PROCEEDINGS OF THE 1964 CLINIC ON LIBRARY APPLICATIONS OF DATA PROCESSING

Graduate School of Library Science
University of Illinois
UNIVERSITY OF ILLINOIS GRADUATE SCHOOL OF LIBRARY SCIENCE

PROCEEDINGS OF THE 1964 CLINIC ON LIBRARY APPLICATIONS OF DATA PROCESSING

Held at the Illini Union on the Urbana Campus of the University of Illinois, April 26-29, 1964

Edited by

HERBERT GOLDHOR

Distributed by
The Illini Union Bookstore
Champaign, Illinois
The papers and discussions at this second annual Clinic on Library Applications of Data Processing have demonstrated conclusively that the use of this new tool is not the wave of the future but that it is something which is already here. Over fifty university and special libraries were represented at this Clinic, and they are either already using a computer or are well along in their detailed planning for its use. And of course there is no reason to think that all libraries using computers chose to send someone to this meeting.

These papers and other publications have much to say about the technical aspects and implications of this new machine and of the approach to library operations which it requires. What should library schools do about it? Certainly it would be unfortunate if library schools generally were to ignore developments in this field, but neither should they accept the new tool blindly or uncritically.

We at Illinois are convinced that the first long step in the use by libraries of data processing will be to mechanize their present routines. This is not only more necessary and more obvious but also requires far less new theory than does information storage and retrieval, though the latter is undoubtedly of much more potential importance. In any case we here plan to emphasize this first main stage of development for the long present.

Furthermore we see our role in this field not as theorists or pioneers but as intermediaries between those who are the innovators and those who are the practitioners. We hope to utilize whatever means are open to us to translate theory into terms which are meaningful to the librarians in the field. We hope in time to develop some research projects here of our own, but many other people are doing important new work in this line of activity and we hope always to remain critical of what is being done and eclectic in what we teach.

There are several ways by which we can play our role as mediator or interpreter. For one thing we do offer one course in this field, at the graduate level. The opinion has been expressed that all our students should be required to take this course, but we have left it optional. As a matter of fact, enrollment in it has been good—so much so that we plan to offer the course every spring and every summer. The course was developed and has been taught in the spring by Dr. Frances B. Jenkins, of the University of Illinois Graduate School of Library Science faculty. Guest instructors are used in the
summer, e.g., Dr. Ralph Parker, librarian of the University of Missouri, in 1963.

A second way by which we hope to contribute in this area is by a workshop on the writing of computer programs for library operations. Such a workshop was held in the summer of 1964; it was successful and will be repeated in 1965. This is the sort of technical skill which is needed by at least one person in every library which attempts to use a computer for even routine operations. In the third place, we plan to publish in this field, e.g., the proceedings of the Clinics. In July 1964 we published John Melin’s summary of library use of data processing to date, as Occasional Paper no. 72.

Fourthly, we hope to continue these Clinics on an annual basis. These conferences are called Clinics because they consist primarily of papers recording the experiences of individual libraries. We think that this emphasis on the case approach is valid and appropriate under present conditions.

I wish to acknowledge with thanks the efforts of my colleagues who helped make the 1964 Clinic a success. Dr. Frances B. Jenkins and Dr. Rolland E. Stevens served with me on the planning committee. Mrs. Maija Harris was our administrative assistant, and Miss Jean Somers helped edit the papers. Mr. Hugh Davison and his staff in the Division of University Extension handled the arrangements for the Clinic. The speakers are all owed a word of appreciation for their cooperation and their contribution. In particular I wish to thank Robert Wallhouse of IBM, for showing two films on the 360 computer. The registrants made the whole affair worthwhile by (a) coming, (b) participating, and (c) teaching us the lessons of their own experience. To all of these and others not named, my sincere and heartfelt thanks.

Herbert Goldhor

Urbana, Ill.
May 16, 1964
OTHER VOLUMES IN THIS SERIES

Proceedings of the 1963 Clinic on Library Applications of Data Processing. Pp. 176. $2.00 in paper covers, $3.00 in cloth covers. Order from the Illini Union Bookstore, 715 S. Wright St., Champaign, Ill. 61803.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOREWORD</td>
<td>iii</td>
</tr>
<tr>
<td>IMPLICATIONS FOR LIBRARIANSHIP OF COMPUTER TECHNOLOGY</td>
<td>1</td>
</tr>
<tr>
<td>Robert M. Hayes</td>
<td></td>
</tr>
<tr>
<td>THE COMPUTER-PRODUCED BOOK CATALOG: AN APPLICATION OF DATA PROCESSING AT MONSANTO'S INFORMATION CENTER</td>
<td>7</td>
</tr>
<tr>
<td>W. A. Wilkinson</td>
<td></td>
</tr>
<tr>
<td>Discussion by W. A. Kozumplik</td>
<td>20</td>
</tr>
<tr>
<td>DEVELOPMENT OF COMPUTERIZATION OF CARD CATALOGS IN MEDICAL AND SCIENTIFIC LIBRARIES</td>
<td>25</td>
</tr>
<tr>
<td>Frederick G. Kilgour</td>
<td></td>
</tr>
<tr>
<td>COMPUTER SIMULATIONS AT THE COLUMBIA UNIVERSITY LIBRARIES</td>
<td>36</td>
</tr>
<tr>
<td>Warren J. Haas</td>
<td></td>
</tr>
<tr>
<td>PRINCIPLES OF COMPUTER PROGRAMMING</td>
<td>47</td>
</tr>
<tr>
<td>Kern W. Dickman</td>
<td></td>
</tr>
<tr>
<td>WESTERN RESERVE UNIVERSITY COMPUTER INDEX OF EDUCATIONAL RESEARCH</td>
<td>56</td>
</tr>
<tr>
<td>Gordon C. Barhydt</td>
<td></td>
</tr>
<tr>
<td>Discussion by William P. McLure</td>
<td>73</td>
</tr>
<tr>
<td>DATA PROCESSING PROBLEMS AT THE DEFENSE DOCUMENTATION CENTER</td>
<td>76</td>
</tr>
<tr>
<td>William A. Barden</td>
<td></td>
</tr>
<tr>
<td>APPLICATIONS OF DATA PROCESSING AT THE CANADIAN NATIONAL RESEARCH COUNCIL LIBRARY</td>
<td>105</td>
</tr>
<tr>
<td>Jack E. Brown</td>
<td></td>
</tr>
<tr>
<td>POSSIBLE APPLICATIONS OF DATA PROCESSING EQUIPMENT IN LIBRARIES</td>
<td>112</td>
</tr>
<tr>
<td>John A. Wertz</td>
<td></td>
</tr>
</tbody>
</table>
IMPLICATIONS FOR LIBRARIANSHIP OF COMPUTER TECHNOLOGY

Robert M. Hayes

In this century we have witnessed the growth of an almost unbelievably complex society. We have seen society change from an essentially rural one to an essentially urban one. At one time, each man knew his neighbor, his town, and his role and his station in life. The role of government, if it was felt at all, was simple. We have seen the change from an essentially agricultural society to an essentially industrial one—even our farms are now largely factories. We have seen the increasing integration of our society into a single whole—an integration wrought by modern transportation and communication, an integration requiring a corresponding social organization. We have seen an ever broadening size of scope of that social structure—until now it is close to encompassing the entire world. We have seen the ever increasing impact of technology—until we have learned not only to accept change but even to expect it. We have seen an ever increasing specialization—an essential ingredient of complexity—until the day when one man could legitimately be viewed as a universal genius is long past. We have seen an ever increasing magnitude, in both breadth and depth, of recorded knowledge—until the day when one man's personal library could be the basis for a national library is similarly long past.

Such complexity, by its very nature but even more because of the speed with which it has developed, must pose problems of corresponding magnitude. National economic problems, industrial management, technological development, social change—all involve decisions of great magnitude, and their impact is felt throughout the social structure precisely because of the complex interactions among its component parts.

Robert M. Hayes is Manager, Advanced Information Systems Division, Hughes Dynamics, Inc., Los Angeles, California, and Professor in the School of Library Service, University of California, Los Angeles.
The day is past when the information needed for these decisions could easily be remembered by the decision-maker himself. The day is past when a delay of weeks or months in a decision could be tolerated—our society is just too complex, and without an adequate memory and rapid flow of information, it will go the way of the dinosaur.

We have all kinds of mechanisms to ensure the rapid flow of new information—but the effects of much that happens now will be recognized as significant to future decisions only long afterwards. Furthermore, much of the significance depends upon the very accumulation of information.

Thus, the complexity of our modern society—science, technology, government, business—has become so great that its very existence is made possible only through correspondingly complex mechanisms for communication, processing, storage, and retrieval of information about itself and the results of its functioning. This may appear to be overly dramatic, yet its truth is demonstrated by the ever-increasing number of information centers, "data-banks," centralized files, and special libraries; the evidence for its importance lies in the ever-increasing concern in science, technology, government, and business that these mechanisms meet their needs for information. The nature of those needs, wherever they exist, is that they are relatively ill-defined and represent a great variety of mutually conflicting requirements. The problem is to meet them within severe economic restraints, so that the "information system" does not itself become a burden.

It is this which constitutes the challenge to librarianship, and all the concerns of the moment—with "mechanization," with "centralized processing," with "economic operation," with "system analysis"—are merely symptomatic, merely the evidence of the crying need for professional knowledge of how to meet the demands for information—ill-defined though they are and severe though the economic restraints may be. Because librarianship does represent the sole existing source of professional knowledge and operating experience in the field of information handling as such, it is librarianship which now feels the pressure of these needs. If librarianship does not meet this challenge and fill the need for professional knowledge, someone else will, but in the process they then must develop the same tools and capabilities which librarianship now provides.

It is my belief that, in large part, the implications of the computer to librarianship today are a result of these pressures and that the real aims should be to lead the profession in meeting the needs for professional knowledge. To support this belief, the impact of the computer on librarianship will be discussed under five categories: (1) operational implications—the computer in the library; (2) systems implications—our national information system; (3) professional implications—the need to understand and to control mechanization,
with knowledge and wisdom; (4) educational implications; and (5) theoretical implications.

Operational implications.—The concern of the library profession with the ever increasing costs of operating complex library systems is representative of comparable concern in all information activities throughout the country. And it is natural to search for the answer to such concern in a better solution to operational problems, and particularly to look for it in the techniques of methods analysis, mechanization, and cost control; therefore, this has been perhaps the most evident impact of the computer on librarianship. In addition, external pressures from administrators, engineers, salesmen, and others, all asking, “Why don’t you automate?” have made this impact painfully evident.

However important the application of these approaches may be for the solution of operating problems, they simply represent the tools of good management and not the substance of the problems in librarianship. There may be some particular aspects of them which are significantly different when applied to librarianship, but fundamentally and in general these tools of good management are not going to have any lasting impact—at least not on librarianship as such. At most, therefore, the aim should be one of educating the profession in the use of these tools, in the special problems in applying them to libraries, and in their relation to the more basic problems in librarianship. In this respect, much of the groundwork has already been done—the profession has been educating itself, has carried out analyses of library operations, has experimented with mechanization, and is developing better concepts of cost control.

It is in part for this reason that this speaker would be concerned if this area became the predominant focus of future concern with automation. It is my feeling that at most the need is to complete the process which has already been underway in the profession.

Systems implications.—It is at times difficult to make a distinction between operational problems and systems problems; they overlap and interact, and one man’s operational problem is part of another man’s systems problem. However, within the context of a national library system, the distinction can be made between those considerations which are local—within a single library or university campus, say—and those which are nation-wide. Interlibrary cooperation is, of course, not a new concept, but rarely has it been done on an integrated basis. It is clear that some degree of centralized processing and allocation of resources can produce not only a more efficient total operation, but even a more responsive one. The systems problems arise from trying to integrate the component libraries for maximum efficiency, without degrading the services locally. The problems in reconciling the conflicting requirements of local operation and system integration are difficult ones. The system
implications of MEDLARS, say, or automation in the Library of Congress, or a National Science Library, are great. They must be understood.

Professional implications.—It is this category of problems which probably represents the most significant departure from the apparent views of others. It is my belief that the scope of professional librarianship is potentially far greater than any presently encompassed by the prevailing concepts of it. The tools of librarianship have application in areas where people are now groping for help. There is therefore a professional responsibility to be fulfilled, but to do so will require an active effort designed to demonstrate the utility of the tools of librarianship and their specialization to particular problem areas.

To be specific: (1) The storage and retrieval of engineering documentation (not just in the sense of technical reports, but more broadly) and project data are woefully inadequate. Only the most primitive steps (such as "configuration management") have been taken, and yet the solution of this class of problems will require the most sophisticated tools of librarianship. (2) The storage and retrieval of business data, particularly management information, is in a state of chaos in even the most advanced business organizations. The developers of "operations research" techniques for scientific management have only recently—and suddenly—come to the realization that those techniques are valueless without control of the information on which they are based. Unfortunately, the form of the information is so diffuse and the amount is so great that the unsophisticated techniques successfully used on inventory files are completely inadequate. Again, the tools of librarianship are essential to the solution. (3) The storage and retrieval of information about geographical regions, such as metropolitan areas, and related political and economical information, are essential to good government—both long-range planning and day-to-day operation. Again, the developers of "economic models" have tended to ignore their dependence upon accurate data, but more immediately have ignored the need for control of it—control which requires the type of professional knowledge librarianship provides.

These examples could be added to by the dozens. In each case, the complexity of a management, control, or research problem has made evident the need for adequate handling of the necessary information. In each case, the only tools used have been the most unsophisticated because professional knowledge was not made available. In each case, these tools have worked only so long as the needs were well structured and the size of the files small. But in each case, the needs have become more diffuse, the size of the files immense, and most important the economic restraints severe.

This represents an enormous challenge to librarianship—the
challenge to apply the tools of librarianship, with confidence that they are necessary, to a broad spectrum of information problems. To do so will require a willingness to handle a correspondingly broad spectrum of physical forms, intellectual content, and classes of users. It will require a willingness to specialize—not just in terms of subject content, but in terms of types of professional tasks.

There is, in addition to the professional responsibility already defined, a social one as well, and one which librarianship is uniquely capable of assuming. In a society as complex as ours, the control and management of information represents a powerful tool which can be put to many uses. The social problems posed by the accumulation of information, readily available, are great. The computer age itself is now only twenty years old and yet we are half way to 1984! It is important that it be viewed with social responsibility, and librarianship as a profession has demonstrated the ability to do so.

Therefore, it is my belief that a professional and social responsibility exists and that librarianship is the best suited to assume it. It will require research into the application of sophisticated library tools to a great variety of forms and types of information. It will require research into the social value of information. And it will require research into the implications—not in the technical sense, but in the social sense—of automation as applied to information files, since in large part it is automation which has raised these areas as ones of immediate significance.

Educational implications.—Recognition of the educational implications of the computer has led university after university to initiate an educational program in information science. However, the burden which library education must carry is already greater than the existing library school curriculum can easily handle. If we now add to it education in the newer methods for analyzing and solving operational problems, in the methods of system analysis, in the extension of subject matter into areas of business and government, in an increased degree of specialization in library functions, in critical social problems in the use of information, and in theoretical foundations, it is clear that a completely new look must be taken. The existing curriculum is not able, either in content or in length of time, to handle the added burden which the computer implies.

It is my suggestion that the change in library education—or perhaps better stated, the addition to library education—will come in three ways: (1) Through increased recognition of the need for specialization—in subject matter and in function—with corresponding orientation of the curriculum toward a limited set of "core courses" followed by a sequence of increasingly detailed specialty courses; (2) Through increased recognition of the need specially to educate library administrative personnel—not so much in the tools of librarianship as in the tools of good management; and (3) Through
increased recognition of the need to develop theoreticians, with broad knowledge in mathematics, logic, linguistics, economics, and engineering in addition to deep understanding of librarianship.

Modern librarianship—and its theoretical discipline, information science—has been called “inter-disciplinary” and indeed it is, but in two senses, quite different from each other. Librarianship is interdisciplinary in the sense that it serves a great variety of disciplines—scholarly, scientific, governmental, business. In this sense, its education must be comparably multi-disciplinary; the steps between the special librarian and the subject specialist (the “information specialist”) ought to be easy ones.

On the other hand, librarianship is also inter-disciplinary in the sense that its theoretical foundations lie in a diversity of fundamental disciplines. It is this which makes clear definition of information science so important. Mathematics, logic, linguistics, economics, psychology, engineering—each has its contribution to make in developing our theoretical understanding of the processes in communicating with large files of stored information.

Theoretical implications.—These, of course, are the substance of information science as a theoretical discipline and, as research areas, represent the greatest long-range implication of the computer. The importance of these problems must not be under-estimated nor subordinated to the pragmatic pressures of the moment, since from their solution will come any true advancement in our abilities to handle information better.

What are these problems? In a sense, half of their solution is in their very definition, so it would be foolhardy to attempt here any but the broadest characterization. But they include problems in valuation (How do we measure information and its utility? How do we measure performance?), problems in communication (How do we process natural language so as best to represent or derive information content?), and problems in system design (How do we represent the processes, both mechanical and judgmental, in the handling of information? How do we assign them and sequence them for performing specified functions?).

In summary, the implications of the computer for librarianship are far greater and of more lasting significance than simply good management or even theoretical problems in information retrieval. They are at the heart of the profession and of the society it serves.
THE COMPUTER-PRODUCED BOOK CATALOG: AN APPLICATION OF DATA PROCESSING AT MONSANTO'S INFORMATION CENTER

W. A. Wilkinson

Data processing techniques have been applied at Monsanto's Information Center for several reasons: (1) To lower operating costs, (2) To meet future growth requirements with minimum staff and budget, (3) To provide multiple copies of catalogs and other records for distribution to library users, (4) To use a systems approach in improving operations, and (5) To provide greater accuracy in all records.

The computer-produced book catalog of the Center illustrates many of these points.

A paper which appeared in Special Libraries¹ in 1963 described the semi-automated book cataloging system which Monsanto was using at that time. An efficient, successful system had been developed for producing the catalog via unit record (punched card) machines. And the catalog had proven itself to be a completely satisfactory index to the book collection.

What were Monsanto's problems? They were the inconveniences or weaknesses which are present in most unit record systems, as compared with computer systems, such as:

A. Large numbers of punched cards were handled, sorted, or filed.
B. While most of the sorting was done by machines, some hand filing was necessary.
C. Revisions to the punched card deck were time-consuming. The cards under each entry (authors, title, subjects) had to be removed from the file, revised, and replaced for any change in the body of an entry (a new edition for instance).

Why then had Monsanto started with the unit record approach and not a computer system? In the first place, a suitable computer was not available to them. Among the other reasons were:

---

¹ W. A. Wilkinson is Head, Information Center, Monsanto Company, St. Louis, Missouri.
A. Because the library users had never seen a book catalog, its acceptance was unknown; therefore, Monsanto hesitated to invest in (expensive) computer programs at first.

B. It was not known at that time if the rate of additions would be great enough to justify (monthly) computer time.

C. Because Monsanto was in the process of learning to use punched cards, we hesitated to plunge into the intricacies of computer systems right away.

D. It was believed that a good semi-automated (unit record) system could be developed so that it could be converted later to a computer system without recreating the punched card input.

Systems Study

About eighteen months after the semi-automated system started, it was decided that it was time to study the feasibility of converting to a fully automated computer system. A preliminary design for an IBM 1401 computer system was made and cost estimates were prepared to show possible savings in keypunch time, card handling, and filing.2 It was shown that sufficient savings would be obtained in these operations during the first year to pay for the cost of programming and computer time. (A total of six days per month would be saved in keypunching and filing operations.) Additional benefits which would be derived were:

A. Catalog entries could be revised more easily.

B. A shorter time would be required to produce the catalog and supplements, i.e., the catalog would always be more up-to-date.

C. The build-up of punched card files would be arrested.

D. There would be more flexibility available in the catalog format.

E. There would be greater filing accuracy via complete machine sorting.

IBM 1401 Cataloging System

The heart of Monsanto's cataloging system is the master file. This is a magnetic tape record in accession number order, consisting of one 285-position record for each book. The information on this tape might be likened to a file of unit catalog cards, in accession number order, with each card containing the descriptive cataloging
and tracings for one book. All additions, changes, and deletions in the book catalog are made via the master file.

A simplified flow chart has been prepared to show each step in the machine preparation of the book catalog (see Fig. 1). Two permanent tape records are maintained: the master file and the headings file. Content of the master file was explained above. The headings

![Figure 1](image)

file is a record of all subject headings and cross references which have been used in the catalog. The master and headings files are brought up-to-date by processing new punched cards through the IBM 1401. Then a new, up-to-date catalog is created by extracting information from the master file and headings file to print author, title, and subject catalogs.

A pre-printed IBM card was designed to accept all punching for additions, changes, or deletions in the master and headings files (see Fig. 2). Three types of cards provide input to the master file: (1) A "1" card carries the call number and author information, (2) A "2" card carries the title information (title, edition, volume(s), publisher, date, and series note), and (3) A "3" card carries the subject and series tracings and location codes. In Figure 2, note the numbers 3, 2, and 1 in the right hand margin of the card. Reading across the card at each level, you can see the information that each different type of card contains. Common to each type of card is the information found in columns 1-13;
Figure 2

Columns 1-2  “action”  
51 = addition  
52 = change  
53 = deletion  

Columns 3-8  accession number  
Column 9  card type  
1 = author(s)  
2 = title, etc.  
3 = subjects, location  

Columns 10-13  card count  
Cols. 10-11 = total number of cards of a single type  
Cols. 12-13 = card number within total  

The “action” code, which appears in columns 1 and 2, indicates whether the information in the card should be processed as an addition, change (replacement), or deletion. Any of these actions can be carried out selectively in the three different types of cards. For instance, a change in subject headings can be made by punching the revised subject tracings in a “3” card with the “52” (change) code punched in columns 1 and 2. There is no need in this case to resubmit the author and title information. This change feature is especially helpful when a second copy of a book already cataloged is added at another location. In this case, a “3” card is punched with the revised location code and processed into the master tape. Type “4” cards and type “5” cards provide input to the headings file. The former are used for subject headings, the latter for “see” or “see also” references.  

