OE FORM 6000, 2/69

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

ERIC ACC. NO. ED 030 776				IS DOCUMENT COPYRIGHTED? YES □ NO ☒
CH ACC. NO. AA 000 383	P.A. 52	PUBL. DATE Jul69	ISSUE RIEDEC69	ERIC REPRODUCTION RELEASE? YES □ NO ☒ LEVEL OF AVAILABILITY I ☒ II □ III □

AUTHOR
Maron, M.E.; And Others

TITLE
An Information Processing Laboratory for Education and Research in Library Science: Phase I.

SOURCE CODE CIQ11405	INSTITUTION (SOURCE) Calif. Univ., Berkeley. Inst. of Library Research.
SP. AG. CODE RMQ66004	SPONSORING AGENCY Office of Education (DHEW), Washington, D.C. Bureau of Research.

EDRS PRICE 0.75;7.15	CONTRACT NO.	GRANT NO. OEG-1-7-071085-4286
REPORT NO.		BUREAU NO. BR-7-1085

AVAILABILITY

JOURNAL CITATION

DESCRIPTIVE NOTE
141p.

DESCRIPTORS
*Library Science; *Library Education; *Library Research; Computer Assisted Instruction; *Information Processing; Information Science; *Training Laboratories; Laboratory Training; Computer Based Laboratories; Systems Approach; Library Technical Processes; Automation; Educational Needs; Information Retrieval

IDENTIFIERS

ABSTRACT
Study, research, and development were undertaken in the first 18 months of a program (Phase I) to design and implement an Information Processing Laboratory for teaching and research in the field of librarianship. Work during this period was concerned with the planning of the Laboratory and its development according to plan. The planning resulted in definition of initial topics within librarianship to be supported by the Laboratory in relation to the educational needs of the field. This, in turn, led to the development of computer programs for on-line interrogation and search and data files upon which to "exercise" these techniques, as well as other Laboratory elements. The Phase I work included assembling and checking out these initial pieces of the Laboratory; however, no true operational activities were undertaken in the sense of students using the Laboratory on a regular basis. The Laboratory was designed to include capabilities relating both to intellectual access (e.g., associate searching, automatic indexing, automatic abstracting) and to more traditional course content (e.g., subject cataloging). Future directions and plans for the Laboratory were included in the Phase I report. (JH)

INTERIM REPORT
Project No. 7-1085
Grant No. OEG-1-7-071085-4286

AN INFORMATION PROCESSING LABORATORY FOR EDUCATION AND RESEARCH IN LIBRARY SCIENCE: PHASE I

by

M. E. Maron A. J. Humphrey J. C. Meredith

Institute of Library Research
University of California
Berkeley, California 94720

July 1969

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

INTERIM REPORT
Project No. 7-1085
Grant No. OEG-1-7-071085-4286

AN INFORMATION PROCESSING LABORATORY FOR EDUCATION AND RESEARCH IN LIBRARY SCIENCE: PHASE I

By

M.E. Maron
A.J. Humphrey
J.C. Meredith

July 1969

The research reported herein was performed pursuant to a grant with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education pos:tion or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

Page

TABLE OF CONTENTS (Cont.)

APPENDICES

LIST OF FIGURES

FOREWORD

This report contains the results of the first 18 months (June 15, 1967 - December 14, 1968) of effort under a grant (OEG-1-7-071085-4286) from the Bureau of Research of the Office of Education, U.S. Department of Health, Education and Welfare. In addition, the University of California provided contributory support to the project. The principal investigator was M.E. Maron, associate director of the Institute of Library Research, and the project directors were Allan Humphrey and Joseph Meredith.

The three co-authors are jointly responsible for the contents of this document; each of us was involved to some extent in all parts. However, section 6 on MTM is almost exclusively the work of Joseph Meredith, section 5 is mainly the work of Allan Humphrey and sections 1 through 4 were authored by M.E. Maron.

