridge, Mass.: 1965) remains the best overview of the system's design and major goals.
13. Andries Van Dam and David E. Rice, "On-Line Text Editing: A Survey," *Computing Surveys* (June 1972), p.65–79. Of particular interest are *Text 360 Reference Manual Operation Guide* (New York: IBM Corp., 1968) and Andries Van Dam, *FREES (File Retrieval and Editing Systems) User's Guide* (Barrington, R.I.: Text Systems Inc., 1971), the former for its general availability; the latter for the originality of its facilities.
14. William W. Cooley and Paul R. Dohnes, *Multivariate Data Analysis* (New York: John Wiley, 1971).
15. W. R. Schucany, Paul D. Minton, and B. Stanley Shannon, Jr., "A Survey of Statistical Packages," *Computing Surveys* (June 1972), p.65–79.
16. Gerard Salton describes the system and its use in *Automatic Information Organization and Retrieval* (New York: McGraw-Hill, 1968).