A complete set of punched cards for one book is shown in Figure 3.
## SET OF PUNCHED CARDS

FOR ONE BOOK

<table>
<thead>
<tr>
<th>ACTION</th>
<th>DEWEY DECIMAL</th>
<th>AUTHOR CODE</th>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>SET</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>OF</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>PUNCHED</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>CARDS</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>FOR</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>ONE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACTION</th>
<th>DEWEY DECIMAL</th>
<th>AUTHOR CODE</th>
<th>TITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>SET</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>OF</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>PUNCHED</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>CARDS</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>FOR</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>ONE</td>
</tr>
</tbody>
</table>

---

### Figure 3
Sample author, title, and subject catalog pages are shown in Figures 4, 5, and 6. Based on our experience with the book catalog, we have made several changes in the over-all page format. One

SAMPLE AUTHOR CATALOG PAGE

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Publisher, Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conway HM</td>
<td>WEATHER HANDBOOK</td>
<td>Conway pub 1963</td>
</tr>
<tr>
<td>Cooke NM &amp; Markus J</td>
<td>ELECTRONICS &amp; NUCLEONICS DICTIONARY</td>
<td>McGraw Hill 1960</td>
</tr>
<tr>
<td>Coolidge JL</td>
<td>INTRODUCTION TO MATHEMATICAL PROBABILITY</td>
<td>Dover pub 1962</td>
</tr>
<tr>
<td>Cook MS</td>
<td>CONSTRUCTION ACCOUNTING &amp; FINANCIAL MANAGEMENT</td>
<td>Fw Dodge 1958</td>
</tr>
<tr>
<td>Cooper JD</td>
<td>HOW TO COMMUNICATE POLICIES &amp; INSTRUCTIONS</td>
<td>BNA 1960</td>
</tr>
<tr>
<td>Coppack JD</td>
<td>NORTH ATLANTIC POLICY THE AGRICULTURAL GAP</td>
<td>Twent cent fund 1963</td>
</tr>
<tr>
<td>Conson DA</td>
<td>MICROWAVE HEATING</td>
<td>Avi pub 1962</td>
</tr>
<tr>
<td>Cooke HR &amp; Laque FL</td>
<td>CORROSION RESISTANCE OF METALS &amp; ALLOYS</td>
<td>2 ed Reinhold 1963/acs monograph 158/</td>
</tr>
<tr>
<td>Corey ER</td>
<td>INDUSTRIAL MARKETING PRENTICE HALL 1962</td>
<td></td>
</tr>
<tr>
<td>Cotton FA</td>
<td>CHEMICAL APPLICATIONS OF GROUP THEORY INTERSCIENCE 1963</td>
<td></td>
</tr>
<tr>
<td>Cox EB</td>
<td>TRENDS IN THE DISTRIBUTION OF STOCK OWNERSHIP</td>
<td>Pennsylvania U 1963</td>
</tr>
<tr>
<td>Crisp RD</td>
<td>MARKETING RESEARCH</td>
<td>McGraw Hill 1957</td>
</tr>
<tr>
<td>Crisp RD</td>
<td>SALES PLANNING &amp; CONTROL</td>
<td>McGraw Hill 1961</td>
</tr>
<tr>
<td>Crossfield LTD</td>
<td>CAUSTIC SODA &amp; CHLORINE IN THE SOVIET UNION</td>
<td>Crossfield 1959/east european chem ind 2/</td>
</tr>
<tr>
<td>Crossfield LTD</td>
<td>COST &amp; PRODUCT DISTRIBUTION IN THE HUNGARIAN CHEMICAL INDUSTRY</td>
<td>Crossfield 1962/east european chem ind 8/</td>
</tr>
<tr>
<td>Crossfield Ltd</td>
<td>EASTERN GERMANY CROSSFIELD 1959/EAST EUROPEAN CHEM IND 3/</td>
<td></td>
</tr>
<tr>
<td>Crossfield Ltd</td>
<td>HUNGARY CROSSFIELD 1958/EAST EUROPEAN CHEM IND 1/</td>
<td></td>
</tr>
<tr>
<td>Crossfield Ltd</td>
<td>POLANDS TRADE IN CHEMICALS 1958 CROSSFIELD 1963/EAST EUROPEAN CHEM IND 9/</td>
<td></td>
</tr>
<tr>
<td>Crossfield LTD</td>
<td>SOVIET UNION CHEMICAL EXPORTS 1955-1959 CROSSFIELD 1960/EAST EUROPEAN CHEM IND 4/</td>
<td></td>
</tr>
<tr>
<td>Crossfield Ltd</td>
<td>SOVIET UNION CHEMICAL IMPORTS 1955-1959 CROSSFIELD 1961/EAST EUROPEAN CHEM IND 5/</td>
<td></td>
</tr>
<tr>
<td>Crossfield LTD</td>
<td>SOVIET UNION CHEMICAL TRADE 1959-1960 CROSSFIELD 1962/EAST EUROPEAN CHEM IND 7/</td>
<td></td>
</tr>
<tr>
<td>Crossfield Ltd</td>
<td>TECHNICAL PROGRESS &amp; ECONOMICS IN THE SOVIET NITROGEN INDUSTRY CROSSFIELD 1961/EAST EUROPEAN CHEM IND 6/</td>
<td></td>
</tr>
<tr>
<td>Cross PC &amp; Allen HC</td>
<td>MOLECULAR VIBRATORS</td>
<td>Wiley 1963</td>
</tr>
<tr>
<td>Crosswell CM</td>
<td>INTERNATIONAL BUSINESS TECHNIQUES LEGAL &amp; FINANCIAL ASPECTS</td>
<td>Oceana pub 1963</td>
</tr>
</tbody>
</table>

Figure 4
seemingly minor change was to move the page number from the bottom to the top of the page. This small change resulted in an increase of several lines of print per page and an over-all reduction of almost 10 per cent in the total size of the catalog. A limitation in the automatic page numbering routine while printing numbers at the bottom of the page had caused the short pages. Incidentally, you will note that each section of the catalog carries a prefix in the page number, "a" for author, etc. This feature was added after pages 35
and 36 of the author and title catalogs had been inadvertently interchanged in the first edition by the bindery.

In earlier editions of the catalog, it was felt that it would be wise to approximate card catalog format for the convenience of library users. For this reason the call number had always been
placed on the left, beginning on the first line of each entry. This is no longer the case. There are two reasons for the change. First, it was believed that it would be logical for the first word in each entry to be the filing word, and all other information would follow.\textsuperscript{3} As you can see, this is now the case in each part of the catalog. Second, there are two pieces of information in each entry which together tell where the book is shelved: the call number and the location (library branch) code. This information was separated when the call number was at the beginning of the entry, but now can be found in one area at the right. Monsanto finds that it is a good reminder to catalog-users that it is a union catalog and that they must note both call number and location.

Several features which have been programmed to appear automatically in each entry even though they are not punched into the input cards are:

A. Asterisks are inserted at both ends of each subject heading to make the heading stand out better on the page.
B. Joint authors are punched with two blank columns between them in the "1" card (see Fig. 3). The program causes the authors to appear once in this order and once in reverse order as two separate entries in the catalog. Also, during the print program an ampersand is inserted between them.
C. No decimal is punched in the classification number. It is inserted automatically during the printing step.

**Schedule of Operation**

A completely revised catalog is produced yearly. Cumulative supplements are issued every two months. A subject listing of new books is issued every month. During the month, a card file is maintained in the library to locate cataloged books not yet listed in the catalog or supplements. This schedule is flexible, so that revisions or supplements can be made more or less often depending on the need. It is felt that the present schedule is quite satisfactory.

During the month, as new books are cataloged, punched cards are prepared for the monthly run. On the eighteenth working day of each month (generally about the twenty-fifth day of the month) all additions, changes, and deletions cards are processed against the master and heading files (tapes). Then all new records are selected to produce the listing of new books, which is distributed widely as Monsanto's monthly Library Bulletin at the end of the month. Selection of records from the master tape is controlled by "keys" in the master record for each book, one for the monthly new book listing and one for the year-to-date supplement. The keys are erased after
the new book listing and the final year-to-date supplement are issued.

After the listing of new books has been made, the year-to-date supplement is printed, during alternate months. In the twelfth month, a complete revision of the catalog is prepared, instead of a year-to-date supplement. Provision has been made via a control card, for adding older books to the master file without selecting them for the new book listing.

Copies of the monthly Library Bulletin are distributed to about 500 individuals, departments, and libraries within Monsanto. About seventy-five copies of the book catalog (and supplements) are distributed to libraries, departments, laboratories, and some individuals. Those who have the catalog keep a copy of the Library Bulletin for reference during the alternate months when no supplements are issued.

Conversion to the Computer System

As part of the systems evaluation study, consideration was given to the conversion of existing punched card records into a format acceptable to the computer system. If it had been necessary to re-punch the records for the 7,000 volumes already cataloged, justification of the change would have been more difficult. Programming for the conversion turned out to be almost as difficult as writing the operating programs.

Monsanto's problems resulted from devices which had been used in programming efficiently for unit record equipment, especially the Document Writer (IBM 870 Document Writing System). The most serious of these was a lack of complete card control in the existing punched cards. A "1" punch in column 1 of the first card in each set of cards and a "2" punch in column 1 of all other cards in the set for each book had been used. This had been done because of the very limited ability of the Document Writer to recognize controls. If one thing was learned from the whole project, it was the importance of adequate card control.

Another problem encountered was the elimination of card control characters which had been punched into the original cards to control printing on the Document Writer. For instance, a non-printing % symbol had been punched at the end of the title to cause a carriage return. These special characters were not needed in the new system, and had to be removed. They were removed during an editing and move-up step in the conversion program.

Many other problems were solved during the conversion either by programmed routines or by error messages. Where very complicated programming and/or an unreasonable amount of machine
time would have been necessary to correct problems in a small number of entries, provision was made to recognize the problems and print messages to say where they were. Then corrections were made later to the master file by the normal change card routine.

Future Plans

Plans were made several months ago to integrate the cataloging system back to the purchasing step. A flow chart was developed and a five-part purchase order was designed and ordered in cooperation with Washington University School of Medicine Library, St. Louis, Mo. (see Fig. 7). The forms have been received now and a board has been wired for the Document Writer. As this paper is being prepared, Monsanto is preparing to start the purchase order routine on an experimental basis.

Under the new system for writing purchase orders, punched cards are created as the first step in ordering, instead of typing purchase orders. The cards are punched in the format necessary
for the cataloging system with the exception that at this stage the call number, accession number, and subject tracings are not yet available. An additional card is punched with all the information specific to the purchase order (vendor, order number, number of copies, etc.). By feeding the cards to the Document Writer, purchase orders are written with copies for vendor, requester, order record, follow-up, and a cataloger's work copy. The latter is on card stock and eventually becomes the shelf list card.

After the book has been received, the cataloger verifies the information already printed on the work copy (author, title, etc.) and adds the call number and subject tracings. This added information is keypunched into the original cards, the accession number added, and the cards are ready for addition to the catalog. While this system looks good on paper (eliminates two typing steps), it remains to be proven in actual use. Other projects planned or under way include:

A. The addition of the complete holdings of three branch libraries to the catalog. (At present, only books added to the branches since 1961 are included.)
B. Editing certain subject headings, "see" and "see also" references and abbreviations of corporate authors for more consistency.
C. Optimizing publication schedule and methods of printing and binding to suit needs.

Conclusions

Although most of the lessons that Monsanto has learned have already been indicated, some deserve another mention in closing:

A. Use fixed fields in the card format if at all possible.
B. Provide adequate card control by card identification and card count.
C. Data control is extremely important. Many hours of programming or machine time can be wasted by careless errors in input data.
D. These problems can be reduced by programmed checks of input data with error messages when appropriate.
E. Use a "systems" approach; do not just automate existing methods.
F. Know your costs. It would be bad to build a heavily automated system on a weak cost structure, subject to withdrawal later when costs are re-examined.

And lastly, plan your system with both an immediate and a long-range goal. It is not possible to wait for the ultimate system; one area should be isolated and worked on at a time. When all the
problems in that area have been solved, move on to another, always keeping the long-range goal in mind. In that way, benefits of improved methods are obtained all along the way, disruptions in library operations are minimized, and encouragement will be gained from each success.

REFERENCES


2. All of the computer programs for Monsanto's Information Center were written by W. A. Shank of Monsanto's Business Systems Department. Without his courage, patience, and thorough knowledge of programming, "3" by "5" cards might still be typed.

3. Also suggested by F. W. Holzbaur, IBM Data Systems Division. Personal communication, July 16, 1963.

Additional References


2. KWIC Indexes. Using the KWIC (Keyword-In-Context) indexing technique, an index to Monsanto Marketing Reports has been prepared and revised. Other KWIC indexes are in preparation.

3. Monsanto List of Serials. This publication is in its fourth edition (since 1958) and includes the holdings of nine Monsanto libraries.

4. Subscription and Standing Orders List. A punched card record for each title includes information such as expiration data, supplier, cost, frequency, where shelved, how checked in, retention, whether cataloged, etc. Renewal lists, budgets, expiration check lists, check in records, etc. are prepared from these cards.
Appendix

Since accurate cost information was not available when the manuscript was written none was included in the paper. The following costs have been gathered and are now added to make the paper more complete.

Annual Cost of Book Catalog System*

1. Amount of IBM 1401 Computer Time Used:

<table>
<thead>
<tr>
<th>Service</th>
<th>Hours/Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Library Bulletin</td>
<td>6</td>
</tr>
<tr>
<td>Catalog Supplements</td>
<td>5</td>
</tr>
<tr>
<td>Annual Catalog Revision</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
</tr>
</tbody>
</table>

@ $50.00/hr.

Annual Cost of Computer Time = $750.00

2. Other Cost:

<table>
<thead>
<tr>
<th>Service</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keypunching Time</td>
<td>$500.00</td>
</tr>
<tr>
<td>Keypunch Rental</td>
<td>$600.00</td>
</tr>
</tbody>
</table>

3. Total Yearly Cost $1850.00

*Provides monthly Library Bulletin, bimonthly cumulative supplements to catalog, and annual complete revision of three-part book catalog. Does not include cataloger's time. Current rate of new additions about 1,500 titles per year.

Discussion

W. A. Kozumplik*

Cost consideration, it is my belief, determines utilization of machines for library operations. The costs we have in mind are concerned with labor, equipment, and space. Of these, labor is by far the most critical, which is the reason mechanization efforts have been so widely applied in our civilization.

*W. A. Kozumplik is Manager, Technical Information Center, Lockheed Missiles and Space Company, Palo Alto, California.
When ways in which to mechanize library operations are considered, there is no doubt that it is the cataloging product which, more than any other operation, holds exceptional promise for cost-reduction. This has been Monsanto's experience. Of all library operations, Monsanto placed initial focus on cataloging, conquered the problem in two phases, which led it from a semi-automatic to a fully automated product, and then set sights on further areas to mechanize. We all look forward to knowing of Monsanto's further experience on the latter in terms of cost and effectiveness of product.

With respect to mechanizing the cataloging product, some libraries have employed the computer to deliver catalog cards as the product. In so doing, 30 per cent savings were achieved over the best available manual method of catalog card producing (utilizing the electronic typewriter).

Monsanto eschewed this step, going directly from catalog card to printed page. When it did this, it achieved additional savings in equipment and space; these are not identified by William A. Wilkinson. Considering equipment alone, savings in the order of 30 to 1 are effected when one supplants card-catalog cases by shelves—even wood shelving, which is double the cost of metal shelving. Savings in space are not so spectacular, being only in the order of 3 to 1. (For a fuller treatment of such comparative costs, one may read the article by Fred Heinritz, which is a condensation of his doctoral dissertation submitted in 1963 to Rutgers University.) To summarize: respecting only equipment and space, the codex catalog is immeasurably less expensive than the card catalog.

You will recall that in 1963 Wilkinson reported Monsanto's comparative costs of manually produced catalog cards versus machine production of the codex catalog. It is my belief that cost considerations were again the dominant determinant in Monsanto's decision to convert its codex catalog production from a semi-automatic system to one fully automated (computer based). Then late in 1963, cost estimates showed Monsanto that possible savings in keypunch time, card handling, and card filing would, in the first year of operation alone, more than pay the programming costs. While Wilkinson does not state it, my conjecture would be that Monsanto was also prompted by two other considerations in deciding to effect this conversion, namely, the potential spin-offs that were so highly desirable and attainable at no extra cost (see Wilkinson's items D and E under "Systems Study") and the attainable improvements in existing products (see Wilkinson's items A, B, and C under "Systems Study").

Cost, not concern for or interest in the reaction of the scientist or engineer to use of the codex catalog, determined the institution and refinement of machine methods at Monsanto. It would be interesting to know how Monsanto's users of technical information reacted
to the codex form. A few undoubtedly grumbled over the change. What we do in such cases is to take these individuals aside and tell them that the codex catalog is what they, as traditionalists, really should be fighting for, not against, because a codex represents a return to the state of affairs before Melvil Dewey. It was his card catalog, you remember, which supplanted the codex catalog in the fourth quarter of the nineteenth century. If there is any doubt that this is not a precious example of the concept "coming full circle," let me remind you that the card catalog was instituted for reasons of economy. Librarians were definitely cost conscious in those days. And resting on our laurels, we found ourselves complacently asleep, from which sleep outsiders chiefly have been trying to arouse us, or at least they have been making the most noise. Clinics like this attest in part to the fact that our profession is indeed aroused and is forging ahead in the role of leadership.

Were we to pursue considerations of cost to their logical end, we should expect that Monsanto would be thinking about taking another step in utilizing computers for its technical information center operations, namely, the storage and retrieval of bibliographic retrieval points in depth. Monsanto may have already thought along these lines and may have discarded the challenge on a cost basis, possibly because of current and forecast low-volume use. In any event, it would appear wise to wait, before any serious, final independent attempt is made along these lines, until the Library of Congress reaches a decision to automate its operations and writes system and hardware specifications. While not expecting to be fully operational until 1972, the Library of Congress system will set the national pattern towards automating research libraries for generations to come. It appears rather mandatory, therefore, for the smaller research library—and this covers about all industrial libraries and all but a handful of university and college libraries—to reconsider expending dollars for systems and hardware that would automate resources, services, and operations. It appears clear that existing programs must be compatible with the system evolved by the Library of Congress if the vast potential for effective utilization of existing national resources (interlibrary cooperation) is to be realized.

In our society, scientific and technical writing constitute a national resource; this resource only becomes effective when it is placed under effective bibliographic control. In the area of scientific and technical disciplines, printed contributions are proliferating at the rate of one magnitude each fifty years.

In fiscal 1963, the Federal Government spent $15 billion for research and development (R&D). The National Science Foundation has reported that the generating federal agencies in 1963 expended $1.5 billion on STINFO (scientific and technical information). So that you are not misled, let me point out that these STINFO dollars
include expenditures for four services or capabilities, namely, (1) publication costs—editing, art work, typing, printing, distribution; (2) travel costs—attending society meetings and sponsoring symposia or clinics (like this one) in order to "acquire" information; (3) library costs—procurement and organization of recorded knowledge, circulation, reference, and literature search services; and (4) computer or data processing costs—development and production of mechanized capability to store and retrieve information rapidly and reliably.

Obviously, there is a need to control literature. Effective bibliographic control, together with timely availability of the literature, should prevent repeated reinvention of the wheel. It was over six years ago, you remember, that L. H. Flett produced the challenging statistics in Information Resources—A Challenge to American Science and Technology that 45 per cent of the R&D expenditure is wasted. Flett's reason is that recorded knowledge was not effectively utilized. I have not personally checked these findings, but if Flett's figures are any indication of the magnitude of the problem, it would appear that several billion dollars are going down the drain annually.

Bibliographic control costs money; such costs will be astronomical by 1975 should we continue to use the traditional methods of cataloging, indexing, storing, and retrieving. These costs would even be excessive today if it were not for our practice of discarding certain works, thus exercising judgment not to catalog for admittedly arbitrary reasons; one overworked reason which you will easily recognize rests on format, particularly that associated with the concept of ephemera.

We just cannot afford to go the route—the rut—of tradition. In a very few years, the cost problem will have been pre-empted by the bigger problem of the chaotic, accelerating, and inundating, "publish-or-perish" paper storm, wherein backlogs of uncataloged (bibliographically unorganized) materials will mount. Recorded knowledge could not possibly be put to effective use; unwanted duplication will abound. It is in such an environment of rising costs and mounting backlogs that computer technology thrives. In an automated system, backlogs normally do not accrue, and the items, as well as their contents, will be under excellent bibliographic control at a cost per title much less than what can be achieved through traditional systems. In addition, computer technology insists on a systems approach which inevitably identifies other technical information center operations that are amenable to mechanization. It has already happened in this fashion at Monsanto, where the semi-automated codex catalog of 1962 has been programmed for fully automated (computer) production in 1963. It is the system analysis approach which quite likely produced Monsanto's decision "to integrate the cataloging system back to the purchasing step," according to Wilkinson. The fact
that the new purchase routine is on a semi-automated basis should not belittle the efficacy of the systems analysis approach. The plain fact is that certain library operations will continue to be accomplished more economically by manual or semi-automated methods; not all stand the cost-test for going fully automatic.

Of the latter variety, two functions and their computer application come to mind: (1) who needs what, that is, the selective dissemination of information (SDI) based on up-to-date user-interest profiles; and (2) deeper and broader identification of contents, that is, a program to store and to retrieve information to a high degree of specificity, almost as though we would be indexing and not cataloging. I should like to hazard the guess that the reason Wilkinson did not mention those two programs was because of their cost, due in part to the size of the collections and to the volume and kind of use which did not warrant deeper specificity and more rapid retrieval at this time.

If this is the case, we once again note that cost rules. But we also note that user needs appear to be receiving more serious consideration.

REFERENCES


DEVELOPMENT OF COMPUTERIZATION OF CARD CATALOGS IN MEDICAL AND SCIENTIFIC LIBRARIES*

Frederick G. Kilgour

Various scientific libraries are computerizing their card catalogs; some produce catalog cards and others have gone to book catalogs. At least one library associated with the military establishment is developing an information retrieval system for catalog card information employing a large, high-speed computer. Still, relatively little work is being done on computerizing the retrieval of catalog and index information—well-known projects being the Medical Literature Analysis and Retrieval System (MEDLARS) at the National Library of Medicine and the information retrieval system of the American Society of Metals. Both of these systems employ sequential searching of magnetic tape. However, this paper will not attempt to survey these burgeoning activities completely but will report only on the Columbia-Harvard-Yale Medical Libraries Computerization Project.

It is in searching the file that the Columbia-Harvard-Yale Project differs from most others. It is intended that the file will be in a random access memory device and will be on-line for each of the libraries. One of the specifications for the design of the system is that the answer to an inquiry should begin to come out of the system within a minute or two after the question has been inserted. Furthermore, it is hoped that articles indexed by MEDLARS in some 260 journals supplying upwards of 75 per cent of recorded use in the three libraries will be included in the computer file. However, this paper will be confined largely to the discussion of the computerization of book cataloging.

Frederick G. Kilgour is Librarian, Yale Medical Library, Yale University, New Haven, Connecticut. The author is grateful and indebted to his colleagues, Thomas P. Fleming, Librarian, Columbia Medical Library, and Ralph T. Esterquest, Librarian, Harvard Medical Library, for making possible and directing with imagination and wisdom the Project described in this report.

*A national Science Foundation grant supports the work reported in this paper.
The Project is based in large part on the fact that there exists in scientific and medical libraries a relatively small core of the book collection that supplies upwards of three-quarters of the use of the library. The Yale Medical Library possesses some 350,000 items of which more than 110,000 are European theses. Still, a study of recorded book usage in the Yale Medical Library done three years ago showed that books published in the previous twelve years, or about 10,000 volumes, furnished 79 per cent of the recorded use. More recently, a study at Columbia and Yale has shown that of the 2,000 journals then being received at Columbia and of the 1,500 at Yale, 262 furnished 80 per cent of the recorded use of recently published journals. It is these heavily used cores that make the on-line computerization of catalogs and indexes economically feasible. The Columbia-Harvard-Yale Project began to get under way in the autumn of 1961. Important for its initiation and present prosecution was a suggestion made by a group at ITEK Corp. whose most prominent members were Lawrence Buckland, Ben-Ami Lipetz, and David Sparks. This group pointed out that it had become possible to produce cataloging information in machine-readable form that could be used to continue the production of card catalogs and could be accumulated over several years to be put in a computer file when sufficient cataloging information was available to justify using a computer. Columbia, Harvard, and Yale drew up a request for a grant from the National Science Foundation (NSF) late in the summer of 1962 and revised this request at the end of 1962. The purpose of the grant was to finance the development and initiation of a computerized library catalog system. NSF made the award in the summer of 1963. However, the new procedures had been initiated in the late winter of 1962/63 so that at Yale all books possessing an imprint of 1963 and later have been processed in the new procedures.

The goal of the system is to increase the speed and completeness with which a user obtains catalog and index information in a library. The Project is attempting to play a significant role in the development of computerized catalogs which will undoubtedly be the next major step towards increased speed and completeness of library services following the nineteenth-century introduction of the card catalog, and the abstract and index journals. The purpose of the Project is to design a computerized catalog system and to demonstrate the feasibility of such a system. In addition, it is anticipated that the system, including computer programs, can be taken over and used by a majority of conventional libraries.

As far as library-like activities are concerned, it is convenient to think of information retrieval from a large corpus of information as being in three categories. In one category is the library supplying heavily and moderately used information rapidly—often within a few minutes. Next there is the bibliographical variety furnishing
relatively little-used information slowly—perhaps in a day or two. An example would be the MEDLARS project at the National Library of Medicine. Finally, there is the documentation category of information retrieval wherein specific, detailed data are furnished, usually after a period of time which may amount to a week; the documentation project of the American Society of Metals is an example of this category. The Columbia-Harvard-Yale Project is in the first group.

**Information Retrieval System**

The main goal of the Columbia-Harvard-Yale Project is information retrieval—the rapid and complete retrieval of cataloging and indexing information. As already mentioned, only that cataloging and indexing which relates to the relatively small core collection in medical libraries will be computerized. Studies done at Yale indicate that books supply approximately 40 per cent of recorded usage, while journals furnish nearly 60 per cent. It will be the cataloging of that part of the book collection supplying upwards of 75 per cent of use which will be computerized. The activation of the information retrieval system cannot occur before 1966, but it is hoped that it will begin operation in that calendar year.