ACKNOWLEDGMENTS

This project began with only a few people but grew rapidly during the next 18 months so that eventually several dozen students, faculty, and Institute staff members were drawn in and actively participating in varying ways. Their contributions ranged from long range planning, criticisms and questions, and active encouragement, to the writing of detailed computer programs and indexing of experimental documents. Thus, this report which summarizes those 18 months of work represents the contributions of many hands and minds, and it is a genuine pleasure for the authors to pause here and thank them. We cannot mention all but we must give special thanks to the following individuals:

Dean R. C. Swank of the School of Librarianship, University of California, Berkeley, has been a constant source of encouragement and help. We are grateful to Grete Frugé, Lecturer, School of Librarianship, for the generous help that she gave to our work on course 201x (subject cataloging). And also we owe thanks to Mary Whouley, Lecturer, School of Librarianship, for design work on a computer-assisted reference course. A number of our doctoral students contributed suggestions, ideas, advice and useful criticisms. Specifically we thank Marcia Bates, Kelley Cartwright, Michael Cooper, Ed Mignon and Irene Travis.

Much of our work has involved considerable computer programming at all levels. William Schieber and Stephen Smith of the Institute of Library Research designed and implemented the Institute's monitor program upon which the operation of our network of remote terminals relies. LABSRC 3 was designed and written by C.V. Ravi, and he also wrote the description of LABSRC 3 which appears as Appendix 5 of this report. Lauro Lazo wrote LABSRC 2.

Rodney Randall and Vivian Frojmovic developed and maintained many of the utility programs.

At the Computer Center of the University of California Medical Center in San Francisco, Sean McGrath deserves special thanks for his extensive help in coder-training and execution of our MTM courses in the PILOT language. And to David Rose, Operations Manager, Computer Center, U.C. Medical Center, San Francisco, we acknowledge the fine help in scheduling and implementing MTM in the UCMCCC system.

Among others on the ILR staff who helped with the MTM work, we thank Helen Yen, Jean Circiello, Katsuhiko Kurano, and Paul Green.

And among those who assisted with the preparation of the experimental corpus, we especially thank Ruth Patrick, Marco Fiorello, Tom Hargrove, Marianne Broz, Harriet Zais and Betty Robinson. We especially thank Connie Farrow for substantial contributions to the technical content of section 5.

J.L. Kuhns of The RAND Corporation served as a consultant to the project and made a number of important suggestions concerning our use of measures of statistical correlation for associative retrieval and we thank him.

Our thanks go to two people who labored with special zest and vitality. Patricia Barkley edited this report and coordinated its preparation. Ralph Shoffner, who is coordinator of projects at the Berkeley branch of the Institute of Library Research, was involved in the activities of this project from the very beginning and without his advice and assistance this work would not have progressed and prospered as well as it did.

Finally, we wish to thank and to commend the work of the Institute personnel who prepared these pages for publication: Joan Chan, Linda Child, Kitty Colburn, Bettye Geer, Linda Horton and Judi Sutliff.

M.E.M.
A.J.H.
J.C.M.

1. INTRODUCTION AND SUMMARY

1.1 The Motivation

This Final Report (Phase I) presents the results of the first 18 months of study, research, and development toward the design and implementation of an Information Processing Laboratory for education and research in librarianship. The purpose of this Laboratory, which we have conceived of primarily as an on-line information processing facility, is to provide a new, active, and powerful vehicle for teaching and for research in the field of librarianship. The impetus behind this work is a realization of the enormous impact that information processing technology (e.g., the digital computer, digital communication systems, video display terminals) will have on future library systems and on the profession of librarianship.

The technology of the digital computer (and the associated conceptual techniques that are presupposed by its use) has a dual significance in this field. On the one hand, it provides the means to store, interrogate, analyze, and retrieve library data, and hence it will be used for automating library services. On the other hand, it provides an ideal vehicle to teach advanced library students about new principles of library science, and hence it can be used for education. The technology and the associated conceptual principles needed in order to use this technology are changing this field. And one of the consequences of this is the need for a restructuring of the education for future library scientists. This, then, is the forcing motivation behind this research.

1.2 The Setting for This Study

This Laboratory project has been conducted in an intellectual setting that we believe provides a number of factors essential for success.

The work has been done within the Institute of Library Research of the University of California at Berkeley and in intimate contact with the School of Librarianship - a professional school that offers the M.L.S. and the Ph.D. degrees in librarianship. The contact with the School of Librarianship has provided the realism, the awareness of current educational problems and policies, the stimulation and critical advice on the part of faculty and students, that have helped to guide this project.

The Laboratory is not an abstract idea and it is not intended as a facility to teach some ideal student. The Laboratory is intended to be part of an on-going graduate school of librarianship. The Laboratory has to "fit" and eventually its effectiveness must be tested in a real operational library school setting.