At least a half-dozen significant achievements are expected of the information retrieval system. As already mentioned, it will increase and make more complete the supplying of cataloging and indexing information as compared to the present tedious card-by-card search. The primary approach to the catalog in the computer file will be by subject, but any given subject can be coordinated with perhaps up to four more subjects, with the date of publication, the language, and the place of publication. An example might be a search for books discussing the use of computers for information retrieval in science and published in English after 1962. In a card catalog there probably would not be an excessive accumulation of entries under any of the headings equivalent to these subjects. However, a search on the relationship between cancer and enzymes in the card catalog of the Yale Medical Library would involve going through 600 cards under cancer and 100 under enzymes. Another important advantage of the information retrieval system involving the catalogs of three libraries is that those three catalogs will be searched as one for users in each library. Since 55 per cent of book holdings are in but one or two of the three libraries, users in each library will enjoy increased access to literature, albeit that some would not be available for a day or two. In addition to the coordinated subject searches, there will also be an increased depth of subject cataloging—a third benefit. As is well-known, libraries now keep to a minimum the number of subject
headings for each title in order to slow the engulfing growth of the card catalog. At the Columbia and Yale medical libraries, subject headings per title average 1.7;\textsuperscript{4} at Harvard, 1.8;\textsuperscript{4} and at the Library of Congress, 1.6.\textsuperscript{5} Since this need for repression will disappear with the advent of the catalog in a computer file, it will be possible to do more adequate subject analysis in depth. During the early months of the application of the new procedures in the Yale Medical Library, the number of subject headings assigned to each book rose from 1.6 to 3.2. At the present time the figure is higher, with the goal being an average of five subject headings per title.

A study of the use of the subject cards of the catalog of the Yale Medical Library was carried out last autumn and yielded results\textsuperscript{6} which seem to indicate that present-day subject cataloging is somewhat less than adequate to meet the increasing demands for information. The subject cards were used but 12 per cent of the time when catalog use by the technical staff of the library was included. When such utilization was excluded, use of the subject catalog was 18 per cent.\textsuperscript{7} These low percentages indicate that the present subject card catalog is not the favored tool of users. However, it is hoped that with speeded access to the subject catalog and with a greater depth of subject cataloging this tool will increase in usefulness.

Greater completeness in catalog searching can be assured because it will be possible to teach the computer always to search “see also references.” It is impossible to teach users to search “see also references,” and they thereby miss useful titles at times.

Another achievement of the information retrieval system will be a relative ease in printing out the catalog in book form. Certainly author and title catalogs will be produced in book form, and it would be equally possible to print subject catalogs. Copies of these book catalogs could be placed outside the library in various laboratories, thereby increasing the availability of materials in the library. Moreover, book catalogs are far easier to use than card catalogs and stimulate catalog “browsing.” Finally, a computerized information retrieval system will also provide for selective periodic dissemination of new cataloging information. “Current awareness listings” of certain subjects could be furnished on a periodic basis. Indeed, the first product—actually a by-product—of the Project was the mechanized production of the monthly Bulletin of the Yale Medical Library which lists accessions for the previous month.

There are also several administrative benefits which will accrue from a computerized catalog system housing the catalogs of two or more libraries. Cataloging expense will be reduced, or amount of cataloging increased, because such a computerized catalog is in effect a union catalog. For instance, an investigation of duplication amongst the Columbia, Harvard, and Yale medical libraries revealed that 66, 84, and 83 per cent respectively of each collection
is in one or both of the other two collections. If acquisitions of books were completely at random in the three libraries, each library would need to catalog only half the percentage of its collection duplicated in one or both of the other two libraries. It therefore appears reasonable that perhaps one-third or one-quarter of present cataloging costs will be eliminated. Of course, the more libraries that participate in one system the greater will be the decrease in cataloging costs providing all libraries acquire works at approximately the same rate and in the same subjects. Another administrative benefit is that such mechanized procedures are faster and more accurate. Once a decklet of punched cards containing cataloging information about a title has been produced and verified, it can be employed for each subsequent activity rapidly and without need for repeated proofreading. Another administrative advantage, albeit perhaps not a major one, is that the ballooning card catalog could be housed elsewhere than in a library's busiest and most desirable location. A computerized catalog will be in the computer file, presumably in a computer room outside the library and enormously reduced in size. Catalogs in book form could be widely available, but as is well-known, book catalogs occupy far less space than card catalogs.

An on-line computer catalog is made possible by locating in each library an information station possessing a telecommunication connection with the computer. It is likely that the computer will be located in New Haven, but actually the Yale Medical Library will be no closer to it electronically, than the Columbia or Harvard medical libraries. Present specifications for the information retrieval system include a time lapse of no greater than one minute at any information station—no matter how remote—between the end of the inquiry going into the computer and the start of the reply from the computer. These information stations will probably consist of an electric typewriter and a card reader. The typewriter will be used to put questions to the computer and will type out the replies in the form of catalog references and will include call numbers. The card reader will process cards to add to the computerized catalog.

A second basic technical feature of the system will be a random access memory unit that will hold the catalog files. Information sought in such files can be found in a fraction of a second. There has not yet been agreement upon the specifications for the arrangement of information in such a file, but one possible configuration would be to have the equivalent of catalog cards in one section of the file. A second section would have the addresses of subjects, and under each subject would be the number of the catalog card on which that heading occurred. The analogy to a card catalog would be an author catalog not filed by subject but with each card having, perhaps, a sequential identification number. Each identification number would then be listed on a subject card which would be filed by a coded, numbered
address in a different drawer. In a computer when a work on three different subjects is sought, the numbers under those three subjects would be brought from the random access files into the core of the computer where the numbers would be compared. Each number that occurred under all three subjects would then be brought from the equivalent of the card file and dispatched electronically to the remote information station.

Mechanized Catalog Card Production

Although the information retrieval system is in the design period, the mechanized production of catalog cards is in the early stages of operations. The procedure starts with the cataloging of a book on a 8-1/2" x 11" worksheet (see Fig. 1). The cataloger makes a format of the card on the worksheet and the keypunch operator punches one card for each line on the sheet. The resulting decklet of cards, with many other decklets, is then fed into a computer behind a program which expands the decklet into the number of cards required and puts these cards on magnetic tape. After a computer
has sorted these records by filing entry, they are punched out on punched cards. These punched cards drive an electric typewriter which produces the cards in their final form and in alphabetical order ready for filing in the catalog.

The computer programs, except for the sorting program, are designed for an IBM 1401, 4K core, 2-tape drive computer. The census of computers in the March 1964 Computers in Automation showed that 37 per cent of all computer installations were 1401's. The nearest competitor was the IBM 1620, but the 1401 had nearly five times as many installations as the 1620, a small scientific computer not particularly suitable for non-numerical data processing. Clearly, the 1401 is the most widely available computer and because of this fact the Columbia-Harvard-Yale Project programs have been written for this computer so that the programs would be of the widest possible use to other libraries.

In the cataloging procedure the only change from the operations in many libraries is the use of a worksheet (see Fig. 1). The worksheet shown is one designed for a cataloger who prefers to write out the catalog card. At Columbia a worksheet which can be used in the typewriter is employed with success. At Yale the catalogers find that the worksheet enables them often to establish the main entry in its final form at the catalog, thereby eliminating some recopying. The format of the catalog card on the worksheet mimics almost exactly Library of Congress format, the only difference being that topical subject headings, name subject headings, and other tracings each begin on a separate line.

Certain positions on the worksheet must be precise and flags or signals to the computer programs must be placed in a half-dozen locations on each sheet. The “Directions for Use of the Cataloging Worksheet” occupy about 5-1/2 typewritten pages; in other words, they are not tremendously long and complicated. Nevertheless, the location of the call number and of the beginning of the main entry, the title, the collation, and other groups of information on the card are precisely defined, as are the locations of the flags. For instance, if a short title tracing is employed, a delta must be inserted in the space preceding the first letter of the first word of the short title, and similarly, in the space following the last letter or punctuation mark. A “less-than” sign is always placed at the end of the title or the title added entry and a “greater-than” sign before the imprint.

The worksheet in Figure 1 makes it possible to use one set of subject headings on the printed catalog cards and another set for the information retrieval computer. As an example, the Yale Medical Library is continuing to use Library of Congress subject headings in its card catalog, but of course employs Medical Subject Headings (MeSH) for the information retrieval computer. The MeSH are written on the lines at the bottom of the worksheet. The
programs which produce the catalog cards disregard these headings.

The completed worksheet goes to a keypunch operator who prepares a punched card for each line on the worksheet with the exception of the call number. The call number is punched on the first card of the decklet with the delta separating each line of the number. At the Yale Medical Library the same person keypunches who formerly stenciled cards, and the cards are punched more rapidly than they could be stenciled.

Next, a group of punched card decklets are fed into a 1401 computer behind a program having a card punched to set up the number of packs of catalog cards which will be needed. For instance, the Yale Medical Library needs to have one pack of main entry cards to go to the National Union Catalog, two packs of main entries for the Yale University Library, one pack including main entry and all subject and added entries, and finally two packs of shelf list cards—one for the library shelf list and the other for insurance purposes. This first program expands each decklet into the number of catalog cards required to make up the group of packs and writes the data for each catalog card on magnetic tape. A second program prepares the data on tape for sorting and the sort is carried out on an IBM 709. The 709 produces a magnetic tape with catalog card data alphabetized by the filing entry within each pack.

The magnetic tape produced by the 709 is then put on a tape drive of the 1401 and manipulated by a program which, like the first, has a control card. This card can be punched to determine the format of each of four types of headings: (1) topical subject headings, (2) name subject headings, (3) title, short title, and series added entries, and (4) other added entries. The control card can be punched so that any one or all of the headings will appear at the top or bottom of the card, at the left-most position, first indention or second indention, in upper and lower case or all upper case, and in black or red. Since the required heading is printed in the proper location on each card, tracings appear only on the shelf list card. The data are read from the magnetic tape, each line of the card formatted in the computer, and the characters recorded on punched cards.

The punched cards produced by the third 1401 program are then placed in an IBM 870 Document Writer. This contrivance is basically a card punch electronically coupled with an electric typewriter. The typewriter which the system uses has eighty-eight characters including sufficient diacritical marks to enable the system to handle twelve different languages in entirety. The cards punched by the 1401 are placed in the card feed of the 870’s keypunch whence they travel through the reading head on the keypunch. Information read off is communicated to the typewriter which types out the cards on a continuous card form which was originally designed by Phillip Bagley of the Mitre Corporation working with the Dennison
Manufacturing Co. of Framingham, Massachusetts. As the cards come out of the Document Writer, they are in final form and in alphabetical order for filing.

The Project has on order together with Florida Atlantic University and the University of Toronto Library an upper and lower case printing chain for the IBM 1403 printer, the printer in the 1401 configuration. The three institutions are sharing the cost of developing special characters for such a chain and each is acquiring its own chain. When the chain is available, the third 1401 program will be altered so that the catalog cards will be printed out directly and much more rapidly on the computer. However, it will not be possible to have red headings, but the catalog cards will be produced at the rate of perhaps thirty per minute instead of at the rate of one in somewhat less than a minute on the 870 Document Writer.

At Yale the original decklets of punched cards have been used for the past half year to produce the monthly Bulletin of the Yale Medical Library which lists the accessions of the previous month. The only extra work which involves the catalogers is indicating by a letter in column 66 of the worksheet whether or not that title is to go into the Bulletin. A 1401 computer prepares copy for the Bulletin. The program used in this operation necessarily takes out all special characters and flags since the printing now must be done all in upper case. Formerly it took one person one week of each month to prepare copy for the monthly Bulletin, but it is now prepared in less than thirty minutes of computer time and at a cost in the vicinity of $25. Moreover, the present Bulletin is 50 per cent larger than the earlier manually-prepared issues.

Perhaps the most difficult problem to solve in the computer processing of bibliographic information is the alphabetical sorting of entries. There is no possible way that a computer can be programmed so that it will know when to alphabetize “St.” as “saint” and when as “street.” The Columbia-Harvard-Yale solution to this problem is somewhat inelegant but appears to work. Other libraries employ essentially the same technique. Whenever the sorting characters of the filing element differ from the actual characters—which is not often—the cataloger writes out the sorting characters on a line at the bottom of the worksheet and assigns that line a special code. In this line numbers and abbreviations are spelled out. The sort program uses this line to alphabetize the entry.

The system is being designed in anticipation of the inclusion of acquisitions and circulation activities in mechanization procedures. Indeed, at Columbia work has already progressed to the trial stage in the use of punched cards in acquisitions with the same cards being subsequently employed in the cataloging process. Similarly, it should be possible to expand information retrieval systems based on the present procedures.
Conclusion

The Columbia-Harvard-Yale Project has developed several principles of small magnitude that are, nevertheless, effective guides past various pitfalls. First of all, flags or signals should not be characters used in routine print-out and should occupy only one column on a punched card. Also, signals which are special signals for the computer should be avoided except in computer output at the very end of the computer operation. Among the flags used by the Project are the “at sign,” delta, less-than sign, greater-than sign, lozenge, and group mark, but the group mark appears only in the cards to go into the 870, these cards being punched out at the end of the computer processing. Another principle is that sub-fields within the overall record length should not be a fixed length. The third principle is that the data used for sorting should be as long as possible; in the Columbia-Harvard-Yale Project system the sort control includes the first fifty characters from the filing entry. Another principle is that it is most desirable to have at least one application from as near the beginning of a project as is practicable. The Yale Medical Library’s Bulletin has served this function and served it well, for several difficulties were detected in writing the program for the Bulletin production, and in the processing of the cataloging data which goes into the Bulletin. Some of these stumbling blocks would have become major entanglements had they remained undetected until the information retrieval system was being activated.

The Columbia-Harvard-Yale Medical Libraries Project is attempting to design a fast, on-line catalog and journal index information retrieval system together with mechanized catalog card production that can be used as a base to expand to total computerization of only the catalog and indexing of the relatively small, heavily-used core collections of books and journals furnishing upwards of 75 per cent of recorded usage. Finally, the goal of the present system is to increase speed and completeness in supplying users with cataloging and indexing references.

REFERENCES


The woods are full of those who have harnessed machines—the professional literature of the past few years reports many breakthroughs of far-reaching importance and is filled with one success story after another in such diverse fields as auto-abstracting, file loading, machine translation, and successful search strategies for electronically massaging huge masses of stored bibliographic data.

For some reason, evidence of difficulties or of failures does not seem to float to the top as easily. While we have not been at this business long enough to be counted failures, we can certainly lend perspective as far as difficulties go. To characterize Columbia’s libraries will help put the description of both our activities and our problems in context.

First, a few notes on size. The general cataloged collections include about 3.2 million volumes, and they grow by about 100,000 volumes each year, only about half of which are in English. Nearly 50,000 serial titles (including documents) are acquired on a current basis. A million or more manuscripts, and an adequate number of items in other typical categories such as technical reports, maps, scores, and microtext are also on hand. The full- and part-time staff, in full-time equivalent, now numbers over 400. Over a million and a half books are charged for outside use each year, and a hundred thousand overdue notices are written to get them back. On an average day during the academic year, readers enter one or another of the thirty-two library doors on the campus 16,000 times—a figure about equal to the full-time enrollment in the university proper.

The annual operating budget, supplemented by special funds, is close to $3,000,000. It is estimated that over 60 per cent of this amount goes to support research while something less than 40 per cent goes to support the instructional program of the university. Roughly one-fourth of this sum goes for books, journals, and binding;

Warren J. Haas is Associate Director, The Libraries, Columbia University, New York City.
4 per cent goes for expendable supplies, leaving approximately 70 per cent for direct wage and salary payments—about one-third for technical service departments and two-thirds for reader service departments.

Organizationally, all library units, including law and medicine, are administered by the Director of Libraries. Two library units are operated on contract—one for the National Aeronautics and Space Administration and one for the National Institutes of Health.

While still not as large as Harvard University or the University of Illinois, Columbia is out of the special library class, and it is already apparent that leading a library of this size down the path of automation is a difficult task.

How do we begin to automate significant portions of this somewhat unwieldy organization? One thing that seems certain is that the adage "the bigger they are, the harder they fall," is valid beyond question. A library with sharply limited subject responsibilities, or with a small (or at least a homogeneous) group of readers, or one without a catalog of both rational and irrational procedures and practices shaped by years of history has a far easier path to follow in instituting radical changes than does a general research library. It took a team of experts two years to decide if it was feasible to begin to plan how to automate the bibliographic processes of the Library of Congress. It will probably, and unavoidably, take two more years before it is decided whether or not to take the next step—that of planning (or more accurately, inventing) the system required to accomplish automation. Without dwelling further on this fact, it may be asserted that large general research libraries, unlike specialized libraries, are not transformed overnight.

But we cannot sit back and do nothing simply because what needs to be done is difficult and slow. Columbia, like many other institutions, has been dipping its toes in the water of automation in recent months, principally to test the temperature before actually committing itself to taking the bath.

A related project to the Yale-Harvard-Columbia Medical Catalog project is centered in Columbia's engineering and physical science group of libraries. A detailed systems analysis has been under way for a short time. The object is to create a record system that will take up at the point an item is selected for the collections (whether before or after acquisition) and be used for all subsequent transactions and processing activities. As a first step (and as evidence that Columbia is serious), those science units not using Library of Congress (LC) classification were switched July 1, 1963, to help implement a concept of collection mobility judged to be an inseparable part of automatic record generation. A draft of a universal process form has been developed, and during recent weeks it has been walked through the various phases of processing to eliminate some of the
more obvious "bugs." Among the things aspired to are printed book catalogs, a weekly printout report called "status of selections," perhaps a Selective Dissemination of Information (SDI) system, and a number of other output products. Two specifications for this project are that it be compatible with the medical program and that it be flexible enough to be extended to other subject fields. We are not walking here yet, but we do seem to be beginning to crawl.

Another example of Columbia activity, and one in which progress might come quite quickly, is in what might be called one of our special libraries. In January 1964, Columbia contracted with the National Institutes of Health to develop and operate a national information center on Parkinsonism and related diseases of the basal ganglia. Without going into detail, it may be noted that the services of the center, which is a possible prototype for other disease-oriented research and information centers, are to include on-demand searches of literature to produce bibliographies as well as substantive data, publication of critical reviews of reports of work done in pertinent subjects throughout the world, organization of symposia, creation and maintenance of a "who knows what" type of file, etc. Work on a thesaurus of terms is under way as a first step towards creation of a machineable file of bibliographic information, and initial planning for a comprehensive information system has started. A distinctive characteristic of this project is the provision that a portion of the salary of each doctor and scientist attached to the Parkinson Research Center is charged to the information center contract—a device designed to stimulate participation of the scientific staff in the work of the information center.

A fair amount of spade work in other areas has been done in recent months—for the most part, it has been directed towards learning more about what is already known. For example, Columbia has a descriptive inventory of all currently maintained records—bibliographic, personnel, process, statistical, etc.—in the library system. They total about 1,000. Much information about the flow of material through the system by the use of log sheets inserted in several hundred sample items as they were unwrapped in the shipping room has been gathered. As a matter of fact, about 10 per cent of these forms have not yet returned to home base—but it has been only a year.

These examples, along with several others that might be noted, suggest perhaps that a crash program to automate Columbia's libraries is gaining momentum. Such is not the case. Columbia's objective is not automation. It is rather to provide effective support for each of the many and diverse instruction and research programs that constitute the work of a complex university. The library services required must be appropriate in type, in quantity, and in quality. They must be flexible to meet changing needs, and they must at the same time offer continuity and incorporate perspective.
Automation will certainly help us achieve service with these characteristics, but at the moment, we are more concerned with what we do, rather than how we do it. Neither Columbia nor any other library can fulfill its obligations by doing better and better what need not be done at all.

As libraries grow in size, the process of program development and performance evaluation becomes more complex and less subject to critical administrative review. In itself, this is not necessarily bad because responsibility for this kind of review can be shared on a wider base. But this same element of size makes it difficult for the larger group of operating policy makers (40 or 50 people at Columbia) to be aware of all the facts pertinent to the problem at hand.

The problem alluded to here is deceptively simple, and can be stated in many ways, but essentially it is this: How can we make certain that we select a proper course of action from among a number of alternatives to achieve an objective that is itself related to a whole complex of other objectives?

Several months ago, Columbia embarked on a type of operations research program known as Simulation of the Columbia University Libraries (SCUL), in an effort to see if a way could be devised to give insight into this fundamental problem of library operation.

Briefly stated, the specific objective of SCUL is to study the comprehensive research library as an economic system. This approach has been successful in some business applications, and the fact that much work has been done in the study of economic systems using computer simulation and mathematical modeling techniques has enabled the SCUL study to capitalize on the experience of others in the field.

At this point, only a part of the first phase of the project, essentially a limited feasibility study, has been completed. The product of this initial effort includes a computer program that simulates the interaction of readers with materials in Columbia’s Engineering Library, an outline of proposed mathematical approaches to the task of creating an economic model, a distinctive questionnaire designed for the collection of some of the required data, and a fuller realizations of the magnitude of the job we have proposed for ourselves. At the moment, funds required to get on with the main job are being sought.

The form that SCUL finally takes is certain to differ from its present state as general concepts are molded to fit library application. In brief outline, the project incorporates development of a probabilistic simulation model of a library in the form of a computer program that will be used to game with patron sets to study the nature of the interaction in varying situations between categories of patrons on the one hand and categories of library materials and library facilities on the other. The output from the simulation model will be a measure of the "satisfaction" experienced by each patron category for any given
mode of library operation (actual or hypothetical). This “satisfaction function” for a single category of readers will be adjusted in the context of the total patron population using the technique of multiple regression and will be associated with relevant cost information. Comparable information for each alternative course of action will be similarly developed. This information will be used as input to an economic model yet to be devised, and will be analyzed using linear programming to determine the mix of alternatives that best satisfy some stated goal.

In the sections that follow, the major parts of the SCUL project are described as they have been developed thus far.

I. The simulation model.—Most SCUL project time has been devoted to the development of a computer program that will serve as a prototype for a general library simulator model. In essence, the program that has been written is used to create a “computer duplicate” of the public service side of Columbia’s Engineering Library.

A dynamic replica of the library is created by playing library patrons, library facilities, and library stock against each other to analyze the complex relationships that exist between these three elements to learn more about the demand on stock and to establish the satisfaction of patron groups in any given mode of library utilization. The model can be operated under different conditions in order to (1) analyze in detail the real-life library, (2) to determine the effect on this library of a shift in the composition of the patron group using it, and (3) to investigate the effect on service (patron satisfaction) of changes in management policies affecting facilities or stock.

As a first step in formulating the simulation, each of the three operating elements in the model were categorized in the following manner.

I. Patrons, or the population using the library

<table>
<thead>
<tr>
<th>Major Categories</th>
<th>Minor Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undergraduates</td>
<td>Chemical engineering</td>
</tr>
<tr>
<td>Graduate students</td>
<td>Civil engineering</td>
</tr>
<tr>
<td>Teaching staff</td>
<td>Mechanical engineering</td>
</tr>
<tr>
<td>Research staff</td>
<td>etc.</td>
</tr>
<tr>
<td>etc.</td>
<td>etc.</td>
</tr>
</tbody>
</table>

II. Facilities

A. Those provided for the comfort and convenience of patrons:

Furniture
Microtext readers
Photocopy equipment
etc.
B. Library intermediaries between patrons and stock, including:

Card catalogs
Indexes and abstract journals
Reference librarians
Clerks
etc.

III. Stock, or objects in the library containing information used by the patron population:

<table>
<thead>
<tr>
<th>Major Categories</th>
<th>Minor Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>Format</td>
</tr>
<tr>
<td>Journals</td>
<td>Full size</td>
</tr>
<tr>
<td>Technical reports</td>
<td>Microform</td>
</tr>
<tr>
<td>Theses</td>
<td>Date</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>pre-1951</td>
</tr>
<tr>
<td></td>
<td>1951-1960</td>
</tr>
<tr>
<td></td>
<td>1960-</td>
</tr>
<tr>
<td>Language</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
</tr>
<tr>
<td>Romance</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
<tr>
<td>Type of loan</td>
<td></td>
</tr>
<tr>
<td>Non-circulating</td>
<td></td>
</tr>
<tr>
<td>Overnight</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
<tr>
<td>Use</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Reserve</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
<tr>
<td>Subject</td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
</tr>
</tbody>
</table>

The second step in developing the model was the construction of detailed flow charts tracing the paths of patrons entering the library, performing one or a number of possible functions, and then leaving. Following completion of the charts, the program which translated the flow charts into computer code was written.

A deck of punched cards, representing a set of patrons, is processed through the simulator program, duplicating the flow of a set of real patrons through a real library. The route each "patron" takes through the simulator model is established by a gaming process. At
each decision point, the program compares a known probability that the specific patron will perform a specific function with a random number generated by a subprogram within the simulator. If the random number is equal to or smaller than the known probability, the decision is "yes"; otherwise, it is "no." Step by step through the program, courses of action are determined by probability tables tied to each decision point.

In another mode of operation, the simulator can process "patrons" in a nonprobabilistic manner—in effect, specifying that a patron will follow a specific path or will use a specified facility.

For each run of a set of patrons, summary reports of patron action and library performance for each patron category are prepared. From these reports, the "satisfaction function" already referred to is calculated for use as input into the economic model. Because the required data has not yet been collected, runs so far have been limited to small sample sets, and the probability tables have been artificially generated.

While simulator output is generated primarily for use in the economic model, it is hoped that it will be useful in itself, since the model produces an analogous account of how the library's facilities are being utilized by the patrons and how well the demands of the various patron categories are met. The model will also predict the changes in stock demand resulting from a change in the proportions of patron categories utilizing the library. Further, the simulation model is also a laboratory library, because it makes possible tests of alternative management decisions and thus provides a way to assess changes before they are actually made.

II. Data gathering.—There are two types of probabilities involved in the library simulation. The first describes the order, or sequence, of patron activities, and the second describes the patron's probability of success. To gather those facts about present library operation that are required to develop the probability tables, a questionnaire in the format of a flow-chart has been developed. The questionnaire has been tested in the Engineering Library, but is has not yet been put to large scale use. It is also possible that the results of work being carried out at the Massachusetts Institute of Technology will provide probability information that can be used in this phase of the SCUL project.

III. The economic model.—The second model implied but not yet developed for the project is an economic model that will hopefully provide insight into a wide range of administrative problems by answering questions of the following type. Given a set of alternatives in library service to various patron groups and a specific allocation of funds to the library (the library budget) and given a set of requirements imposed on the library (service goals), what is the optimum distribution of the allocated funds to satisfy the requirements set?
In brief, this model will characterize the library operation as an economic system. The output of the simulation model (e.g. the derived “satisfaction function” for any or all alternative methods of operation) is coupled with cost information and analyzed by a linear program to determine the mix of alternatives that maximizes the effectiveness of the library for every dollar spent.

Only tentative approaches to the construction of the economic model have been taken. The entire process promises to be an undertaking of great complexity. The determination of cost information for existing modes of operation requires extensive and imaginative study; to establish meaningful costs for projected or hypothetical changes makes the task even more difficult. Areas of operation that seem particularly fruitful will have to be identified. Establishing relationships between, and constraints on, variables and expressing objectives and policies in quantitative terms will require a kind of analysis and a point of view that is new to library administration. The actual formulation of the problem will present complexities of many kinds, but this is to be expected simply because the nature of a library is itself complex.

From this brief description of the SCUL project it is evident that we have far to go before we can determine the utility of this approach, but thus far the promise of the project is such that we hope to continue what we have begun.

Benefits of many kinds will inevitably come from this kind of intensive research into library operations, even if the final results differ from those looked for at the beginning of the project.

It is already obvious that any significant success of this project implies major administrative and operational changes. For example, program objectives of the library will have to be carefully related to every segment of the university program and stated with more precision than has been the case in the past. The mission of the library will have to be reviewed and understood by all concerned parties in the university as well as within the library. Because a university library is in many ways a microcosm of its parent body, this very process might have interesting and useful supplementary effects.

Second, it is evident that a management team of a type new to libraries will have to be developed to employ effectively and utilize fully the results of management techniques of the kind contemplated.

Finally, because success of the SCUL concept is dependent on a continuing flow of data to make the models honestly reflect the real-life situation, it is evident that an integrated and automatic system to generate information as a by-product of every important library operation will have to be devised. Planning for an output of useful information should be an important part of every system component designed to carry out library operations.
Thus far we have described by example some of the Columbia projects that have already, or will soon, involve us in the use of data processing equipment. In the course of the next two or three years, several of these activities now in the formative stage will be fully operational. But as we have moved along in recent months, we have been reminded again and again that we are not coming to grips with some of the basic problems that must be solved if the promise of data processing machines applied to library operations on a nation-wide scale is to measure up to the visions we have been induced to accept.

First, it seems unlikely that most members of a staff of a large research library—a staff already responsible for carrying on a substantial load of day-to-day operations—can put their regular work aside for the time required to become conversant with machine techniques, and then devise, install, and operate a new system. A mountain of undone work would quickly grow and bury them. For example, we have done some detailed work in flow charting serial processing, but so far no one on the serials acquisitions or cataloging staff has found a way to create the new world while coping with the old—the simple process of handling the half-million items that come their way each year dominates time and energy. How do we surmount this dilemma? Do we have to create a parallel system, including a duplicate staff, to move from the world of the 3" x 5" card and the visible record to that of magnetic tape and printed holdings lists? Or should we break up our central serials acquisitions system into smaller subject-oriented units and revamp them one by one? Is it possible that a very large library must break up into a federation of libraries before changes of the magnitude we envision can be accomplished? Is there a limit to how big or old a dog can be if he is to learn new tricks?

A second, and related, question concerns the amount of what might be called "risk capital" that an academic library should spend to assure a progressive program of operational evolution. The SCUL project, just described, much of which was done by a private firm, has cost about $15,000 already, not counting substantial amounts of library staff time or computer time—and this is for pure research devised simply to test a methodology, with no guarantee of a pay-off. My question—are academic institutions too conservative generally in investing capital on a planned basis to improve operations? Higher education, judged on the basis of dollar expenditures, is big business and is growing bigger. Perhaps both libraries and the institutions of which they are a part need to provide in their regular budgets for more research into their way of operating.

Next, is it reasonable or even rational for every library to go off on its own to establish a type font and design a format for what should be generally useful and useable bibliographic information? Johns Hopkins is now hunting a way to convert its shelf list to tape.
The Library of Congress might one day go to work producing machineable records for current publications and might ultimately find itself involved in converting The National Union Catalog (NUC) and the Union List. The New York Public Library has major catalog problems and might have to get into converting some of its records into machineable form. The Yale-Harvard-Columbia medical group is working on format and establishing requirements for type fonts. One group in the government is out to establish a nation-wide information network for scientific material published in journals. This list could go on and on. Many other organizations and individuals are involved. It will be a little short of tragic if some sort of national machinery is not soon created for coordinated development of a generally acceptable format for basic bibliographic information and a standard font of characters for printing with data processing equipment. I shall go all the way and suggest that perhaps the President of the United States might properly create a permanent Commission on Access to Recorded Knowledge to tie together the multitude of activities in this area, in and out of government. The acronym ARK is itself symbolic of the flood of uncoordinated (and therefore both competitive and redundant) solutions to bibliographic control. When this fundamental problem is solved, I can suggest others just as basic to keep this commission active for some time to come. In the long run a high level official organization responsible for optimizing access to recorded information might prove as important to us all as the Atomic Energy Commission or the Fish and Game Commission. After all, when we are dealing with recorded knowledge, we are dealing with the cumulated product of the brain power of the human race. I cannot think of anything that deserves more care.

In short, individual institutions can and should move to try new methods of operations and analysis—everyone can learn from such efforts. They can, and should, seek new ways to handle administrative and operating services such as circulation control, collection maintenance and inventory records, and business and fiscal aspects of acquisitions. But in the field of bibliographic control, to say nothing of text storage, it seems both impossible and unrealistic for any large general research library to step out on its own. The research resources of this country need to be linked by more than transitory ad hoc committees or a complex of professional associations.

A final problem, one that must be solved if coordinated cataloging is ever to be achieved on a massive scale, involves the matching of a book in hand to a remote bibliographic record. How can I pick up a book in Hungarian (one of the several languages I do not read), on a subject I know nothing about, and locate the descriptive and analytical bibliographic information for that book? At present, one has practically to catalog the book before he can begin to search for this information. We need an internationally understood and
automatically derived means to describe a printed item. For example, if imprint information were noted as day, month, year, instead of year alone, it would be easy to devise a number made up of the imprint, the number of the last numbered page, and a code for the language of the title page. Thus anyone, anywhere in the world, could pick up a book, and describe it by the same number that would be established by any other person working with the same book. This number could be attached to a bibliographic record, wherever and whenever this record was created—and we would be on the way to tying the item in hand to a remote, machine-stored record. A little work on the probability of generating duplicate numbers for different items would need to go into the composition of the code. Some duplication would seem acceptable, because this would mean that one would simply select the right information for the book in hand from two or three records produced out of the system.

The problems I have isolated here are large ones—they are not concerned with machine configuration or programming shortcuts, or even such important questions as how are enough people to be trained to meet the demand for the special skills required. But the solutions to these larger questions and others of similar magnitude are required, or must at least be on the way, before large, general, research libraries can join fully and without reservations in this revolution in the methodology of librarianship.
Many years ago an advertisement appeared frequently in popular magazines which displayed a photograph of a man or a woman seated before a piano. The caption below read: "I learned to play in five easy lessons." We are going to learn the principles of computer programming in one easy lesson.

We know, of course, that it is possible to learn to play a simple tune, "Twinkle, Twinkle, Little Star," and a few others, over a period of five weeks, with practice and five easy lessons. This is, however, a limited repertoire. This hoax—that one can learn anything in five easy lessons—will be perpetuated by this talk. A few simple, but typical, procedures will be examined, and the corresponding computer programs will be outlined. The characteristics of programs in connection with these exercises will be discussed, but it takes years to learn to perform well on a computer.

Let us begin with a simple, but typical, computer program: the calculation of a mean or average. We may wish to calculate a student's grade point average, or the average daily amount of money spent on food, or, for income tax purposes, the average number of miles traveled in a car, or, for the library, the average cost of cataloging a book.

The mean, of course, is found by adding the scores together and then by dividing the sum by the number of observations, N, as shown in equation (1).

\[(1) \text{MEAN} = \frac{1}{N} (X_1 + X_2 + \ldots + X_N) = \frac{1}{N} \sum X_i\]

To add to the clarity of the example, the variable X shall be replaced with some real numbers as in equation (2).

\[(2) \text{MEAN} = \frac{1}{4}(-16 + 20 - 4 + 8) = 2\]
A characteristic of an equation is that it is in balance. One side of the equation is equal to the other. Therefore, we may regard an equation as a static statement of a relationship. Moreover, if each of several persons were asked to evaluate the right side of equation (2), a great variety of procedures would emerge.

Some persons first would add the positive numbers, then the negative numbers, and finally the difference would be found and divided by the number of observations. Others would hunt for easy number combinations. In equation (2) they might notice that -16, +20, and -4 add to zero. So +8 would be divided by +4. Many would begin by rewriting the numbers in a column, or in two columns, one for positive numbers and one for negative numbers. It is also possible to begin by dividing each number in the sum by +4. Then the sum of the quotients is equal to the mean.

An equation is a static relationship, for it does not tell us what steps to take or even that we should take the steps. A computer program, on the other hand, is a set of dynamic statements which imply action. It tells a computer to perform a series of steps, and it specifies exactly in what order these steps are to be performed.

Let us introduce at this time a dynamic relational symbol to replace the equal sign. This symbol, :=, has been borrowed from ALGOL, a computer programming language. It means more than "equal"; it means "set equal to." Thus, the statement, SUM := 0, means set SUM equal to zero. Notice that this relational symbol permits such statements as

\[(3) \ I := I + 1.\]

As an equation, (3) is nonsense, for we can transpose I from one side of the equation to another, yielding 0 = 1. As a dynamic statement of relationship, however, (3) makes sense, for it means that I is to be set equal to the previous I plus one.

Now let us suppose that we have some hypothetical computer. Like all computers, our computer has an arithmetic unit capable of performing addition, subtraction, multiplication, and division, and other logical operations which shall be introduced later as needed.

Like all computers, our computer has a memory consisting of a large number of locations or slots in which information can be stored. These locations are numbered from zero through L, and they can be referred to by these address numbers. In our computer, each of these locations can also be referred to by a symbolic address. Thus, SUM might be the symbolic address of one such location.

However, we can call them anything we like: SUM, MEAN, CHARLIE, ADA, or BOOK. In fact, we can refer to a series of them as BOOK (1), BOOK (2), BOOK (3), and so on. Such a list shall be used in our program to calculate a mean. The scores will be referred to as X (1), X (2), X (3), and X (4), or more generally, as X (I) where I
<table>
<thead>
<tr>
<th>Address</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>4</td>
</tr>
<tr>
<td>I</td>
<td></td>
</tr>
<tr>
<td>SUM</td>
<td></td>
</tr>
<tr>
<td>MEAN</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>X (1)</td>
<td>-16</td>
</tr>
<tr>
<td>X (2)</td>
<td>20</td>
</tr>
<tr>
<td>X (3)</td>
<td>-4</td>
</tr>
<tr>
<td>X (4)</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>L - 1</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1—Memory Locations

can take on successively increasing values (see Fig. 1). For our program we shall need a location to contain the sum as we add the numbers together. We shall refer to it symbolically as SUM. We shall want to keep track of the numbers as we add them by increasing our tally location \( I \) so that exactly \( N = 4 \) numbers are processed.

Since computers can be instructed to perform almost an infinite number and variety of tasks, there is a correspondingly large variation in computer programs. A characteristic of nearly all programs, except those of the most trivial kind, is that they are made up of one or more iterative loops. Any sequence of computer instructions which is operated upon repeatedly may be called an iterative loop. Our program to calculate a mean, which we shall begin to assemble shortly, has only one such loop. It will serve, nevertheless, to illustrate the properties of iterative loops.

It is often necessary to set up initial conditions prior to entry into an iterative loop. In our program, for example, we shall want both locations, SUM and I, to be cleared of the results of any previous computations. Otherwise, our program may miscalculate, and we shall have a program error. Let us use the word initialize to refer to this part of a computer program. For our program, the initialize section will consist of two statements:

\[
\text{SUM : = 0} \\
\text{I : = 0}
\]
The iterative loop for our program will consist of three statements. Let us label the first statement \((L)\) so that we can go back to it for each iteration. First, we shall write the loop, and then we shall discuss it.

\[
\begin{align*}
(L) \quad I & : = I + 1 \quad \text{Increment} \\
\text{SUM} & : = \text{SUM} + X (I) \quad \text{Recurrence relation} \\
\text{IF} \quad I & < N, \ \text{GO TO} \ (L) \quad \text{Test}
\end{align*}
\]

The first statement of loop \((L)\) sets \(I\) equal to the previous \(I\) and tallies one. But the previous \(I\) was made zero by the initialize section. Therefore, at this point \(I\) is merely equal to one. The second statement, which is called a recurrence relation, says to set \(\text{SUM}\) equal to the previous \(\text{SUM}\), which was cleared to zero by the initialization, and to add \(X (I)\). Since \(I\) is equal to one, this means to add \(X (1)\) or -16. Finally, we arrive at the test to determine if we have added \(N = 4\) scores. Notice that we have underlined \(\text{IF}\) and \(\text{GO TO}\). This underlining is done to avoid confusion with a symbolic address which might be called \(\text{IF}\) or \(\text{GO TO}\). The tally \(I\) will be less than \(N\), and so we follow the directions of the statement and return to the beginning of loop \((L)\).

Let us examine what will happen on successive passes through the iterative loop. Each time \(I\) will be increased by 1, and so each time \(X (I)\) will refer to the next score in the list. \(\text{SUM}\) will take on values as follows:

- **First time:** \(\text{SUM} = 0 + X (1) = -16\)
- **Second time:** \(\text{SUM} = -16 + X (2) = -16 + 20 = 4\)
- **Third time:** \(\text{SUM} = 4 + X (3) = 4 - 4 = 0\)
- **Fourth time:** \(\text{SUM} = 0 + X (4) = 0 + 8 = 8\)

When \(I\) becomes 4, it will no longer be less than \(N\), and so the test will not return control to the beginning of loop \((L)\), but will continue to the next part of the program.

The next part of the program will be another process section, but this time no iterative loop is involved. The process consists merely of dividing \(\text{SUM}\) by \(N\) to find the mean.

Finally, we shall want to print the result, so we shall add an instruction to our list of permissible ones. Our hypothetical computer will interpret \text{WRITE JOE} as a command to activate the printer by printing the contents of memory location \text{JOE}, or whatever other location is specified.

Similarly, we shall need to add a \text{READ} instruction to our list to input the list of scores. This will have the form, \text{READ N, X (I)}, and will be interpreted to mean that \(N\) scores will be read and placed at symbolic locations \(X (1)\) through \(X (N)\).

At this point we can assemble the complete program to calculate a mean.
Program No. 1

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>START</td>
<td>Initialize</td>
</tr>
<tr>
<td>READ N, X(I)</td>
<td></td>
</tr>
<tr>
<td>SUM : = 0</td>
<td></td>
</tr>
<tr>
<td>I : = 0</td>
<td></td>
</tr>
<tr>
<td>(L) I : = I + 1</td>
<td>Process; iterative loop</td>
</tr>
<tr>
<td>SUM : = SUM + X(I)</td>
<td>Increment</td>
</tr>
<tr>
<td>IF I &lt; N, GO TO (L)</td>
<td>Recurrence relation</td>
</tr>
<tr>
<td>MEAN : = SUM / N</td>
<td>Test</td>
</tr>
<tr>
<td>WRITE MEAN</td>
<td>Process</td>
</tr>
<tr>
<td>STOP</td>
<td></td>
</tr>
</tbody>
</table>

You will notice the underlining to distinguish instructions from symbolic addresses. Also the label (L) is enclosed in parentheses for the same reason.

Program No. 1 is a set of dynamic statements which specify exactly how to calculate the mean. Furthermore, it is a perfectly general program, for it will operate as successfully when N = 10,000 or 100,000 as when N = 4.

Notice how compact the program is presented despite the many thousands of operations the program may imply if N were to be large. Notice also how easy it has been to write the program using such a language. Of course, the language we have been using does not exist in any well-developed form, for we have been making it up as we needed it. However, there are such kinds of languages. Examples which are well known are FORTRAN, COBOL, and ALGOL.

Before our program to calculate a mean could be used on a real computer, it would be necessary, of course, to translate these symbolic instructions into machine language. Fortunately, programs exist which do the translation for you, and FORTRAN, COBOL, and ALGOL compilers are now available on most machines.

Let us return to our program, however, for it should be pointed out that it is easy to make mistakes when writing a program. For example, in the iterative loop, the increment statement preceded the recurrence relation. If these were to be interchanged,

\[
\text{SUM} : = \text{SUM} + X (I) \\
I : = I + 1
\]

then the numbers to be processed would run from X (0) through X (N - 1). We do not know what number is contained in location X (0). It may cause only a trivial difference in the result, and especially if the list were a long one, the result may appear to be quite reasonable. The program would be accepted as correct, and
others would begin to use the program. Such an error could remain undetected for a long time. The consequences could be serious if crucial decisions were made on the basis of the results. In any event, it is embarrassing when such results are published.

Let us also consider an effect from writing the wrong test statement. Suppose that this statement is

\[
\text{IF } I \leq N, \text{ GO TO (L).}
\]

This would result in an extra passage through the loop and so X(N+1) would be added to the sum. Again the error may or may not be readily detected depending upon what might be contained at this location. Of course, there are many other ways to make errors, but the point in these illustrations is to show how easy it is to make an error. The warning is always to check the program carefully.

If you think that this warning is an overemphasis, let me remind you that some years ago, when the United States was trying to close the missile gap, a missile blew up at a cost of 40 millions of taxpayer dollars as a result of a trivial programming error. A hyphen was omitted from the program.

There are many ways to check a program, but one way is actually to trace the steps in the same order as they will occur during the execution of the program. To save time, let us illustrate this by checking the iterative loop of our program only, as illustrated in Table 1.

<table>
<thead>
<tr>
<th>Steps</th>
<th>I</th>
<th>SUM</th>
<th>X(I)</th>
<th>Test</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
<td>?</td>
<td>Initial conditions</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-16</td>
<td>1&lt;4</td>
<td>Yes, return to (L)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>20</td>
<td>2&lt;4</td>
<td>Yes, return to (L)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>-4</td>
<td>3&lt;4</td>
<td>Yes, return to (L)</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
<td>4&lt;4</td>
<td>NO, continue</td>
<td></td>
</tr>
</tbody>
</table>

It appears from Table 1 that the result + 8 at the end of the loop is correct, and this confirms that the program also is correct.

One reason that we have spent a large part of our time in discussing a program to calculate a mean is that it has served as a vehicle to introduce concepts and principles used in programming.
However, it will also serve as an outline for a large number of other procedures. For example, by modifying the recurrence relation we could use the program to find the mean of differences, or mean of products, or mean of quotients. Similar iterative loops are quite typical of other programs, and complex programs often consist of the assembling of a series of similar iterative loops. Thus, we have examined in detail a basic building block typical of many larger programs.

Before we end our discussion of programming, we should examine a program of a different kind—one involving more decisions. To illustrate this variety of program, let us pose a simple problem.

We are given a set of N observations, \( X(I) \), which may be the scores of students, the costs of books, or the weights of hogs. These also might be the titles of books since, as far as a computer is concerned, letters are merely special numbers. We wish to locate and print the largest number in the set, call it \( L \), and the smallest number in the set, call it \( S \).

Although this may appear to be a somewhat trivial exercise, like the mean, it is related to other more important problems. It is not unusual to want to place a set of numbers in order of size, or to alphabetize a set of book titles. To do either, it may be necessary to continue to locate the next smaller or next larger number.

We shall not repeat our step-by-step development as was done for the calculation of the mean, for by this time we know the main ideas connected with writing a program. Instead let us begin immediately to write our program. Later we shall test it to ascertain if there are errors.

### Program No. 2

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>START</td>
<td>Initialize</td>
</tr>
<tr>
<td>READ N, X(I)</td>
<td>At the point where only one number has been examined, it is both the largest and smallest.</td>
</tr>
<tr>
<td>L : = X(1)</td>
<td>The next number to be examined will be ( X(2) ).</td>
</tr>
<tr>
<td>S : = X(I)</td>
<td></td>
</tr>
<tr>
<td>I : = 2</td>
<td></td>
</tr>
<tr>
<td>(A) IF X(I) &gt; L, GO TO (C)</td>
<td>Test if larger than ( L ).</td>
</tr>
<tr>
<td>IF X(I) &lt; S, GO TO (D)</td>
<td>If not, test if smaller than ( S ).</td>
</tr>
<tr>
<td>(B) I : = I + 1</td>
<td>If neither, increment and test for end of loop.</td>
</tr>
<tr>
<td>IF I ( \leq ) N, GO TO (A)</td>
<td></td>
</tr>
<tr>
<td>WRITE S</td>
<td>If end of loop, process.</td>
</tr>
<tr>
<td>WRITE SPACE</td>
<td>We do not want to print ( L ) on top of ( S ),</td>
</tr>
</tbody>
</table>
WRITE L
STOP

(C) L := X(I)
GO TO (B)

(D) S := X(I)
GO TO (B)

so we cause a carriage advance on the printer.

If X(I) > L, replace L with X(I).
Return to loop.

If X(I) < S, replace S with X(I).
Return to loop.

Notice that this program is longer than the program for the mean despite the fact that no calculations are involved. The complexity of a program is more closely related to the number of decisions that it is required to make than to the number of calculations.