Diversity of talent is another important benefit that is obtained by developing this Laboratory in a library school environment within a great university. We have been able to engage the talent of advanced students in disciplines ranging from electrical engineering and computer science, to philosophy, statistics, business administration and, of course, librarianship. These bright, active students have been contributing to the development of this Laboratory and, at the same time they have been learning about problems and new methods in the field of librarianship.

1.3 Summary of Laboratory Status

The work during this phase has been concerned with the planning of the Laboratory and its development in accordance with this plan. The major planning task has been that of evolving the concept of the Laboratory in relation to the education needs of the field, while the development tasks have been to provide the computer programs, data files, equipment, space, etc., which together constitute the Laboratory.

To determine the educational needs of the field, we had to assess the state of the field, its past, its present, and its future. Librarianship has been primarily an applied science concerned with immediate operational problems, and its practices have too often been based on precedent without any explicit logical rationale. Today librarianship is in a state of transition, transition being caused by a number of factors, but mainly due to the increasing demand for information and to the enormous information processing capability of digital computer technology. These factors are causing those concerned with librarianship to look at the field more closely in order to identify key conceptual problems, to discover the rationale for its present practices and to start to develop, if possible, a coherent theory of librarianship.

In short, the field is moving from a pre-scientific to more of a scientific discipline. At the same time, the practice of librarianship is changing rapidly as digital computers are incorporated in library operations. This is creating a demand for librarians who grasp the nature and techniques of systems analysis and automatic information processing. Education in librarianship, then, must provide for the dual nature of the field - the applied and the theoretical - and we plan for the Information Processing Laboratory to fill both needs with emphasis on the theoretical.

The planning has resulted thus far in definition of initial topics within librarianship to be supported by the Laboratory and this, in turn, has led to the development of computer programs for on-line interrogation and search, and to data files upon which to "exercise" these techniques. Thus far, these initial "pieces" of the Laboratory have been assembled and have been checked out.

However, no aspect of the Laboratory is as yet operational in the sense that students are using it on a regular basis.

The Laboratory capabilities relate both to intellectual access (e.g., associative searching, automatic indexing, automatic abstracting) and to more traditional course content (e.g., subject cataloging.

In both teaching and research, there will be multiple modes of providing these laboratory capabilities (e.g., on-and-off-line access through mechanical and CRT terminals, printouts, books, micro-production), but the main one will be on-line access over remote CRT (cathode ray tube) terminals to a computer system. We did not have CRT's during this period, however, and the initial on-line mode we have established uses remote mechanical terminals.

We have developed an initial prototype capability in the two topics (associative search and subject cataloging). To support the teaching of associative search, we have created and programmed three retrieval routines, any one of which can be used with any one of three known statistical measures of term association. These routines are run from on-line terminals on an experimental corpus of approximately 300 deeply indexed documents in the field of library and information sciences. The major contribution of the work in associative search will be in the ability to make comparative studies of query and retrieval effectiveness. Further, we believe that the most complex of these retrieval routines will itself be a contribution to retrieval technique once it is fully debugged and documented.

To support the teaching of subject cataloging, we have written the prototype version of a subject cataloging course to be taught through the CRT terminals. This course is programmed in PILOT, a high-level terminal interaction language written in PL-1, and it operates from the Institute terminals on the computer facility of the University of California Medical Center, San Francisco. (Acoustic couplers are used to connect two of the Institute's mechanical terminals via phone lines with the San Francisco computer, an IBM 360/50, 256 K memory.) The subject cataloging course will be reviewed and revised before it is used on an experimental basis.

We expect that the most important contribution of the subject cataloging work will be to the methods of using computer systems to support education in librarianship. That is, before we can make the course available to students, we have had to extensively analyze and define the course content, the sequence of presentation and indeed the very nature of teaching and computer presentation of topics. In addition, our work in methodology has led to an extension in the methods of programming such courses. As with the associative search routines mentioned earlier, these methods are not yet documented for general presentation.

To support the teaching and research in both areas, we have two remote mechanical terminals connected via data sets and telephone lines to an IBM 360 Model 40 in the Berkeley campus computer center. This machine has a 128 K core memory and 2314 disc storage, plus the normal peripheral equipment. The terminals interact with the computer under the control of a monitor system developed by the Institute.* See Appendix 3 for a more complete discussion.