Further, errors are more likely to be made in problems with a large number of decisions so we shall check our program by tracing the steps one-by-one with a set of four numbers, as follows:

\[ X(1) = +2, \quad X(2) = -2, \quad X(3) = +4, \quad X(4) = +1 \]

TABLE 2

<table>
<thead>
<tr>
<th>Steps</th>
<th>I</th>
<th>X(I)</th>
<th>S</th>
<th>L</th>
<th>Tests and comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Initialize</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Read X(I)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>First number is both S and L</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Process loop</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Is -2 &gt; L? No, continue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Is -2 &lt; S? Yes, replace</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Test for end: Is 3 \leq 4? Yes, return to (A)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Is +4 &gt; L? Yes, replace</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Test for end: Is 4 \leq 4? Yes, return to (A)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Is +1 &gt; L? No, continue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Is +1 &lt; S? No, continue</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increment</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Test for end: Is 5 \leq 4? No, continue End of loop; next process</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WRITE S = -2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Space</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>WRITE L = +4</td>
</tr>
</tbody>
</table>

It appears that our program to find the largest and smallest number of a set is correct, for we have found the correct results.
It has often been remarked that one learns more by teaching a course than by taking a course. It also seems to be true that one learns more in trying to develop a program for a set of operations than he would either by teaching these procedures to others or by being taught.

Now that we have learned to write programs in one easy lesson, we can apply our talents to automating certain library applications. However, at the beginning of this paper, you were warned that it takes time to develop into an experienced and proficient programmer. Indeed, it has been said that an experienced programmer is one who has already made all of the mistakes. In that sense, we are still amateurs, for we wrote two programs without errors.
In 1961, the U. S. Office of Education (USOE) contracted with the Center for Documentation and Communication Research (CDCR) of the School of Library Science of Western Reserve University to develop a pilot information service of educational research materials. That this contract represented a new direction in information retrieval is illustrated by D. J. Foskett's preface to his *Classification and Indexing in the Social Sciences*:

> ... Enormous resources are devoted to the advancement of science and technology, and the dissemination of scientific information, without which these advances lose most of their significance, has been studied systematically for several years.

> This is not the case, however, in the field of the social sciences themselves... Yet hardly any studies have appeared of information of dissemination and retrieval in social science. New techniques of classification and indexing are only just beginning to make an impression, although they have already become commonplace in science.¹

The Center therefore welcomed the opportunity to apply its experience in documentation and information retrieval to the field of education, since it was felt that what could be learned about information retrieval in education might be applicable, at least in part, to the whole social science field. The USOE, at the same time, was faced with very practical problems in retrieving and disseminating educational research information and realized the potential contribution of an information service.

The product of the merger of the interest of the Center and the need of the USOE is reported in this paper.

Some explication of the specific nature of information problems in educational research will be useful as background for explaining the objectives of the project.

¹ Gordon C. Barhydt is Manager, Educational Research Information Projects, Center for Documentation and Communication Research, School of Library Science, Western Reserve University, Cleveland, Ohio.
Information Problems in Educational Research

Educational researchers are hampered by inadequate bibliographic control and infrequent and uncertain dissemination of educational research literature. They are further hampered by the absence of comprehensive and exhaustive collections of educational research. Bibliographic control is practically non-existent. The most frequently used index in education has employed more than 20,000 subject headings since it first appeared in print, and the number is rapidly increasing. Few indexes and fewer abstracting publications cautiously and randomly select a small sample of completed research from hundreds of journals regularly publishing research, and from the doctoral outpourings of innumerable colleges and universities. Almost totally ignored by standard indexes are written reports of sponsored research of foundations, of many agencies of the government, of state and local boards, and the unsponsored research of scores of individual researchers.

In the educational media field alone it was demonstrated that much research of interest was going unabstracted, unindexed, and probably unnoticed. Tauber and Lilley in their Feasibility Study Regarding the Establishment of an Educational Media Research Information Service stated that "...reports of research relating to new educational media are not represented satisfactorily in the existing bibliographic controls..." They further reported, that within those controls, considerable duplication of coverage existed.2

Small collections of research are scattered among libraries, research organizations, and individuals; few attempts have been made to gather them into a comprehensive and exhaustive whole.

A search conducted for the Center by the Science Information Exchange (SIE) early in 1963 produced abstracts of 686 current research projects in education. If one conservatively estimates each project as receiving $20,000 of support, the total is over $1,370,000, and the real total, because of the limited coverage of SIE at that time, is probably much higher. The value of this research is obviously wasted unless results can be adequately disseminated to other researchers, and ultimately translated into practice.

Based on the Center's knowledge of the information problem in education, certain specific objectives were formulated for the pilot phase of the project, 1961-62:

a. Analysis of subject content significantly deeper, more detailed and more flexible than that provided by existing systems.

b. Control, or cross-referencing, of terminology more flexible and more interdisciplinary in nature than that provided by existing systems.
c. A mechanism for exploiting the body of literature indexed in the manner described above which will permit the system to function on both a centralized and decentralized basis.3

The Center's work since the initiation of the project has been directed toward the furthering of those objectives. The balance of this paper reports our progress and is divided into three sections:

I. The CDCR Education Project.
   A. Orientation.
   B. Specifics of the System.
II. Current Research.
III. Future Research.

Since our current work is centered on a retrieval system for media and media-related educational research, most of the examples given are from this area.

The CDCR Educational Research Project

Orientation

The educational research project at the Center has three unique advantages:

1. Because it is only one of many research activities at the Center, it benefits from a substantial research effort in many fields and from the extensive work in information retrieval theory.
2. Because of the Center's contact with the research activities of other documentation centers in the U. S. and Europe, it is in close touch with many related efforts.
3. Because of the establishment of a pilot user group in October 1963, it benefits from the advice and experience of twenty key educational researchers.

Of the Center's varied research activities, the comparative systems laboratory, established by a grant from the National Institute of Health in June 1963, has perhaps most significance to the education project. Here components of several information retrieval systems are being isolated and compared under experimental conditions. Included in these comparative tests are system components applicable to a system for educational research literature.4

Complexities of system development demand involvement with every facet of documentation research, and two research activities outside the Center are of particular importance. The first is the work of the Classification Research Group in England, in particular
the faceted classification for education developed by D. J. Foskett, librarian of the University of London's Institute of Education.\(^5\)

We are fortunate that Foskett will be in the United States in the summer of 1964 and will act as consultant to the project at the Center during a portion of his stay. Of equal interest is the work of the Centre Nationale de la Recherche Scientifique in Paris, in the application of SYNTOL (Syntagmatic Organization Language) to social science literature.\(^6\)

Although advisory groups in information retrieval are not new, the Center's pilot user group provides the advantage of critical evaluation of experts based on their specific knowledge and use of the system. A conference was held in Cleveland in October 1963 to familiarize this group with the system and to seek their advice on several important problems, among these, criteria for the inclusion of material in the file. During the next eighteen months, they will submit questions and provide evaluations of the relevance of the responses.

These three advantages have proved to be invaluable supplements to the research activities in the project and have contributed substantially to the development of the system described below.

**Specifics of the System**

These can be grouped into eight divisions or processing steps:

1. acquisitions and selection,
2. analysis,
3. terminological control,
4. recording of results of analysis on a searchable medium,
5. storage of records or source documents,
6. question analysis and development of search strategy,
7. conducting of search,
8. delivery of results of search.

**Acquisitions and selection.**—The base point for acquiring media and media-related research was William Allen's bibliography for his summary of audio-visual communication in the Encyclopedia of Educational Research.\(^7\) A "citation index" search was conducted restricting selection where the material did not appear to be within the loosely defined limits specified by Title VII of the NDEA. Preliminary criteria for inclusion were then developed. Since this area is one of direct concern to educators and possibly one of only peripheral concern to librarians, a complete discussion may be found in the final report for Title VII Project B-170a.\(^8\) Basically the criteria are as follows. "Research," as we have defined it, means controlled experiment, the reporting of which is accompanied by quantified data. Included are research reviews if they make a contribution to the analysis or synthesis of a particular area. The file of "research" includes studies of and related to the utilization of the newer educational media (those made possible by technological advances, e.g., educational television (ETV), motion pictures, teaching machines, etc.) within intentional, human learning situations employing meaningful materials.
At this point in the development of criteria for inclusion we are in the role of judge; a judge, as defined by H. L. Mencken, is "A law student who marks his own examination-papers." We do have what can be considered preliminary criteria: a base for extension or reduction of the contents of the file.

Analysis and terminological control.—The Western Reserve University semantic-coded telegraphic abstract approach has been applied to the research studies in the file. In view of our present and past work in other fields and our current work in other techniques, the Center feels that this approach is a reasonable one. It offers the capability of providing specific, generic, and other relationships necessary in dealing with educational research literature. We are prepared to modify the system if it seems advisable, and to incorporate, where appropriate, the results of our own research and the research of others.

The first step in analysis is to prepare a telegraphic abstract (TA) (see Fig. 1). The TA is designed to provide a detailed machine-readable index to a research study. An abstracter selects those words from a document which have a high indexing value. Although the abstracter is free to select any indexable term from the document, freedom of selection is limited by well-defined rules governing the inclusion of certain types of information for a particular kind of study. These terms appear in the right hand column of the TA form.

The abstracter then establishes several kinds of relationships among these terms by the use of role indicators. Role indicators indicate logical relationships between terms,

KEJ = population
KAM = process
KQJ = agent of process (by means of)
KWJ = device or material prepared

or facets of the study,

KEC = subject matter taught
KAP = dependent variable(s)
KAL = independent variable(s)

or provide descriptive information.

KAB = type of material or study
KIT = date of study

Punctuation or level indicators are also incorporated into the TA. These symbols (., .), (.), (, ) are signals of the closeness of association between the elements of a TA: role indicators, words, etc., preventing cross-talk between separate portions of the TA during the searching operation. The level indicators underlined, separate
## TELEGRAPHIC ABSTRACT

<table>
<thead>
<tr>
<th>Col. 6-8</th>
<th>Role Indicator (Col. 28-80)</th>
<th>Col. 6-8</th>
<th>Description (Col. 9-27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>..KAB,</td>
<td>2</td>
<td>Research</td>
</tr>
<tr>
<td>3</td>
<td>.KIT,</td>
<td>4</td>
<td>1962</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>..KEJ,</td>
<td>8</td>
<td>Experimental Group</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>..KEJ,</td>
<td>12</td>
<td>Control Group</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>..KAM,</td>
<td>16</td>
<td>Matching</td>
</tr>
<tr>
<td>17</td>
<td>.KQJ,</td>
<td>18</td>
<td>Mathematics</td>
</tr>
<tr>
<td>19</td>
<td></td>
<td>20</td>
<td>Aptitude</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>..KEC,</td>
<td>24</td>
<td>Electronics</td>
</tr>
<tr>
<td>25</td>
<td>.KQJ,</td>
<td>26</td>
<td>Filmed</td>
</tr>
<tr>
<td>27</td>
<td></td>
<td>28</td>
<td>Programmed</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td>30</td>
<td>Instruction</td>
</tr>
<tr>
<td>31</td>
<td>.KQJ,</td>
<td>32</td>
<td>Lecture</td>
</tr>
<tr>
<td>33</td>
<td></td>
<td>34</td>
<td>Demonstration</td>
</tr>
<tr>
<td>35</td>
<td>.KQJ,</td>
<td>36</td>
<td>Textbook</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>..KAM,</td>
<td>40</td>
<td>Testing</td>
</tr>
<tr>
<td>41</td>
<td>.KQJ,</td>
<td>42</td>
<td>Psych Corp EPSAT</td>
</tr>
<tr>
<td>43</td>
<td>.KQJ,</td>
<td>44</td>
<td>Achievement</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td>46</td>
<td>Test</td>
</tr>
<tr>
<td>47</td>
<td>.KQJ,</td>
<td>48</td>
<td>Mathematics</td>
</tr>
<tr>
<td>49</td>
<td></td>
<td>50</td>
<td>Test</td>
</tr>
</tbody>
</table>

**Abstracter**

---

Fig. 1
KAP, (dependent variable)  Experimental Group
     Reading
     Listening

KAL, (independent variable)  Language Laboratory
     I.Q.

from

KAP,  Control Group
     Reading
     Listening
     Lecture
     Demonstration
     I.Q.

KAL,  

Where in this example the independent variables (.KAL,) are associated with their appropriate groups (.KAP, experimental group or .KAP, control group) and can be so specified in the search program. (A complete list of role indicators and punctuation levels is given in Appendix A.)

The next step is the encoding of the TA by the application of the semantic code to each word listed. If the word has previously appeared in a TA and been coded, this may be accomplished by mechanical means. If not, the process is as follows.

The semantic code is comprised of semantic factors—three letter combinations representing concepts; alphabetical infixes, which show the relationship of the factor to the word being coded; numerical infixes, which delimit a concept; and numerical suffixes, which establish the uniqueness of each code. For example, the code for the Minnesota Multiphasic Personality Inventory (MMPI) is

DACM MUSR MYMT 1017 3102.

Breaking this down we have the semantic factors

D—CM = printed document
M—SR = measurement
M—MT = emotion

and adding the alphabetical infixes appropriate for each factor, we have

A = categorical infix
U = productive infix
Y = attributive infix

The code tells us that the MMPI
is a document = DACM
is used for measurement = MUSR and that
the concept emotion is an important characteristic of the word (s) coded - MYMT.
Since there are many aspects of the concept emotion, a numerical infix has been assigned to the factor M-MT to designate, in this instance, the concept of personality.

MYMT 1017 - Personality

Thus far we have DACM MUSR MYMT 1017. Since other closely related tests may be coded in the same way, e.g., The Rorschach Ink-blot Test (RIT), a numerical suffix is added to the end of each complete code to establish the code as unique.

\[
\begin{align*}
\text{MMPI} & = \text{DACM MUSR MYMT 1017 3102} \\
\text{RIT} & = \text{DACM MUSR MYMT 1017 3304}
\end{align*}
\]

A search, therefore, can be made on any generic to specific level retrieving all tests of this type (by programming for DACM, MUSR, MYMT 1017) or by specifying the unique code for a specific test. Utilizing the semantic code and combining it with the relationships established by the TA, a very powerful searching tool can be constructed.

Concurrent with the preparation of the TA, the abstracter prepares a conventional abstract of the original document (see Fig. 2).

Figure 2
Conventional Abstract


The Keesler Mathematics Test and the Psychological Corporation Electronic and Physical Sciences Aptitude Test are used to match three groups of Air Force trainees during six weeks of a course on the principles of electronic communication. The experimental group consists of a randomly selected set of 14 students with scores in the middle 60 per cent of the distribution. The control group is a matched set of 14 students who are aware of their participation in a research project, but who are taught by lecture-demonstration. The blind control group, another matched set of 14 airmen, is also taught by
lecture-demonstration, but is wholly unaware that its performance is under experimental consideration. The experimental group receives all of its instruction from 35 mm film projected with the AutoTutor Mark I. The film is organized along the principles of intrinsic programming. Three progress tests are administered at two week intervals and scores are analyzed by F ratio, analysis of variance and t-test. No significant differences are found between control and blind control groups. While examination scores for control groups are somewhat higher than scores for the experimental groups, the differences are not great. A replication of the original study produces results which are not significantly different.

Recording of results of analysis on a searchable medium.—Each role indicator along with its punctuation, and each word on the TA are punched on separate Hollerith cards. The words are matched with a card reproduction of the code dictionary and where a word has previously been encoded the proper code is gang-punched from the dictionary card into the word card. Codes are assigned by an individual to new words entering the system, and these new words and their codes are added to the code dictionary. All cards for role indicators and coded words are then sorted in the order in which they appear in the TA. Processing in blocks of 100 abstracts, the detailed index (TA) is transferred from the cards to storage on magnetic tape.

Storage of records or source documents.—The original document is shelved by accession number. It is hoped that hard-to-get documents will be available on demand, although the cost is somewhat prohibitive. Conventional abstracts are filed according to accession number and await the results of a search.

Question analysis and development of search strategy.—Allan Rees, assistant director of the Center, in a paper for the American Documentation Institute conference in October 1963, makes some illuminating observations on the real problems of question analysis. He points out that there is frequently a distinction between:

1. What the questioner needs...
2. What he thinks he needs...
3. What he wants...
4. What he is prepared to read...
5. How much of what he gets he is prepared to read...
6. How much time he is willing to devote to it all...
7. In what sequence he would like to read what he gets...
8. What value he will attach to what he gets...10

The best method for determining the answers to the questions raised above is as yet unknown; no research has been done relating to the nature of the question-asking process, although increasing attention is being devoted by Rees and others at the Center to precise
identification of the areas of investigation. It is obvious though, in the light of our experience, that question analysis must be approached with a great deal of care.\textsuperscript{11}

The education project at the Center asks each questioner to:
1. state his question on three levels—specific, more generic, most generic;
2. define the terms in the question,
3. list those terms he associates with the question terms, and
4. describe the purpose of his research.

In instances where this outline is followed rigorously and completely, the Center's question analysts have a good beginning. The real problem is whether the questioner can define his research need so precisely. To further the more complete analysis of a question, telephone contact with the questioner is very desirable, and frequently used.

Once the analyst has what appears to be a complete statement of the question, the question is analyzed for searchable concepts; these are translated into the indexing language of the system and are organized so that they correspond to the logic of the question. Identification of searchable concepts involves the isolation of question concepts which correspond to the indexing concepts used by the system, and the addition of generic, specific, and associated concepts derived from the analyst's knowledge of the file or from conversations with specialists. One of the computer listings of the semantic code dictionary is arranged alphabetically by code so that the thesaural relationships established by the code are apparent.

The concepts thus identified are translated into the semantic code, and further structured by the application of appropriate role and level indicators.

In formulating the logical structure of the question program, the following connectives can be used.

\[
\begin{align*}
A \cdot B & = A \text{ and } B \\
A + B & = A \text{ or } B \\
A - B & = A \text{ but not } B
\end{align*}
\]

Any question therefore can be expressed as an algebraic polynomial of logical sums, products, and differences of semantic codes.

Let me briefly illustrate the search structuring by providing an example. The question submitted by a researcher is "Give me abstracts of all studies dealing with the use of educational media in teaching biology at the below college level." The concepts identified as "searchable" are media, biology, educational institution, and college.

\[
\begin{align*}
\text{Let } A & = \text{ media} \\
B & = \text{ biology} \\
C & = \text{ educational institution} \\
C^1 & = \text{ college}
\end{align*}
\]
Using the logical connectives we have,
\[ A.B.(C-C^1) \]

Applying the appropriate role indicators,

- KQJ = agent of process (by means of)
- KEC = subject taught
- KIS = location of population

the program becomes

\[ KQJ.A.KEC.B.KIS.(C-C^1) \]

Since KQJ.A must be associated with KEC.B and not with any other word of the telegraphic abstract, the level indicators must be added. Additional level indicators are then included to designate for the computer the precise grouping of all the terms to be searched.

- 4 level - a role indicator and the word to which it applies
- 5 level - a group of terms closely associated within the study
- 6 level - all words relating to the same study

Our complete program is:

\[
6\left[4\right.\left( KQJ.A \right) \bullet 4\left( KEC.B \right) \bullet 4\left[ KIS.(C-C^1) \right]^{4/6}
\]

Conducting of search.—The question program is keypunched and the question transferred to computer memory. The computer, a GE-225, compares the analytics of each document on tape with the analytics of the question and where they match prints out the document accession number.

Delivery of results of search.—Conventional abstracts corresponding to the accession numbers identified by the computer are pulled manually from the file and mailed to the questioner.

The above is intended as an elementary summary of the structure of the system. For a detailed explanation and analysis I refer you to the various Center reports listed in the references.

Current Research

A study recently completed for Cooperative Research has indicated several fruitful areas for research. The purpose of the first part of the study was to compare the relative effectiveness, in terms of relevancy and recall, of three different approaches to searching the file. The second part attempted to determine what differences,
if any, existed in the assignment of relevance by different evaluators of the same question. For Part I, twenty-four questions, selected from the more than 400 submitted during the initial year of the project, were used as the sample. The questions were programmed using three searching strategies: (1) narrow semantic code programs, using the maximum discriminatory features of the system, (2) broad semantic code programs, derived from the narrow programs by eliminating role indicators, by omitting conjuncts, by adding disjuncts, etc., and (3) faceted classification programs. Number 3 requires some explanation. Along with the semantic-coded telegraphic abstract approach, a machine searchable faceted classification, developed at the Center and based on the Tauber-Lilley faceted classification for media literature, was applied to all of the documents used for this investigation. It was felt that comparative testing would benefit from the application of a classification scheme different in concept from the semantic-coded, telegraphic abstract approach.

Responses to the questions were evaluated as relevant or peripheral by CDCR staff members and as relevant, peripheral, or nonrelevant by the questioner.

Although any conclusions about the comparative effect of the three searching strategies would be unwise because of the inadequate size of the sample, the first part of the study made several important recommendations.

1. The structure and application of the semantic code should be examined in more detail, to determine the desirability of modification.
2. Greater terminological control should be exercised in the telegraphic abstract and more attention should be devoted to the consistency of its preparation.
3. Intensive investigation should be made of the nature of question formulation and analysis.
4. An attempt should be made to establish more precisely the appropriate level of information content of conventional abstracts.
5. The faceted classification should be further developed to provide a suitable tool for researchers wishing to organize their own collections.

In Part II of the study, answers to fourteen questions were evaluated as relevant or peripheral by CDCR staff members and as relevant, peripheral, or nonrelevant by the questioner. Four of the fourteen questions were given two outside evaluations (by two questioners who posed the same question). The results of these evaluations indicated a wide variation in the assignment of relevance for a particular question between CDCR evaluators and the questioner, and a wide variation between two questioners who posed the same question.
Included in Part II was a preliminary investigation to determine whether relevant answers are characterized by some objective properties of relevance and whether those properties can be isolated. One very interesting aspect of this portion of the study was the application of probability theory to the problems of relevance. I refer you to the final report, since only a complete reporting of the method and the results would be of value.

Current research activities are based on the experience of the Center during the last three years and on the recommendations made by the effectiveness study outlined above. Key to the conduct of current research is the pilot user group. They will analyze and evaluate the system in four areas: (1) coverage (within the user's own subject area or field of interest), (2) usefulness (in relation to the user's own research needs), (3) relevance (of abstracts received in response to questions), and (4) recall (missed known answers).

On the basis of the questions submitted by the user group, the Center will analyze the system objectively in four areas: (1) further development of the techniques of question analysis, (2) revision and testing of telegraphic abstracting techniques, (3) comparison of relevance assessment by different evaluators (questioner, staff member, and expert), and (4) development of operational administrative procedures. As contact with the field of education has increased, most importantly through representatives of the USOE and the present user group, so has the necessity for expanded investigation. Plans are now being made for the expansion and extension of current research activities.

Future Research

From our current research we have selected three areas which we feel could contribute most, at this time, to the refinement and development of the system: coding, development of inclusion criteria, and conventional abstract preparation. Our experience in question programming and searching has revealed that some of the present codes are either incorrect (through human or machine error) or do not establish the desired thesaural relationships among terms. These codes must be corrected or revised. In addition, we wish to determine whether the development of additional semantic factors, elimination of some of the existing factors, or changes in conceptual meaning of existing factors, would have any appreciable effect on relevance and recall. Any code revision or modification will be tested under operational conditions.

Development of criteria for the inclusion of material in the file has been approached pragmatically by establishing a pilot user group,
by analyzing the needs of this group as expressed by their pilot questions, by scanning thousands of research studies, and by analyzing the citation patterns in well-known studies. As I indicated earlier, our current work is focused on the establishment of a comprehensive media and media-related research file. Although criteria must eventually be established for the total field of educational research, much will be gained by concentrating at this time on the media field. The Center is enlisting the help of an experienced media researcher to examine the problems of inclusion from a theoretical point of view, hopefully providing a rationale for the inclusion of media and media-related research. The practical experience of the Center will then be merged with the rationale to provide inclusion criteria for an operational file.

The preparation of conventional abstracts poses some as yet unexplored questions.

1. What level of information content of conventional abstracts is most appropriate to the needs of educational researchers?
2. What kinds of data should be included in abstracts and at what degree of specificity?
3. What is the effect of various levels of information content on the users' assessment of relevance?

A tentative experimental design has been worked out for an investigation of the above. The results of this experiment will contribute to the establishment of precise rules governing the amount and type of information to be included in a conventional abstract.

Conclusions

The fact that we have made and are continuing to make progress in no way implies that the system is ready for operation. Sufficient evidence of our realization that many tasks remain is given in the outline of current and future research. We are, however, confident that the system can be developed to an operational level.

Our experience tells us that we must proceed slowly, so that any operational service will have the full benefit of a concerted research effort. Research is extremely difficult when one is faced with the many day-to-day problems of operating a large information system.

The interest and criticism of educational researchers have been invaluable. Currently the most interested and most critical of these are the twenty members of the pilot user group. Without their help as users, critics, and advisors, our work in the education project would be much more difficult.
We have also benefited from the advice and counsel of representatives of the USOE. Without the support (both moral and financial) of the U. S. Office of Education, the project would have been impossible.

REFERENCES


Additional References


APPENDIX A: ROLE INDICATORS AND LEVEL INDICATORS

Role Indicators

<table>
<thead>
<tr>
<th>Role Indicator</th>
<th>Functional Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAB</td>
<td>Type of study</td>
</tr>
<tr>
<td>KIT</td>
<td>Date of study</td>
</tr>
<tr>
<td>KIS</td>
<td>Geographical or environmental location</td>
</tr>
<tr>
<td>KEJ</td>
<td>Population acted upon or studied</td>
</tr>
<tr>
<td>KAM</td>
<td>Process carried out on, by, or in relation to KEJ</td>
</tr>
<tr>
<td>KEC</td>
<td>Subject taught</td>
</tr>
<tr>
<td>KQJ</td>
<td>Agent of process (of KAM or KEC)</td>
</tr>
<tr>
<td>KWV</td>
<td>Attribute given</td>
</tr>
<tr>
<td>KAH</td>
<td>Condition of process</td>
</tr>
<tr>
<td>KUP</td>
<td>Attribute or behavior determined</td>
</tr>
<tr>
<td>KAP</td>
<td>Dependent variable; attribute or behavior influenced</td>
</tr>
<tr>
<td>KAL</td>
<td>Independent variable; influencing factor</td>
</tr>
<tr>
<td>KEW</td>
<td>Person interviewed or answering questionnaire</td>
</tr>
<tr>
<td>KWC</td>
<td>That toward which an attitude is noted</td>
</tr>
<tr>
<td>KWJ</td>
<td>Device or material prepared</td>
</tr>
</tbody>
</table>

This list comprises all the role indicators used in the TA. Their sequence and use in a TA are dependent on the characteristics of the individual document.

Level Indicators

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Space ()</td>
<td>To separate two or more role indicators on a single line of the TA.</td>
</tr>
<tr>
<td>(,)</td>
<td>To separate a role indicator from the word or words to which it applies.</td>
</tr>
</tbody>
</table>
To separate one role indicator—word(s) combination from the next.
To separate one group of related role indicator—word(s) combination from the next.
To separate groups of unrelated role indicator—word(s) combinations from each other.*

*In instances where a document contains or discusses two or more unrelated or loosely related experiments or surveys.

Discussion

William P. McLure*

I have been asked to react to the usefulness to the field of educational research of the project which Barhydt describes.

His purpose is laudable, and the general idea is clearly stated. A tremendous amount of careful work is evident. Barhydt shows proper restraint and modesty in describing the project. He states that the system is not operational, but he expresses confidence that it can be developed to this stage.

In such an early stage of research and development, one can only speculate on the usefulness of this system. I am sure Barhydt would agree that now we can apply primarily the tests of logic and common sense to the probable usefulness. As I read the paper I feel a recurring desire to talk with some of the members of the pilot user group. One of them and not I perhaps should be making this reaction. I am sure that as the project develops the experience of the user group will be evaluated constantly for feedback into the system.

While Barhydt's paper quite properly concentrates on the technical aspects of the information retrieval (IR) system, the ultimate test will be its usefulness. It must meet certain needs of the user so well that its expense and operation are justified.

My bias is strongly hopeful that a useful system can be developed. Comprehensive services of bibliographic control and abstracting would be invaluable in the field of educational research to both

*William P. McLure is Director, Bureau of Educational Research, University of Illinois, Urbana.
producers and consumers. But these are only steps to aid in the identification of materials. What ultimate goal is contemplated? Barhydt alludes to the problem of "uncertain dissemination" of educational research literature and to absence of comprehensive collections. Does he envisage a system of reproduction of materials in the original form? Can the system, however, comprehensive, be made available to the thousands of centers for use, such as college and university libraries, public school libraries, public libraries, and others?

It seems to me, therefore, that the ultimate test of the system is its value to the user not only to identify materials but to make careful discriminations. I am concerned about two groups of users, the researcher or producer and the general consumer. The former is also a consumer but a special one. These two groups have different needs. Indeed within each one there is a wide range of needs.

Is the approach to the development of the IR system grounded in theory of learning and human behavior? Or is it dominated by criteria which satisfy theories of mathematics and electronics primarily and only secondarily those of learning and behavior? For example, how much knowledge of the system does the researcher, and the general consumer, need in order to use it effectively? To what extent are his intellectual processes structured in the use of the system?

I am particularly concerned about the researcher and the demands of the system on him to state or to define his research need in the earliest stage when he is attempting to create or to formulate an idea into researchable form. At this time he is engaged in the act of structuring something out of nebulous thought. How much does the system demand of him in this situation? In this connection, the role of the analyst needs to be elaborated. It seems that this person may have a key role in assisting the questioner and perhaps in making judgment about selections.

If my inference is correct, is it not true then that this system may lead to further development or modification of the role of the librarian? I get the impression that this project is of necessity centered at the moment on the warehousing and transportation function of librarianship. But the performance of the system must facilitate the higher functions of interpretation, consultation, advisement, guidance, and specialized forms of teaching. If this is true, my earlier assumption that the development of this system must proceed over bridges of research on use is correct. I am wondering, therefore, to what extent research with the use is made an indigenous part of the process of development of the system? For example, the statement is made "Development of criteria for the inclusion of material in the file has been approached pragmatically by establishing a pilot use group, ..." What does the term "pragmatically" mean? Does it mean that the approach is limited to a priori knowledge of the group of users? Or is research being done on the experience of users with
the system? If so, then, I should like to know what is being built into the system and what results are being obtained from the experience of users.

I am sure that Barhydt and his colleagues face many puzzling problems in this venture. For example, in the preparation of abstracts, it is not clear whether he is selecting a narrow field and concentrating on it, sampling from a broad range of materials, or taking everything available to him. The three questions in the paper suggest that he is assuming a big responsibility of deciding what is "appropriate," "how specific," and "relevant."

I cannot see how the problem of choice or selection can be avoided, given the volume and range of material. Much of the literature on research, for example, as in other fields, reflects the advancement of people as well as of knowledge. What may be new to the neophyte may not be new to a field of knowledge. Each may have ample justification for publication but differential demands for use. Thus it seems that the IR system should make it possible to improve the rationality of choice which now exists in the selection of materials.
DATA PROCESSING PROBLEMS AT THE DEFENSE DOCUMENTATION CENTER

William A. Barden

This account deals with the trials and tribulations encountered at the Defense Documentation Center (DDC) during the development stages—and later the expansion stages—of a data processing system for indexing and retrieving scientific and technical documents. But, before reviewing various aspects of the Center's computer operations, some highlights concerning the mission, organization, and functions of the Center should be presented.

DDC provides a central service for the interchange of scientific and technical documentation for the Department of Defense (DoD). The Center receives, stores, and announces all technical reports prepared as the result of Defense research, development, test, and evaluation activities. As you know, the cost to the Federal Government for Defense research and development activities is approximately seven billion dollars annually. DDC provides copies of the research and development reports to the entire Defense community on a secondary distribution basis. Other DDC services include the functions of providing bibliographic searches and maintaining a file of R&D current tasks of the DoD. The Center provides these services to the Defense Community at no cost to the user. Reports which are free of security or proprietary restrictions are released for sale to the public through the Office of Technical Services in the Department of Commerce.

Until March 1963, DDC was known as the Armed Services Technical Information Agency (ASTIA). The ASTIA was formed in 1951 and was assigned to the operational control of the Air Force to provide an effective service for all DoD components seeking copies of reports derived from Defense research and development.

DDC moved to Cameron Station, Alexandria, Virginia, in July 1963, after five years at Arlington Hall Station in Arlington, Virginia. Four months later, the operational control of DDC was transferred from the Air Force to the Defense Supply Agency. The move was made as a part of DoD's rapidly developing technical information program and was designed to provide a more direct channel of

William A. Barden is Directorate of Automated Systems and Services, Defense Documentation Center, Alexandria, Virginia.
communication through which DDC could provide wide documentation services to its DoD. Headquarters for the Defense Supply Agency are also located at Cameron Station.

The Defense Documentation Center is one of twelve major field activities of the Defense Supply Agency, which reports directly to the Secretary of Defense. As with the earlier organization, ASTIA, DDC continues to receive policy direction from the Director of Defense Research and Engineering. DoD's Director of Technical Information, Walter M. Carlson, is currently responsible for policy direction of the DDC program. His responsibilities were not affected by the shift of DDC from the Air Force to the Defense Supply Agency.

Concurrent with its transfer to the Defense Supply Agency, DDC was reorganized with Dr. Robert B. Stegmaier, Jr., assigned as Administrator. With its new organization, DDC has a rather typical staff and line structure (see Fig. 1). The latter consists of the three operating directorates as follows: Directorate of Document Analysis and Processing, Directorate of User Document Services, and Directorate of Automated Systems and Services.

The Directorate of Document Analysis and Processing handles input to the DDC system. This Directorate (1) performs accession, selection, review, cataloging, scientific analysis and related
processing of documents received and incorporated into the collection for announcement and services, (2) develops and controls the DDC Thesaurus of Descriptors and the DDC Thesaurus of Identifiers, and standardizes a technical vocabulary compatible with automated storage and retrieval systems, (3) monitors the application of approved terminol-ogy in the announcement, storage, search, and retrieval of docu-
ment references, and (4) organizes and provides information and indexes of current DoD research, development, test and evaluation programs.

The Directorate of User Document Services handles the output of the system. This Directorate (1) publishes an abstract-index journal, bibliographic tools and other media to announce the existence, accessibility, and availability of documents in the DDC collection to authorized users, (2) provides reference services requested by authorized users, including documents, references to documents, and referral to other document sources and Information Evaluation Centers as appropriate, (3) provides graphic arts and reproduction processing, and (4) operates seven regional Field Offices to provide extended user document services.

The Directorate of Automated Systems and Services handles the internal massaging of data and management information in the DDC system. This Directorate plans and operates the automatic data processing (ADP) services in support of DDC documentation operations.

**Background of the Automatic Data Processing System**

In February 1960, ASTIA began operational use of its first automatic data processing system (ADPS). It has had a 100 per cent increase in work load for its original applications. These consisted of request processing, inventory control, indexes for the Technical Abstract Bulletin, and accountability records for security classified documents. We had hoped to experiment with information retrieval, but we had no basis for computing the load which it would represent. Our experimental work in information retrieval, however, was sufficiently successful that we were running bibliography searches operationally by January 1961.

During 1961, the document request processing work load increased from 500,000 to 700,000 per year. The bibliography work load jumped from 1,300 to 2,500 requests per year. In the same period, the Office of the Director of Defense Research and Engineering (ODDR&E) requested that we implement the cataloging, storage, and retrieval of information on current tasks which are represented in the Research, Development, Test and Evaluation (RDT&E) Basic Research Projects. This amounted to some 7,000 records and was to
be followed by the Applied Research Projects and the Development and Test Projects, building to a total of about 40,000 such records per year. Again, there was no way to estimate the work load in terms of searches to be run. We assumed 1,000 searches per year and were sure this figure would not be exceeded because all requests had to be approved by the ODDR&E.

At almost the same time the Department of Defense issued a directive requiring automatic time-phased downgrading of its classified information except for a few special categories. In this case "automatic" simply meant the action was to be taken at a given time with respect to the date of the report in each case, e.g., certain secret reports were to be downgraded to confidential at the end of three years while other reports are downgraded at different time intervals. At that time that portion of our collection which was under ADP control amounted to about 250,000 reports of which about 80,000 were classified either confidential or secret. It was fortunate that we had our ADPS. We designed a program which would accomplish the necessary downgrading, and this application has been running monthly ever since.

No attempt has been made in this paper to cover the myriad activities required to collect and to validate the information needed to create the many files which are essential to such a variety of system applications. My purpose up to this point has been to portray the very rapid build-up of work load much of which could not have been anticipated with any degree of accuracy if, indeed, it could have been anticipated at all. In July 1961, I wrote a paper entitled "ASTIA's Retrieval System: An Interim Appraisal." In the final paragraph, I stated:

Since we have come this far, one might assume that we could relax a bit. However, such is not the case. Already, work loads for which we had planned, augmented by some which we did not foresee, are rapidly approaching the absolute limits of our present ADPS. We expect to install supplemental equipment later this year [1961]. But even this is only a stop-gap measure. We are now actively planning for a much more powerful system. We hope the next one will be capable of all the expansion we may need. At this point in time we know we will be assigned additional responsibilities, and we are reasonably certain that whatever our next ADPS it will have to be expanded from time to time in order to keep pace with requirements.1

Feasibility Study of the Expanded System

We started our feasibility study in April 1961 by preparing specifications covering the applications and the anticipated work
loads. At the same time we requested the Air Force Systems Command (AFSC) to provide us a list of manufacturers whose equipment might satisfy our requirements. Late in May we received the list and mailed invitations and specifications to thirteen manufacturers. Late in June we held our first briefing. However, our situation was changing so fast in terms of both applications and work loads that it was necessary to hold additional briefings and question and answer periods.

From our experience, we knew one of the most important things to be done was to determine system requirements before doing anything else. Our next step was to evaluate the proposals for supplying the equipment, make our own selection, and forward our feasibility study with our evaluations and recommendations to Headquarters USAF through AFSC for approval. At that point in time, December 1962, we settled down to the seemingly endless series of tasks that must be accomplished in preparing for any automatic data processing system.

Make-Ready

Preparing for a large scale system on an extremely tight schedule is a major undertaking. It was necessary to:

1. Maintain existing programs and operations on the two UNIVAC Solid State 90's.
2. Accomplish systems analysis and design for all the planned applications.
3. Schedule and conduct training for systems analysts, programmers, and operators of the new ADPS.
4. Plan and schedule site preparation.
5. Provide for necessary conversion of existing files and creation of additional files.
6. Write, test, and debug the programs as individual programs.
7. Conduct a system test which involves all the hardware, software, and the operating programs as an integrated system of applications.

The supplemental equipment mentioned in the background section above was a UNIVAC STEP system which could accomplish some portions of the existing applications. It was installed in October 1961, and within a month a request was made to upgrade this system to another full scale UNIVAC Solid State 90 identical to the first one. This was approved, and the second Solid State 90 became operational in March 1962. This was necessary in order to be able to cope with the machine loads.
The ADPS Milestones Chart (see Fig. 2) shows the schedule that was worked out. Although the feasibility study was formally concluded in November 1962, it really did not end there. Additional requirements were being imposed which meant reanalysis of the ADPS to assure adequate hardware capability. As a result the actual feasibility study was not terminated until July 2, 1963 (some twenty-seven months after it was started). Official approval of our study for the procurement and installation of a UNIVAC 1107 Thin Film Memory Computer for DDC came on July 22, 1963.

### ADPS Milestones

<table>
<thead>
<tr>
<th>EVENTS</th>
<th>MAR</th>
<th>APR</th>
<th>MAY</th>
<th>JUN</th>
<th>JUL</th>
<th>AUG</th>
<th>SEP</th>
<th>OCT</th>
<th>NOV</th>
<th>DEC</th>
<th>JAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feasibility Study</td>
<td>1 APR 61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 JULY 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personnel Training (Operators &amp; Programmers)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 NOV 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systems Design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 JULY 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Programming</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 NOV 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>File Creation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6 NOV 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Debugging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 DEC 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Site Preparation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 NOV 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Installation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25 DEC 63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Operational Checkout Period</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>15 JAN 64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2
Automatic Data Processing System Milestones Chart

Personnel training for systems analysts, programmers, and operators started early in March 1963. The training of analysts and programmers was completed by mid-July. Computer operators were trained during the time remaining until November 15. The broken line beyond that point merely acknowledges the fact that training is an ongoing process that never terminates. Personnel who have kept the former system going must be retrained to operate the new system. New personnel have to be indoctrinated in the operations (and mission) of the organization and then trained in programming or operating the ADPS.

The systems analysts were the first to receive training. This was completed during March 1963. From April 1 through July 15,
1963, they designed the applications to be programmed. This involved frequent discussions with and briefing of personnel in the operating divisions to assure mutual agreement on the capabilities to be programmed into the system. Mutual agreement on what is to be designed into the system is absolutely essential to the operational success of any ADPS. Of equal importance is agreement on the data and information to be incorporated into the various files including the format of the data and information which are to be furnished as input to the ADPS. A corollary requirement is agreement about the format and content of the data and information to be provided as output by the ADPS. Some of the details as well as the broad general requirements can be worked out during the system design phase.

However, many of the details cannot be settled until programming is underway. With major aspects of the system design frozen and programmer training completed as of July 15, 1963, the task of writing the programs began in earnest. DDC's staff of programmers was not adequate to accomplish the needed programs in the time allocated. This brings me to the next major principle, i.e., prepare your own staff to the maximum extent possible. Requirements must be established, and personnel must be secured and trained. The overall program package was divided into four parts, and an individual in each of four groups was designated to be responsible for the output of his group. The Chief of the Systems and Programming Division had the awesome task of coordinating the groups as well as planning for file conversion, file creation, and debugging.

A crucial component of the system was a Master Accessioned Document (MAD) file which we appropriately referred to as the MAD file. The creation of this file was such a stupendous task we could not hope to accomplish it ourselves. Consequently, a contract was let for the creation of this file. The file itself will be described subsequently. Since this discussion has to do with milestones, it is sufficient to say that the file creation slipped beyond the established target date of November 6, 1963. The target date was changed to January 31, 1964, by mutual consent of DDC and the contractor. The file covers some 350,000 documents, all those processed by ASTIA and DDC from March 1, 1953, through February 1964.

Almost two man-years of effort were invested in a review of punched-card information and in standardizing the form of entry before releasing the files to the contractor to be used in file creation.

When the MAD file was completed and sample printouts were made, some startling problems were discovered. Many errors of various types were found. Some were a matter of incorrect information; others were a matter of omission. It was decided that a detailed review of sample printouts was required. This led to the conclusion that the information must be obtained from other sources. In some cases, it meant going back to the original work sheets for the data.
The MAD file had to be redone. Since it is the source of information for the inverted (retrieval) files in the UNIVAC 1107 system, the effects were reflected in those files. This unforeseen development extended the transition from the Solid State 90 to the 1107 by two months.

When we converted the Field-of-Interest Register (FOIR) file, we found discrepancies that had to be checked against the master file of FOIR's. This delayed the transfer of document request processing to the 1107 by a month. An orderly transition under these circumstances was impossible. For anyone who considers a data processing system for the first time or is moving up to a larger system, I have two recommendations: (1) Before installing your own system, do not be too optimistic in scheduling the operational date, and (2) Arrange to rent time needed to check out files for accuracy so that programs can be run against files of known validity. This brings me to the next point.

Debugging of programs is the only other task which slipped beyond the target date. As fast as programs could be written and desk-checked, they were keypunched and transmitted to UNIVAC at St. Paul, Minnesota, by means of a UNIVAC 1004 Card Processor. At St. Paul, a 1004 received the program data and produced punched cards which were processed by the UNIVAC Data Processing Center on an 1107 by means of their compiler known as SLEUTH. After compiling the programs, the 1107 prints out in parallel the machine coded instruction on the left and the instruction as written by the programmer on the right. SLEUTH also identifies errors and codes them as to type on the line in which an error occurs.

Site preparation was a rather simple matter. It had been worked out prior to the move of DDC to Cameron Station. The period shown on the chart for about one month ending November 15, 1963, was for additional electrical power cables that were required for the 1107.

Installation was a somewhat different matter. For the final six months or so prior to delivery of the 1107, we were running the two Solid State 90’s twenty-two hours each per day. One of the Solid State 90’s had to be removed in order to make room for the 1107. We managed to struggle along with normal operations by contracting with UNIVAC for fourteen hours per day on a Solid State 90 at their New York Data Processing Center. We placed a complete set of tapes representing our retrieval file in their hands and transmitted bibliog-raphy requests to New York via the 1004. They ran the searches at night and transmitted the results to us the next morning. The 1107 was completely installed except for one high speed printer and print control synchronizer by the target date of December 15, 1963. The remaining printer had been delayed because at the last minute we decided we could have more effective use of the printers if we had an
additional synchronizer. These two units were installed before the end of December.

The operational check-out period was started on December 15, but was subsequently delayed because of some bugs in the software. This should not have surprised anyone because our configuration was such that it had to be completely installed before there could be any degree of certainty that all the bugs had been identified. In the meantime we were able to start checking out our programs as part of an integrated system instead of one at a time as had been the case at St. Paul.

File Creation

The retrieval files designed for the 1107 are listed alongside those for the Solid State 90 (see Fig. 3). For the latter we had four inverted (subject type) retrieval files covering the document collection. These were the coded descriptor file, the coded asterisk descriptor file, the coded identifier file, and the uncoded identifiers.

OLD SS90 FILES

NEW 1107 FILES

<table>
<thead>
<tr>
<th>SEPARATE INVERTED RETRIEVAL FILES</th>
<th>NEW INVERTED FILE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODED DESCRIBTOR</td>
<td>(UNCODED AND CODED)</td>
</tr>
<tr>
<td>CODED ASTERISK DESCRIPTOR</td>
<td></td>
</tr>
<tr>
<td>CODED IDENTIFIERS</td>
<td></td>
</tr>
<tr>
<td>UNCODED IDENTIFIERS</td>
<td></td>
</tr>
<tr>
<td>RDT&amp;E (DD613) SEARCHABLE FILE</td>
<td></td>
</tr>
</tbody>
</table>

RDT&E (DD613) DIRECT NON-SEARCHABLE FILE

- PROVISIONS FOR SYNONYM CAPABILITY
- PROVISIONS FOR HIERARCHY CAPABILITY
- PROVISIONS FOR ROLES
- PROVISIONS FOR LINKS
- PROVISIONS FOR EXPANDED WEIGHTS
- PROVISIONS FOR STIC REFERRAL SERVICE
- PROVISIONS FOR OTHER COLLECTIONS

- PERSONAL AUTHOR
- SOURCE
- SOURCE ACRONYM
- SOURCE SERIES NUMBER
- MONITORING AGENCY ACRONYM
- MONITORING AGENCY SERIES NUMBER
- CONTRACT NUMBER
- SERIAL NUMBER AND DATE
- PROJECT NUMBER
- TASK NUMBER

- REPRESENTS PROGRAMMED CAPABILITY AND FILE DESIGN TO PERMIT EXPANSION

Figure 3

Retrieval Files for the Solid State 90 System and for the UNIVAC 1107 System
In addition, when we designed the RDT&E management information application, we established a searchable (inverted) file very similar to the retrieval file for the document collection. However, it differed from that for the document collection in that source, monitoring agency, principal investigator, and other approaches could be used in addition to subject for retrieval. Another point of difference was in the nonsearchable (direct) file for the RDT&E. This file is organized by accession number and contains extensive information about each project. This information is broken out into specific segments each of which is identified by a code number. The initial result of a search in the searchable file is a list of accession numbers of projects which fit the search specifications. The next step is locating the project information by accession number in the non-searchable file and selectively printing out the segments of information required to fill the request. The concept of the MAD file for the 1107 is an outgrowth of the direct file for RDT&E.

For the 1107 we designed the retrieval files along the same lines as those for the Solid State 90. There was one important exception. In our desire to capitalize on the capabilities of the 1107 we decided not to substitute numeric codes for the descriptor word statements even though codes had been planned for a very good reason. This was considered no hardship for us in retrieval, but it could be a deterrent in terms of the use that our customers might make of the file. On re-examining the total application, the good reason for using numeric codes was rediscovered. The reason was that in storing the MAD file in our mass storage we had planned to encode all information that could be regenerated by a simple file look-up in order to minimize the storage requirements. For example, it was estimated that by using numeric codes in place of the descriptors in the MAD file we could save about 140,000,000 characters of storage in terms of the present collection. This is more than 25 per cent of our total mass storage. The creation of the new file for retrieval on the 1107 is still giving us problems. It is one of the most complex tasks in going from Solid State 90 to the new system. This problem and debugging which cannot be completed until after the retrieval file is completed are the principal reasons for build up of backlogs in data processing during the changeover.