1.4 Future Directions and Plans

As we move into Phase II on this project, we expect to shift the emphasis from planning and preparation to student involvement in the Laboratory. Of course overall systems planning will continue to some extent and will be directed more toward the student work within the Laboratory and toward determining the educational and research effectiveness of the Laboratory.

We will continue to widen the scope and range of the Laboratory facilities. Specifically, we will be extending and expanding the assortment of programs for the on-line study of methods of intellectual access. First of all, our current routines for handling associative retrieval will be extended and refined. Actual Laboratory exercises designed to illuminate certain features of associative retrieval will be worked out. These exercises and guidelines for the on-line learning of associative retrieval will be coupled with an advanced course in the School of Librarianship thus making the Laboratory a full and integral part of the curriculum.

Programs will be developed that will enable the computer to assist in making comparative evaluations of alternative searches. Thus, not only will the student be presented with the different outputs ("relevant" documents) in response to different search specifications, but the system will assist in providing certain quantitative measures of the differences (e.g., in terms of size and ranking of retrieval items). Also, we will start to select, evaluate, and develop measures of retrieval effectiveness to assist the student in determining the goodness of competing search modes.

In addition to the current emphasis on associative retrieval we plan to widen the scope of topics to include routines for: 1) teaching formal principles of subject analysis and identification, 2) on-line citation coupling, and 3) context searching. This necessitates parallel work on a number of fronts. Not only must we design special routines for presenting, "exercising" and teaching these methods of intellectual access, but we will need special routines for evaluating their behavior.

*The monitor was jointly developed by the Laboratory project and the Institute's File Organization project (OEG-1-7-071083-5068).

We will expand our current experimental corpus of approximately 300 documents, both in terms of size and also in terms of the amount of data that is related to each corpus document. We will be using, for example, not only the standard bibliographic record for each item, but also the citations, abstract, and other contextual data for the corpus items. We will seek out, obtain, and integrate other and different kinds of experimental corpora for comparative studies by the students.

In the work on computer-assisted instruction, we will continue to refine the current computer course on subject cataloging. We plan to test and evaluate the effectiveness of this course by having a relatively large number of students use it. Also we will start to apply these techniques for an on-line teaching of some aspects of the topic entitled "reference."

1.5 Emphasis of This Report

Our purpose in this project is not merely to design and implement an Information Processing Laboratory, but to uncover and clarify some of the key educational issues in contemporary librarianship so as to be able to guide and justify our particular choices and decisions.

We did not start this project with clear and distinct ideas as to exactly what ought to be done. We started with a strong intuitive sense of the problem and we groped, stumbled, sometimes retreated briefly, but eventually made real forward progress. We now feel that we have a firm grasp of the issues and that we can justify the "design principles". Thus, we hope that the results of our inquiry will be of value not only to those students who will be using the Laboratory here at Berkeley, but also to those educators who may be thinking of implementing a similar kind of educational facility elsewhere.

In addition to the detailed presentation of our current status*, the emphasis in this report is on clarifying some of the central issues that must be faced today concerning the impact of digital technology on the future of librarianship and on the education of future librarians.

1.6 The Organization of This Report

This Final Report is divided into six sections of which this Introduction is the first. A discussion of the field of librarianship and its probable future direction is contained in the section 2, entitled "Current Trends in Librarianship". From this view of "what

*For those who would like to have a brief chronology of our activities, see Appendix I.

is happening" flow certain implications for education, and these are discussed in section 3, "The Implications for Education". This sets the stage for a general presentation of the nature and scope of an Information Processing Laboratory, which is contained in section 4.

The final two sections report in detail on our laboratory development effort. The on-line teaching of associative retrieval is contained in section 5. The on-line teaching of traditional subject cataloging is contained in section 6.

2. CURRENT TRENDS IN LIBRARIANSHIP

2.1 Initial Remarks

What kind of education is needed to prepare the new leaders in the field of librarianship?* The answer to this key question can come only after a closer look at what is happening in librarianship today and what the potential implications are for the future. Thus, in this part of the report we will look at some of the forces behind the current changes in librarianship. Specifically we will consider the digital computer and its great impact, present and potential, on both practical and theoretical librarianship.