During the system design phase, while working very closely with the operating divisions, we decided to design into the system a number of provisions which will ultimately permit a more sophisticated retrieval system. These provisions include the capability to use: synonyms instead of a manual look-up of a precise descriptor, hierarchy based on computer search of more specific terms to create artificially higher general roles for precise specification of a search term, links for precise designation of terms which really are related as applied to a given report, expanded range of weights (at
present the asterisk carries a weight of three, absence of an asterisk gives the descriptor a weight of zero), referral service, and search of other collections.

Profiting from our experience with the RDT&E files, we decided to create for the document collection inverted files as follows: subject, author, source (corporate author), project number, task number, and contract number. These files greatly enhance our ability to tailor bibliographic searches to the users' needs.

In Figure 4 are shown two punched-card forms which are designed to create a file that could be searched for duplicate checking and request identification on the Solid State 90 system. The card at the bottom of the Figure was created for documents cataloged from 1953 to 1961. The card at the top was adopted in 1961 and was used until August 1963. The earlier form lacked many items of information that were required for the intended applications. On the other hand, the newer form lacked information concerning personal author, but it did provide information required for the automatic-time-phased downgrading plus information for the inventory file.

**INPUT STANDARDIZATION**

![Punched Cards Formerly Used for Input into the Solid State 90 System](image-url)
These cards provided some input for the MAD file. In February 1961, we began using punched paper tape equipment for producing copy for our announcement bulletin and for catalog cards. Thus, these tapes provided input for abstracts for reports processed during the period February 1961 to mid-August 1963. Information that could not be obtained from the punched-card files or the punched-paper tape had to be keypunched for the period March 1953 to mid-August 1963. At different times during this period, decisions were made to pick up additional items of information. For some of the times, it would have been too costly to go back and start at the beginning. Hence, when bibliography printouts are made, there will be differences in the amount of information available in different time periods—From February 1961 on, all items in each entry of the MAD file are provided for in the punched-paper tape as indicated in Figure 5.

DATA KEYPUNCHED FROM CATALOG CARDS

<table>
<thead>
<tr>
<th>Authors and Title</th>
<th>1953 - 1963</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry Classification</td>
<td>1953 - 1963</td>
</tr>
<tr>
<td>Abstracts</td>
<td>1961 to-date</td>
</tr>
<tr>
<td>PPT.</td>
<td></td>
</tr>
</tbody>
</table>

All Items

15 Aug 63 to-date

Figure 5
Input Items Currently Used in UNIVAC 1107 System

Prior to releasing information to the contractor for creation of the MAD file, it was necessary to analyze the variations in cataloging which had occurred during a period of more than ten years. Figure 6 shows the change in acronym during that time for a major research and development activity located at Wright-Patterson Air Force Base,
Ohio. As shown, that activity was successively identified as Wright Air Development Center (WADC), Wright Air Development Division (WADD), and Aeronautical Systems Division (ASD). Such variations must be identified, and information standardized in order to produce an effective machine system. In standardizing the information, we adopted the most recent acronym and reflected it back through earlier portions of the file. Similarly, any variations in format of originator’s report numbers were considered, and decisions were made as to the format we would adopt for our files. The files must be established with precision, and input for duplicate checking and request identification must meticulously adhere to the same standards.

The data field number, field name, and period for which each item is available for the MAD file are shown in Figure 7. Of these, Number 10-personal authors, Number 13-originating agency acronym, Number 15-contract number, Number 16-project number, Number 17-task number, Number 18-monitor agency acronym, Number 23-descriptor set, and Number 25-identifier set are used as the base for inverted files for information retrieval applications. Other items such as Number 2-FOIR (Div/Sec), Number 20-report classification, Number 29-inventory, and Number 33-limitation (code) are used in the request processing files. Checks in the columns on the right indicate the items of information that have been recorded in machinable form for various periods in the past eleven years.

DDC’s UNIVAC 1107 Equipment Configuration is shown in Figure 8. Every 1107 has 128 words of thin film control memory. The principal memory of DDC’s central processor consists of 49,152 words of magnetic core. To this central processor are connected (1) eight FASTRAND units with a total capacity in excess of 500 million characters used for the MAD file and Inverted Files, (2) two FH 880 drums with a capacity of almost 9 1/2 million characters used as working storage for FOIR tables, User File, sorting, merging,
### Master Accessioned Document File

<table>
<thead>
<tr>
<th>DATA FIELD NO.</th>
<th>FIELD NAME</th>
<th>PERIOD ACCESSION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>DDC ACCESSION NUMBER</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>FOIR (DIV/SEC)</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>CITATION CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>ANALYST INITIAL/OTS PRICE</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>ORIGINATING AGENCY NAME</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>UNCLASSIFIED TITLE OF REPORT</td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>CLASSIFIED TITLE OF REPORT</td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>TITLE CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>DESCRIPTIVE NOTE (PROGRESS REPORT)</td>
<td></td>
</tr>
<tr>
<td>10.</td>
<td>PERSONAL AUTHORS</td>
<td></td>
</tr>
<tr>
<td>11.</td>
<td>DATE OF PUBLICATION</td>
<td></td>
</tr>
<tr>
<td>12.</td>
<td>PAGINATION</td>
<td></td>
</tr>
<tr>
<td>13.</td>
<td>ORIGINATING AGENCY ACRONYM</td>
<td></td>
</tr>
<tr>
<td>14.</td>
<td>ORIGINATING AGENCY REPORT NUMBER</td>
<td></td>
</tr>
<tr>
<td>15.</td>
<td>CONTRACT NUMBER</td>
<td></td>
</tr>
<tr>
<td>16.</td>
<td>PROJECT NUMBER</td>
<td></td>
</tr>
<tr>
<td>17.</td>
<td>TASK NUMBER</td>
<td></td>
</tr>
<tr>
<td>18.</td>
<td>MONITOR AGENCY ACRONYM</td>
<td></td>
</tr>
<tr>
<td>19.</td>
<td>MONITOR AGENCY REPORT NUMBER</td>
<td></td>
</tr>
<tr>
<td>20.</td>
<td>REPORT CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>21.</td>
<td>SUPPLEMENTARY NOTES (COPYRIGHT CITATIONS)</td>
<td></td>
</tr>
<tr>
<td>22.</td>
<td>AVAILABILITY (LIMITATIONS)</td>
<td></td>
</tr>
<tr>
<td>23.</td>
<td>DESCRIPTOR SET</td>
<td></td>
</tr>
<tr>
<td>24.</td>
<td>DESCRIPTOR SET CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>25.</td>
<td>IDENTIFIER SET</td>
<td></td>
</tr>
<tr>
<td>26.</td>
<td>IDENTIFIER SET CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>27.</td>
<td>BODY OF ABSTRACT</td>
<td></td>
</tr>
<tr>
<td>28.</td>
<td>BODY OF ABSTRACT CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>29.</td>
<td>INVENTORY</td>
<td></td>
</tr>
<tr>
<td>30.</td>
<td>ANNOTATION</td>
<td></td>
</tr>
<tr>
<td>31.</td>
<td>ANNOTATION CLASSIFICATION</td>
<td></td>
</tr>
<tr>
<td>32.</td>
<td>GROUP NO. (AUTO. PHASE DOWNGRADING)</td>
<td></td>
</tr>
<tr>
<td>33.</td>
<td>LIMITATION (CODE)</td>
<td></td>
</tr>
<tr>
<td>34.</td>
<td>SERIAL NO. (NO. OF PROGRESS REPORT)</td>
<td></td>
</tr>
<tr>
<td>35.</td>
<td>DOWNGRADING AUTHORITY</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 7**

Coverage of Items by Years for Documents in the MAD File

assembly and processing, (3) four printers each of which has 102 characters consisting of upper and lower case alphabet, Greek alphabet, punctuation marks, and other special symbols and operates at 600 lines per minute for printing hour demand bibliographies, (4) a paper tape reader (for input) and punch, (5) a card reader, (6) a card punch, (7) two UNISERVO III magnetic tape drives, (8) twelve
Figure 8

DDC UNIVAC 1107 ADP Equipment Configuration
UNISERVO IIA magnetic tape drives, and (9) six UNISERVO IIA magnetic tape drives.

Why so many magnetic tape drives of three different types? The UNISERVO IIIC's provide compatibility with other manufacturers' systems. These drives will be used for reading files created on other systems and for writing copies of DDC retrieval files for users who have data processing systems and want to run their own searches.

The UNISERVO IIA's are the working drives of the system. They read or write numeric data at 200,000 characters per second. These drives are used for compiling indexes for individual issues of the Technical Abstract Bulletin, as well as quarterly and annual indexes, all of which are to be created by the 1107. Further, these tape drives are used for updating the FASTRAND files. At each updating the complete information is retained on IIA tapes for use in rewriting the record on the drums if this should be necessary before the next scheduled updating. A task yet to be done is programming the 1107 for retrieval using the tapes instead of the FASTRANDS against the contingency of an extended period of downtime on the FASTRANDS.

The UNISERVO IIA's are low density, low speed drives which give us compatibility with our present tapes and are ideally suited for storing information which is to be printed out on the printers.

Part of DDC's computer installation is shown in Figure 9. It

![Figure 9](#)

*General Layout of the DDC Computer Installation*
would be difficult to represent the installation adequately with one photograph. However, this one does show the general layout.

Integrated Program Package and Applications

The conceptual organization of the program package is shown in Figure 10. So many files and runs are involved, it would be impossible to use lettering large enough to assure legibility of individual items. My purpose in using it is to depict the four major areas, i.e., Updating and Document Accessions (UA), Request Processing (RP), Retrieval (IR), and Indexes (Ind). The heavy lines connecting portions of the different areas portray the highly integrated character of the application. All updating information covering documents processed into the system is introduced via the Updating and Document Accessions programs. This portion maintains the MAD file. The Field of Interest File and Inventory File are essential features of the Request Processing run. The Inventory File is updated from the UA file for information as to releasability of individual documents. The Field of Interest File plus the transactions and accountability portions of the Inventory File are updated during the Request Processing runs. The Retrieval programs search the inverted files and within the limits of the requester's need-to-know as determined from the Field of Interest File provide for selective printout of information from the MAD File following the search. The programs for Indexes working against the inverted files in the Retrieval area and the MAD File in the UA area periodically compile indexes to the Technical Abstract Bulletin (TAB).

Speaking of indexes, a major study has just been completed by DDC. For each issue of the TAB there will be subject, corporate author, and personal author indexes bound separately from the TAB. An AD Locator Index will continue to be bound in each issue of TAB. On a quarterly basis the subject, corporate author, and personal author indexes will be cumulated, and a contract index will be added. Annual cumulations will be published of all these indexes. Phasing in of the indexes has not been completed. The details of the various indexes such as content and format would justify a paper exclusively devoted to indexes. I mention them here only because all the indexes will be compiled by computer based on data provided initially by human analysts. I think that it is important to make the point that this is not machine indexing. Someday, when we have a better understanding of linguistics and when machine translation is an established fact, we shall be able to go into machine indexing. In the meantime, we must be content with machine compiled indexes based on input from human analysts.
One program which is part of the UA package is duplicate (dup) checking of incoming documents. In this application we assign an Accessioned Document (AD) number to each report as it is received. This is a tentative assignment, and the number is used for control purposes. For each distinct report the AD number, the source acronym and report number, personal author, monitor name of acronym, and report and date of report are recorded on punched paper tape.

Dup checking of any document can be accomplished by matching one of the following combinations of data elements (in the preference indicated) against information already in the file: (1) monitor acronym-series, (2) source acronym-series, (3) contract-series, (4) contract-serial-date, and (5) source name-series. However, a related application is document identification which involves identifying the AD number of a requested document when the requester has furnished descriptive information only. As with dup checking, document identification can be accomplished by matching one combination of data elements in the appropriate Inverted File. Since we cannot assume that every requester will furnish the same combination of information for a requested document, the dup check input operation must capture all identifying information. A few other elements of information are recorded at the time of key-boarding information for dup checking. These elements are security classification, special limitation, subject division and section, and number of copies received. Picking up this additional information at the beginning of the pipe line permits DDC to fill requests for reports at the earliest possible time. In the past, it has been necessary to refuse to fill requests for reports until after they had been announced because information concerning releasability was not incorporated into the machine system until that point in time. Thus, scheduling of input to the machine system can be critical in a system serving many users.

The bulk of input information is recorded on a worksheet shown in Figure 11. Each element of information is numbered, and the element number is recorded along with the information itself. These numbers correspond to the field numbers in Figure 7 and constitute the basis on which information is added to the appropriate files in the updating run. These numbers also are the basis for selective print-out of information following a bibliography search. Figure 12 shows a new version of the worksheet designated as DD 1473. Its use is promulgated by DoD instruction 3200.8, dated February 18, 1964, Subject: Standards for Documentation of Technical Reports under the DoD Scientific and Technical Information Program. It is anticipated that the DD 1473 will result in expediting the input analysis of reports because the originators can do much of the work. This should not impose any additional work load on the majority of originators because most of them catalog their reports for their own files anyway.
Figure 11
Worksheet for Recording Input Information
The phenomenon of grain boundary hardening has been explored for the CaCl structure intermetallic compound NiGa. NiGa has a homogeneity range of a few percent and it was possible to examine the effect of stoichiometry upon the grain boundary hardening due to preferential oxygen diffusion down grain boundaries. While some grain hardening was noticeable just below 50% Ga, the effect was much less pronounced than at 52% Ga. It was possible to estimate both bulk diffusion and grain boundary diffusion rates for oxygen. The results suggest that hardening is due to lattice distortions which arise from the formation of a Ga-O complex. (Author)
Unclassified

Securities Classification

<table>
<thead>
<tr>
<th>KEY WORDS</th>
<th>LINK A</th>
<th>LINK B</th>
<th>LINK C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory</td>
<td>8 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grain Boundaries</td>
<td>8 9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hardening</td>
<td>8 7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Hardness</td>
<td>8 7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Penetration</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Microstructure</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Crystal Lattice Defects</td>
<td>8</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Intermetallic Compounds</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Imperfections</td>
<td>3 6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Controlled Atmosphere</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooling</td>
<td></td>
<td>10 6 2</td>
<td></td>
</tr>
<tr>
<td>Quenching (Cooling)</td>
<td></td>
<td>10 6 2</td>
<td></td>
</tr>
<tr>
<td>Aging (Materials)</td>
<td></td>
<td>6 3</td>
<td></td>
</tr>
<tr>
<td>Best Treatment</td>
<td></td>
<td>10 1</td>
<td></td>
</tr>
<tr>
<td>Diffusion</td>
<td></td>
<td>10 1</td>
<td></td>
</tr>
<tr>
<td>Brittleness</td>
<td></td>
<td>7 3</td>
<td></td>
</tr>
</tbody>
</table>

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defence activity or other organization (corporate author) issuing the report.

2a. SECURITY CLASSIFICATION: Enter the overall classification of the report. Indicate whether "Restricted Data" is included. Please use in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that only portions have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parentheses immediately following the title.

4. DESCRIPTIVE NOTE: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHORMC: Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initials. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

5b. REPORT DATE: Enter the date of the report as day, month, year or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

9a. A. & B. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, selected numbers, system numbers, task number, etc.

9b. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified in the originating activity. This number must be unique to this report.

9c. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (either by the originator or by the sponsor), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

1. "Qualified requesters may obtain copies of this report from DDC."
2. "Foreign announcement and dissemination of this report by DDC is not authorized."
3. "U.S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through the..."
4. "U.S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through..."
5. "All distribution of this report is controlled. Qualified DDC users shall request through...

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (spearheading) the research and development, include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (U); (D); (C); or (U). There are no limitations on the length of the abstract. However, the suggested length is from 150 to 250 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designations, trade name, military project codes, national, geographic location, may be used as key words but will be followed by an indication of technical content. The assignment of links, rules, and weight is optional.

Unclassified

Securities Classification

12 Securing Input Information
A survey made a few years ago showed that 66 out of 100 recipients of a given report cataloged that report for their own purposes. Twelve recipients abstracted it. Thus, standardized cataloging of reports at the source should develop benefits for all recipients of such reports. At the same time the freedom to use terms of their own choice will provide input for DDC's lexicographers so they can do a better job in keeping the Thesaurus of Descriptors current.

A machine printout of the information shown in Figure 11 is portrayed in Figure 13 with the identifying number of each element shown. In devising the machine applications, one of our objectives was to achieve improved quality of product. This machine printout

Figure 13
Machine Printout of Information on Worksheet
will be used for verifying the information that is actually recorded in the machine system. Figure 14 shows the entry exactly as it appears in TAB. The item numbers are suppressed, yet they are in the machine record for the purposes already indicated. The long-range objective of this application is to use the computer for compiling TAB itself.

At present TAB is issued semimonthly, listing reports in thirty-three major subject categories. DDC is currently receiving about 50 per cent of the reports which are generated under the DoD RDT&E programs. Within a year or two, document input may almost double since a program is underway to assure that DDC will receive at least 90 per cent of the reports prepared. By that time the use of a single bulletin covering all subject categories would be too cumbersome. In addition, most users are interested in only a portion of the subject matter. Hence, it is probable that the TAB will be produced in a number of selective arrangements. The various issues of TAB will be compiled by means of the computer, and indexes will be compiled which cite the appropriate issue of TAB for each entry in each index. Cumulated indexes will be issued quarterly and annually as already described.

There are several important considerations in the Request Processing portion of the program. At present, document requests average about 5,000 per day. This represents a major work load in key-punching. Consequently, we are exploring the use of mark sense or "Port-a-punch" cards. Experiences of others who have tried these methods have been discouraging. However, we are highly motivated in this area. In the interest of providing requested documents on a timely basis, it is essential that manual effort in processing the requests be held to an absolute minimum. If neither of the alternatives is successful, we shall try optical scanning of requests making use of equipment such as is being developed for the Post Office Department. Speaking of the Post Office, the mailing of 5,000 reports a day represents a rather respectable work load in terms of mailing labels. In the past we have used Addressograph-prepared mailing labels prefilled by user code number. The person who wraps the requested document had to withdraw a label matching the requester's code and apply it to the package. Maintaining the stock of preaddressed labels, and finding and applying the right label are additional work loads which consume valuable man-hours. For these reasons we have designed the request processing run to produce the mailing labels.

The Field-of-Interest Registers (FOIR) are not established for an indefinite period. Normally, an FOIR expires when the cited contract or grant terminates. The planned termination date is shown on each FOIR and is incorporated into each user record. A standard part of the request processing run is to provide expiration notices


Descriptors: (*Creep, Metals), Failure (Mechanics), Deformation, Grain boundaries, Strain (Mechanics), Stresses, Time, High temperature research, Temperature, Copper, Theory, Crystal lattice defects, Diffusion, Crystal substructure, Grain structures (Metallurgy), Anisotropy, Iron alloys, Silicon alloys, Test methods, Recrystallization, Controlled atmospheres, Test equipment, Internal friction, Low pressure research, Tensile properties, Tables, Data, Ferromagnetism, Shear stresses, Hydrogen.

Identifiers: 1963, Activation energies, Curie temperature.

Constant tensile stress creep tests in dry, deoxidized hydrogen and measurements of dynamic Young's modulus in vacuum were carried out at elevated temperatures on polycrystalline sheet specimens of 001 (110) - oriented Fe-3.1%Si, 001(100) - oriented OFHC copper. The dynamic Young's moduli of Fe-3.1% Si decreased strongly with increasing temperature between 500 degrees C and 750 degrees C and thereafter assumed a lesser temperature dependence. This was attributed to the loss of ferromagnetism as the Curie temperature was approached. Creep tests on Fe-3.1% Si showed that a Curie point effect existed such that the ferromagnetic state has a higher creep strength than the paramagnetic. This Curie point effect was shown to be equivalent to that for self-diffusion in iron. Etching of dislocation sites after creep showed that (1) edge dislocations pile up at grain boundaries and polygonize perpendicular to their glide planes, (2) the dislocation density developed during steady-state creep increases markedly with increasing creep stress, and (3) grain boundary serrations are developed at grain boundaries. (Author)
thirty days in advance of the established expiration date so the user can take action to extend the date if extension of the contract is under consideration.

The request processing package also incorporates the program to accomplish the monthly review for automatic time-phased down-grading. In addition, the daily request processing run will provide document accountability records for classified documents plus the necessary shipping receipts. Further, this application provides demand analysis on documents in the system in order to provide the basis for optimum destruction policy and prestocking policy. Finally, it provides inventory control—indicating what reports are available on the shelf, what reports are to be prestocked, and what reports are to be processed for single copy reproduction.

An illustration of inverted file contents is provided in Figure 15. The box at the left represents graphically the structure of the file. In the upper left portion is a synonym or hierarchy code. In the upper right portion is the type code, e.g., descriptor, identifier, personal author, etc. The next entry in the record is the word statement of the term. This will be followed by a numerical code corresponding to the word statement. The balance of the record provides for weight, link, or role as appropriate for the term as it pertains to each AD number, RDT&E Project or Task number, and so on. The inverted file itself is arranged in straight dictionary order with personal author, corporate author, and subject terms intermingled on an alphabetic basis.

The major categories represented in the inverted file are listed as to type, i.e., descriptor (Synonym or Hierarchy), Identifier (Synonym or Hierarchy), *Source, *Contract Number, Personal Author, *Source Acronym, Project Number, Task Number, and *Monitor Acronym. Items designated by an asterisk are used in duplicate checking and document identification when processing unidentified requests.

In the columns at the right, in Figure 15, are indicated the elements of information that will be picked up for various collections under each type of term. Since AD and other collections will be searched for bibliographies, each item is checked. For the Specialized Technical Information Centers (STIC's) or Technical Evaluation Centers, only subject type terms are employed. These could be the basis for specifying the interest profile of appropriate activities. These profiles are intended for use in referral actions and will be used in automatic announcement of documents to major information activities on a selective basis.

The inverted file on FASTRAND and the Reference Address (Index) Table on FH 880 are shown in Figure 16. Searches involving infrared pulses would enter the FASTRAND storage via the index on the FH 880. Similarly, a search involving infrared pulses or infrared
<table>
<thead>
<tr>
<th>SYNONYM OR HIERARCHY</th>
<th>TYPE CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MISSILES</td>
<td></td>
</tr>
<tr>
<td>WEIGHT LINK ROLE AD NO.</td>
<td></td>
</tr>
<tr>
<td>AD 001</td>
<td></td>
</tr>
<tr>
<td>AD 033</td>
<td></td>
</tr>
<tr>
<td>AD 400</td>
<td></td>
</tr>
<tr>
<td>RDT&amp;E 007</td>
<td></td>
</tr>
<tr>
<td>RDT&amp;E 015</td>
<td></td>
</tr>
<tr>
<td>RDT&amp;E 028</td>
<td></td>
</tr>
<tr>
<td>STIC 002</td>
<td></td>
</tr>
<tr>
<td>STIC 015</td>
<td></td>
</tr>
<tr>
<td>STIC 073</td>
<td></td>
</tr>
<tr>
<td>Other 003</td>
<td></td>
</tr>
<tr>
<td>Other 028</td>
<td></td>
</tr>
<tr>
<td>Other 075</td>
<td></td>
</tr>
<tr>
<td>Other 110</td>
<td></td>
</tr>
</tbody>
</table>

**TYPE**

<table>
<thead>
<tr>
<th>DESCRIPTOR</th>
<th>SYNONYM HIERARCHY</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDENTIFIER</td>
<td>SYNONYM HIERARCHY</td>
</tr>
</tbody>
</table>

**SOURCE**

- **SOURCE SERIES**
- **CONTRACT NO**
  - **SERIES NUMBER**
  - **SERIAL NO. & DATE**
- **PERSONAL AUTHOR**
- **SOURCE ACRONYM**
  - **SERIES NUMBER**
- **PROJECT NUMBER**
- **TASK NUMBER**
- **MONITOR ACRONYM**
  - **MONITOR SERIES NUMBER**

**AD RDT&E STIC OTHER**

- X X X X X
- X X X X
- X X X
- X X X
- X X X
- X X
- X X X
- X X
- X X
- X

**Figure 15**
Illustration of Inverted File Contents

Radiation would enter FASTRAND via the FH 880 Index. In order to illustrate the hierarchy feature, I have selected infrared radiation which is shown as having two records on the FASTRAND. If a request specifies only reports of sufficient scope to warrant assignment of the term by the analyst, the top record would be searched. If a request specifies infrared radiation and all terms specific to it, the next record would be searched. Thus, in the DDC system the advantages of the machine applications are realized without the risk of getting reports dealing with terms that are more specific than desired.

Figure 17 shows a rather complex example of the search capability designed into the retrieval system. In effect it says that the requester wanted documents characterized by term A or B or C and D_B and E_B and F_B and G_B (the sub B indicates these terms are specifically linked as applied to the documents) and H or I and not J. In each case the requester can specify collection (i.e., AD or RDT&E), weight, and role. I deliberately said “designed into the system.” The retrieval programs provide for this kind of complex search specification. However, the Directorate of Document Analysis and Processing must provide for this through a more sophisticated analysis of the reports to include appropriate links and roles before this type of search can be put into practice.
Conclusion

In designing this system, we undertook a massive job. The hardware provides a massive capability. As we gain operational experience on a system of this scope, we undoubtedly will encounter problems in software which the manufacturer has not encountered or foreseen. The DDC configuration is about as complex as they come because of the scope and the variety of applications which are needed for the service to be provided. In fact, the DDC system has just about every type of peripheral gear that is available with the 1107. We are working closely with the manufacturer to identify and to correct any shortcomings in both hardware and software in the shortest possible time, in order to serve our users more efficiently.

Our objective is to provide the best service at the lowest cost. Further, we plan to provide copies of our retrieval (inverted) files to established users who have across-the-board approved FOIR, ADP capabilities and will do their own searching. Major features of the new system with a greater ADP capability were adopted in order
to expedite service, and provide reference tools for manual use or machinable files for the use of others at a net savings to all concerned.

REFERENCES

APPLICATIONS OF DATA PROCESSING AT THE CANADIAN NATIONAL RESEARCH COUNCIL LIBRARY

Jack E. Brown

In the world of libraries, the Canadian National Research Council Library is an unusual and perhaps unique institution, for it performs two closely related but often-times conflicting and incompatible roles. One of these roles is that of a science and technology library and documentation center serving a large group of scientists and engineers engaged in pure and applied research in many areas of science and technology. The other role is that of a National Science Library serving the entire scientific community of Canada.

The NRC Library, as it exists today, consists of a main library which houses the bulk of the Library's half-million volumes, and six smaller and more specialized collections serving several divisions of the National Research Council (NRC) which are located four miles from the main building. The main library acts as the nerve center of the Library system with administrative services, acquisitions, cataloging, classifying, and binding centralized at this point. The branch libraries operate primarily as working collections which, in most instances, duplicate parts of the main library's holdings. By means of close cooperation between the various library units, the maintenance of a union catalog at the main library, good telephone communication, and the use of a station wagon which shuttles back and forth several times a day between the main and branch libraries, the entire resources of the system are coordinated for ready access. The Library has a staff of seventy-eight, twenty-seven of whom are professional, and an acquisitions budget of $200,000.

For purposes of this meeting, it is unnecessary to describe in detail the resources and services provided by the Library—suffice to say the NRC Library is much more than simply a repository for the world's output of scientific and technical literature. It is a dynamic organization which utilizes every means at its disposal to provide the Council's scientific and engineering staff with the publications and

Jack E. Brown is Librarian of the Canadian National Research Council Library, National Science Library of Canada, Ottawa, Canada.
information required in their day-to-day work. These same resources and services are extended to scientists and engineers anywhere in Canada, by means of interlibrary loans, through the provision of photocopies, and by means of a Science Information Service geared to compile bibliographies, carry out literature searches, and answer requests for scientific and technical information.

The NRC Library, as with most other scientific and technical libraries worthy of the name, is endeavoring to keep abreast of the latest developments in the field of mechanized storage and retrieval of information. Key members of the staff are encouraged to attend pertinent training courses and seminars, and one member of the staff whose formal training embraces chemical engineering and mechanized systems of documentation, has been designated Library Systems Analyst. His specific assignment is to determine, in collaboration with the librarians, those operations which can be made to function more effectively through the use of automatic data processing equipment.

During the past four years, the NRC Library has been experimenting with the use of electronic equipment to solve specific problems. The scale of experimentation is indeed modest as compared with similar activities being conducted in many United States libraries. However, we must learn to walk before we can run, and attention has been concentrated on the improvement of those essential operations which, because of sheer volume of work involved, were failing to achieve their objectives.

At present, automatic data processing equipment is being used successfully in the following operations:

1. Preparation of a list of subject headings for use in one of the branch libraries, with revised editions to be issued at regular intervals.
2. Preparation of complete lists of serials held by the NRC Library, and issued annually.
3. Preparation of periodic and cumulated lists of NRC publications, together with author and Keyword-In-Context (KWIC) subject indexes.

As I discuss these three operations, I trust you will keep in mind that the work was carried out within the limits set by existing staff and budget. No additional allotment was provided or extra staff hired. A key punch machine (IBM-26) was acquired by the Library, but all other machines required in the operation were available either at the National Research Council or at other government departments in Ottawa.

Our first attempt to employ electronic equipment in a Library operation was in the preparation of a printed list of subject headings covering the fields of aeronautics and mechanical engineering.
Lacking experienced guidance, we made many mistakes before the system was operating to our satisfaction.

Three years ago the Library started work on a revised list of subject headings used in indexing technical reports received by the Aeronautical and Mechanical Engineering Branch Library. The headings and sub-headings were typed on 3"x 5" cards, but the ultimate aim was to prepare a list of headings which could be revised and reissued as often as was necessary and with a minimum of repetitive work. Various techniques for obtaining lists from the cards were evaluated, and it was decided that punched paper tape would best meet our needs. This decision was in some measure influenced by the accessibility of three Flexo-writer machines.

In theory the method should have worked; in actual practice it created more problems than it solved, due largely to our inexperience and inadequate guidance. The tapes, when consolidated, proved to be incompatible, and single letters, parts of words and whole words failed to appear in the printed list. These errors were, of course, quickly discovered and the printing operation halted. Furthermore, we found the updating of the tape, to incorporate new headings, to be a cumbersome and frustrating task.

At this point the Ottawa office of IBM became interested in our problem and offered their assistance. They suggested converting the paper tape directly to magnetic tape from which a printout could be obtained through the use of the IBM 1401 computer. Since it was the first time the IBM office had tackled such a project, they offered to do the work for a nominal sum and we agreed. Many programming difficulties were encountered during the various steps in the conversion from punched tape to computer printout, but eventually they were overcome and the final results justified the adoption of IBM equipment.

New subject headings and corrections are prepared on punched cards and the magnetic tape updated at regular intervals. The preparation of new editions of the list of headings is now a relatively simple and inexpensive operation. The chief cost lies in the multilithing of sufficient copies for distribution to other interested libraries.

The second project, the preparation of a complete list of serials held by the NRC Library, has been described in detail in an article in Canadian Library.\textsuperscript{1} For this reason, and because similar procedures are used in several U.S. libraries, I shall limit my discussion to the main features of the project.

Because of the nature of scientific publishing, periodicals and other serials constitute the major portion of the NRC Library's collection and account for approximately 80 per cent of its total holdings. At the present time, the Library receives more than 10,000 different serial titles. The preparation and publication of up-to-date lists of these serials is a formidable task, but one which must be continued if Canadian scientists are to be made fully aware of the material available to them.
Until 1958 a complete record of periodical titles and holdings was published in book form at three-year intervals. The lists were placed at strategic points throughout the main library and its branches, and in the offices of the various divisions. Later, as the national responsibilities of the NRC Library expanded, copies of the lists were sent free of charge to university libraries and, for a nominal sum, to other interested organizations.

The size of the periodical collection has now reached the point where, with the staff available and by conventional methods, it is no longer possible to issue up-to-date lists at reasonable intervals. The Library examined various mechanized systems to determine which of these, if any, could be used to solve this dilemma and, in June 1963, embarked on a system using IBM punched cards and related automatic data processing equipment. During the preliminary stages of development, it was found that, with a little more effort, it was possible to assign codes to each title which would facilitate the preparation of lists of selected titles on the basis of subject, country of origin, language, subscription agent, and other categories.

The layout of the IBM card is as follows:

Sort groups.—In order to maintain an alphabetical arrangement of titles, and to permit resorting of the file, each card or set of cards is assigned a number. This number sequence (columns 2-6) allows for a listing of 99,999 titles. As new periodical titles are received, additional numbers are assigned to columns 7-8. This allows for the insertion of 99 titles between any two existing titles and in alphabetical order. Thus, up to ten million titles may be listed in the alphabetical-numerical sequence. Since more than one IBM card is required to describe the title, a numerical code in columns 9-10 ensures the proper sequence of cards within a set.

Text.—Columns 11-18 contain the LC classification and Cutter number. Columns 27-80 record the title of the periodical, the holdings of the main library and its six branches, and any necessary notes or "see" references. The information punched in these columns determines the number of cards required for a given title—usually four to five cards per title.

Subscription agent.—The majority of the NRC Library's periodical subscriptions are handled by eleven agents located in various parts of the world. Here, in column 19, an alphabetical-numerical code has been assigned to each agent, leaving fifteen additional letters available for use at a later date.

Subscriptions.—A numerical code, in column 20, is used to indicate whether a journal is received as a paid standing order, as a paid subscription renewed each year, received at irregular intervals, or received on exchange, as a gift, or through membership in a society. A numerical code, in column 24, records the expiration date of each paid subscription by month.
Language.—The twenty-one major languages in which the periodicals received by the Library are printed are indicated by an alpha-numerical code in column 22. International journals which contain papers published in a wide variety of languages are not coded.

Translations, abstracting and indexing services, bound volumes—a yes/no number code, in column 23, records those journals which are English or French translations of journals published originally in some other language, abstracting or indexing services, and journals which are bound on a current basis.

Holdings.—A numerical code, in column 24, records titles held by the main library and/or any of the Library's six branches. The code also indicates duplicate sets held in the reserve collection at the main library.

Country.—An alpha-numerical code, in column 25, indicates the country in which a journal is published. A combination of the information recorded here and in column 22 permits the preparation of lists of journals published, for example, in Russia but printed in another language.

Ideally, all information to be coded should be keypunched in one operation. Because of staff shortages this was not possible, and columns 11-18 and 27-80 were punched first to record the LC classification and Cutter number, the title of the serial, the holdings, and so on.

At this stage, the resulting cards were run through an IBM 407 to obtain a complete printout of all titles and holdings. The first printout was done on 11" x 12-1/2" sheets, with space left at the lefthand column for the insertion of coding symbols. Sets of sheets from this printout were distributed to selected members of the staff for proofreading and the assigning of appropriate codes. Errors or omissions were noted on the work sheets and the sheets forwarded to the keypunchers for the preparation of corrected cards.

Upon completion of proofreading and keypunching of codes and corrections, the complete set of cards was ready for preparing the final and master list of serials, again by the use of the IBM 407. The printout was done on 15" x 18" sheets which were then reduced by Xerox camera to 8-1/2" x 11" duplimat plates. Duplication was carried out by means of multilith machines. If a reduction is not required, the master copy, of course, can be printed directly on to duplimat paper.

New titles, changes of title, or changes in holdings, together with the pertinent coding, are recorded on specially designed 3" x 5" cards. The layout of these cards enables the key punch operator to prepare the IBM cards without further instruction from the librarian.

Supplementary lists of new serial holdings, for internal use, are run off at four-month intervals. The new cards are then incorporated into the master file preparatory to the printing of a revised and complete list of serials.
I have indicated earlier that this mechanized system enables the Library to meet requests for lists of selected journals which hitherto could not be satisfied at reasonable costs by conventional methods. For example, the Library has prepared such lists as journals held by each branch library, mathematical journals (or any other subject), Russian language journals, journals published in China, abstracting and indexing services covering all subjects or selected subjects, journals received on exchange or as gifts, and journals whose subscriptions are handled by a specific agent.

The third and most recent application of data processing equipment in the NRC Library has been to prepare a list of all papers published by NRC personnel, and other publications issued by the National Research Council. The procedures used are similar to those described above and can be dealt with in less detail.

The NRC, as a publisher of scientific and technical information, is best known for its seven Canadian journals of research: Canadian Journal of Physics, Canadian Journal of Chemistry, Canadian Journal of Biochemistry, Canadian Journal of Botany, Canadian Journal of Physiology and Pharmacology, Canadian Journal of Zoology, and Canadian Journal of Microbiology. It is also the publisher of many separate monographs and scientific series, but the majority of the reports written by the Council's scientific staff are published as papers in international scientific and technical journals.

It has been a relatively easy matter for the Library to issue periodic lists of NRC publications. On the other hand, the preparation of cumulated lists of more than 8,000 publications has become a task with which the Library could not cope by conventional methods. The success of the serials operation prompted the adoption of these same techniques to solve this new problem. Because of the descriptive nature of the titles of the papers, it was further decided to compile a KWIC index.

The preparation of IBM cards recording all bibliographical information and NRC numbers for some 3,000 papers published between 1958 (the date of the last cumulation) and December 1963, was completed in three weeks. A preliminary bibliography for proof-reading purposes was run off on an IBM tabulating machine, and a list of 250 non-significant words keypunched. The latter cards and the title cards were turned over to the IBM Data Processing Center and the KWIC index prepared. Punched cards are now being prepared for all papers published between 1918 and 1958, and a complete list of NRC publications will be issued.

Once again, it is worth noting that the preparation of the master file of IBM cards offered no saving in time or money as compared with the typing of cards or lists. However, with the completion of the master file, it requires only a few hours to prepare supplementary lists of NRC publications or complete cumulations, each with author
indexes and KWIC indexes. The compilation of lists of papers by author or by publishing journal is also a simple matter.

In conclusion I must emphasize that, in each of these projects, our aim was not to reduce the costs of existing operations; rather, we were seeking new procedures which would enable the Library to provide several needed services which could not be performed by conventional methods and without an increase in staff or budget. The fact that, through the use of automatic data processing equipment, we were able to provide these services at no additional costs, was all the encouragement we needed to extend our experiments to such areas as circulation control, acquisitions, and the preparation of printed catalogs and accession lists.

The Library staff has become conversant with the techniques and possibilities of data processing machines, and looks forward to the time when mechanized systems of information storage and retrieval will become a reality. We have learned to walk, and hope that we may soon be able to run.

REFERENCES

POSSIBLE APPLICATIONS OF DATA PROCESSING EQUIPMENT IN LIBRARIES

John A. Wertz

Conferences such as this one have made it increasingly evident that anyone speaking of the computer and the library is no longer dealing with possibilities but with probabilities. Historically, the library profession has adapted, if at times with some misgivings, any technological advance that promised to solve its problems. The computer has been no exception to this rule. That the computer is useful in the library has already been demonstrated. The librarian is now concerned in finding new applications for the computer within the library.

The development of the computer has been extremely rapid. The digital computer and its associated technology has been on the commercial scene a relatively short time. In little more than a decade its uses have outgrown the laboratory and become commonplace in the business and academic world.

In fact, as recently as 1958 General Electric’s Computer Department installed industry’s first solid-state computer and first computer system utilized by a bank for electronic bookkeeping. It was called ERMA, for Electronic Recording Method of Accounting. It represented the largest single order ever placed for computers—some $60 million—and set the stage for a complete new generation of solid-state computers.

Then in 1961, Western Reserve University installed a General Electric GE-225 general-purpose computer for information storage and retrieval. It is used by the Center for Communication and Documentation Research to perform literature searches within various technical and scientific areas.

In the process, a whole set of new technologies and new epistemologies have been engendered. It is these subsidiary effects of the computer that have had the greatest import for the librarian. The relationship of the librarian to the new theories of information

John A. Wertz is a Programmer at the Programming Research & Development Computer Department, General Electric Company, Phoenix, Arizona.
propagation and dissemination is a major challenge to the profession. But this is not the time to investigate the metaphysic of librarianship.

What we are concerned with today is the computer's impact upon library technology, not theory. We may take comfort from the fact that the computer is not the first device to have affected library technology. The steel pen, the typewriter, the camera, the printing press, all of these have been assimilated and have become tools for the librarian. In this respect the computer is no different from the others; it too will eventually be looked upon as casually as the typewriter is today.

Before we can explore the possible utilization of the computer in the library we must define the limitations of the device. The computer is a computational device. It was invented to perform numeric calculations. The computer, in its mathematical function, is also capable of quantitative comparison. It is also capable of being programmed to take action depending upon the result of a comparison. It can be made to rearrange data within itself. This is the sum total of the capabilities of the computer, oversimplified; but we must realize that we cannot ask the computer to think, to be intuitive.

From these simple mathematical functions have evolved the techniques of general data processing. As this concept of generalized data manipulation expanded, the librarian and the computer have been brought inexorably together, the basic function of the librarian being, after all, the rational organization of data.

If the librarian is to use the machine, he must be able to express the library problem at a logical level compatible with the capabilities of the machine. This means that operations must be reduced to the simplest possible logical steps. And it is here, in the definition of the process or the problem, that the real difficulty lies.

Library utilization of the computer is limited only by the ability of the librarian to define his process at a sufficiently simple level. In this the librarian has no greater problem than any businessman who might wish to use a computer. It is immaterial to the computer what it is doing. The computer has no awareness of whether it is doing simple accounting or very sophisticated mathematical analysis.

The librarian's problem is reduced, then, to one of defining library techniques. This is the point at which the profession now finds itself, struggling with the reduction of its technique to the machinable level. In a sense this is not a new struggle, it is as old as the profession. There has been a constant dialogue within the field trying to define its intellectual content. This had usually taken the form of trying to define and separate the "clerical" tasks from the "professional" operations. The computer has forced a new rigor into this dialogue. Now for the first time the librarian is faced with the opportunity of the perfect clerk, a clerk with no initiative or judgment but with an infinite capacity for letter-perfect adherence to instructions.
It is this dialogue which gives the first clue as to the possible uses for the computer. Search the library for operations which consist only of the clerical task of rearranging the format of information, the simple comparison of one datum to another, or the creation of ordered lists of data. These are the things a computer can do.

There is little in the operation of a library that does not fit into one of these three categories. There are very few actions that require actual interpretive decision on the part of the library personnel. In fact I can identify only three: (1) the decision to acquire a certain document, (2) the classification of a document, and (3) the reduction of a reference request to terms suitable for effective searching of the library resources. All other tasks—at the moment excluding from consideration the entire area of administration—are clerical dependencies upon these three decision points. There is no reason, therefore, why the abilities of the computer cannot be utilized in every aspect of library technology.

Given these three decision points, the intervening processes must be reduced to computer terms. It may at the moment seem to the librarian to be a Gargantuan task. In some ways it is. There exists, however, a large corpus of technical and intellectual know-how within the computer industry, developed primarily through the analysis of the operation of business and manufacturing firms. But is there very much difference between the ordering of raw materials and the purchase of books, or between the problem of inventory maintenance and the problem of circulation control? What matters here is not the name of the process but the ability to reduce a type of process to machinable form.

There is another aspect to this analysis of processes, the need for a parallel analysis of the forms of data involved. Thus the librarian who would achieve automation is faced with the simultaneous tasks of systems analysis and source data automation. The two studies complement one another, however, contributing to each other's successful completion.

Needless to say, the same problem has long faced industry, and a method of approach has been developed by the professional systems analyst. I believe that the current state of the art on the part of both the librarian and the computer industry is such that a fully integrated computer system for the library is well within our grasp. The existence of such a system for any particular library is only a matter of time. In fact several libraries have already begun the process.

What will the completely integrated computer system for a library include? Everything. It has to. The process of computerization can provide too many efficiencies to be limited in scope. This is not to say that the changeover might not be gradual and absorb one area of the library operation at a time, but the conversion must be complete to be effective or profitable.
Computerization will have to be a co-operative venture. Both the librarian and the manufacturer must be allowed to contribute to the process. The librarian should insist that no violence be done to his technology under the guise of conforming to the computer's needs, but he should also be ready to look at some of his hallowed traditions with a critical eye. Respectful co-operation should be highly beneficial to both parties.

Perhaps the first question is where should automation begin. There has been much talk about the use of the computer as an information storage and retrieval tool. All too often this is conceptualized as a simple putting of the card catalog into the computer and then asking it questions. If we accept the principle of source data automation as a critical criterion, this approach to automation breaks down. True, the card catalog is the source document for reference service, but in the larger view of the library process it is only one of the many intermediate documents. Or for that matter, it can be viewed as the end product of the cataloging function.

All this discussion re-enforces one point: automation must be planned with the total system in mind. The librarian must prepare to find some way to integrate the data used in every process, from the initial purchase request to the final discard notation. In this goal the librarian has the advantage over many prospective computer users, as he is already well versed in the concept of the unit record, a basic computer technique.

It is easy to forecast the integrated computer library system. It is close enough to reality to need no Jules Verne as its prophet. The impact of automation will be greatest upon the technical service and administrative areas of library technology. This is partly because the computer cannot change the intellectual process of questioning (and decoding the question) and partly because the technical processes are most open for improvement.

In this library of the near future, automation will be invoked from the minute a purchase decision is made. From that moment a unit record will accumulate all the pertinent information about the transaction and the document. As each new fact is developed it will be added to the unit record through the medium of punched cards or paper tape or both. At any moment the main files may be interrogated for the status of single items or for batches of data, such as orders outstanding, encumbered funds by department or by vendor, etc. The process of accretion of information to the unit record will continue throughout the usual process of order, receipt, and cataloging. The unit record, replete with cataloging information, will be used in many machine files comparable to the shelf list, author catalogs, etc.

One element which should not disappear from the library scene is the card catalog. It is still the easiest and cheapest way to interrogate the library collection. The cards will be prepared as a
by-product of the computer processing, however, and will be able to attain a new level of accuracy and completeness. By allowing the normal card catalog to carry the burden of the average, simple request, the computerized unit records can be reserved for the more challenging aspects of information (or reference) retrieval. This is where the complex, multi-faceted reference question will be answered. This file will also furnish the comprehensive demand bibliography. Here also is the source of the recurring bibliography of the library’s specialty. But note the difference between this master record and the ones now used in much information storage and retrieval. Whereas the latter are created especially for the purpose, involving a duplication of effort, the master record in the integrated systems library will have been the result of a gradual and programmed accretion of knowledge. Its creation will not have involved duplication of human effort.

Needless to say, the circulation records will also be automated. Here it is harder to project an image of the possible system as every library will have highly individual needs. Here also it is difficult to predict the effect of possible developments in charging machines.

The most fascinating prospect for the automated library lies in the administrative field, however. It is a rare library that has either a sufficiency of administrative statistics or a means of utilizing them efficiently. The systems automated library would be in an excellent position in terms of statistical records. Statistics would be extracted via computer from the various daily working records. They could then be correlated and printed in usable form by the computer.

The possible types of analysis are manifold. For instance, helpful studies might be made of the rate of document use against subject area as a means of maintaining an up-to-date collection, or perhaps a statistical analysis of borrower patterns in order to determine more intelligently the location of a new branch or bookmobile stop. Studies could be made of the internal operations also. Just a few possibilities might be: (1) analysis of vendor or binder performance on the basis of cost or service, (2) programming of personnel scheduling to even out work loads, and (3) the programming of serial binding to take advantage of both cost and time factors. These are just some of the many analytical possibilities in the systems oriented library.

Now one last note on the advantages of the systems approach to library automation. If the library has been carefully analyzed, its size makes little difference to the computer system. There will be certain differences in technique between libraries of widely divergent sizes, but once a particular library system is developed it will be much more elastic for library growth than would any manual system. Another consideration is the slowly changing emphasis in libraries
from the treatment of books to the treatment of report and serial literature. A system which has been carefully automated would be much more amenable to shifts to deeper levels of subject cataloging or indexing than is a manual system. The computer can become a tireless inter-filer of subject lists.

These are not visionary schemes; such totally integrated library automation systems are just around the corner. There is nothing I have spoken of that is beyond the present state of either the librarian's or the computer manufacturer's art. There is no technical reason why such a system could not be operable next year. It is my firm conviction that the librarian, pressed on one side by the information explosion and enticed upon the other by the increasing availability of computers will soon turn to library systems automation.