Librarianship today is a profession in a state of transition. This changing nature of librarianship has implications not only for the future of libraries and library research, but also for the educational requirements for future librarians.

One way to characterize the change is to describe it in terms of a transition from a pre-scientific to a scientific discipline. Traditionally, librarianship has been a strictly practical profession - one that has been looked at (and that has seen itself) primarily as service oriented. In an exaggerated sense, we could say that the profession never fully looked inward at its own subject content in order to formulate and empirically justify some of its general principles. In the pre-scientific state a profession justifies its procedures and practices primarily in terms of rules of thumb, and in terms of its history, its traditions. And this has been the case with librarianship.

As a profession moves to a more scientific state, it begins to identify and explicate fundamental concepts and it begins to formulate principles that can guide and logically justify its practices. Contemporary librarianship is becoming more analytic, self-critical, more scientific and research oriented. It is beginning to see its own subject content more fully and clearly, and it is beginning to apply keener tools of analysis to its own problems.

Thus the subject content of contemporary librarianship, which can be characterized as the general problem of information identi-

*Here as elsewhere throughout this report our concern in librarianship is on the so-called "information science" aspects and not on such specialties as history of the book, history of printing, history of libraries, international and comparative librarianship.

fication, storage and transfer, is beginning to be unfolded and treated in more rigorous ways using new conceptual tools and techniques of logic, mathematics, statistics and systems analysis. Librarianship is a prcfession with both empirical and theoretical content. In its most fundamental aspects, librarianship involves the following kinds of problems: the nature of knowledge and the notion of an information need; the nature of information and how it can be represented, identified, and communicated; the structure of language and how it might be analyzed formally; the meaning of "content" and "about"; and, of course, the meanings and measures of "relevant".

There are numerous important implications and consequences that flow from our account of the scope of librarianship. These implications relate to the design and organization of libraries of the future, they relate to the kinds of research and development that are most relevant to this changing field, and most important for our purposes, they relate to the education and training that is required to prepare a new generation of library scientists who can participate creatively in this rapidly moving field.

2.2 Increasing Demand for Information

The pressure of an exponentially growing population of books, journals, technical reports, etc., certainly can be cited as one of the forces behind much of the current activity in librarianship. And, as the birth rate of books (and other forms of documentary information) booms upward, the problems of storage and access are multiplied at an even greater rate. But this accelerating "input" is only one dimension of the problem; another is the growing population of people who need access to that literature. This greater and more diversified demand for information results in part from a new awareness of the nature of information. Increasingly we are tending to view information as a commodity to which one can assign a value. People want information for pleasure and understanding, for prediction and control; we use information in purposeful ways - whatever our purposes.

Thus, in principle, the value of information can be related to its effectiveness in helping its recipient to achieve his purposes. Whether it be small organizations, large industrial firms, or agencies of our government, there is a growing inclination to view information as an essential resource that must be protected, enriched, and intelligently exploited. Thus, there is pressure from many sectors of our society to develop ideas, techniques, and systems to properly handle these valuable resources of published information. Increasingly, there is a sense of urgency to get on with the important job of designing improved library systems to acquire, identify, store, retrieve, and disseminate information to a growing, diverse class of users.

2.3 The Computer

The major force that is both accommodating and effecting rapid transition in the field of librarianship is the digital computer and its related technology. There has always been a mutual interaction and influence between science and technology. As science develops new physical principles, the principles become embodied in a variety of new technological devices, some of which, in turn, (such as measuring instruments) contribute to the continual development of science.

Digital computer technology is just over two decades old, but the growth and improvements in this field have been most dramatic. We now have really high speed and reliable central processing units; but even more important for library purposes we now have very high capacity magnetic memory systems which can store hundreds of thousands of bibliographic records in digital form. And, of course, there are high density, noneraseable digital photographic memory units that have a capacity of over a trillion bits.

New developments in on-line, time-sharing systems, high capacity communication channels, and inexpensive and easy-to-use terminals allow one to communicate with a central system from remote locations. Because it is possible to formulate very complex rules as a long sequence of very elementary computer instructions, any kind of a task that can be described in complete and unambiguous detail can be implemented as a program for a computer and automated. Thus the computer becomes a powerful tool for analysis, for simple data handling and processing, for teaching and for research.

2.4 The Impact on Applied Librarianship

Applied librarianship denotes that large classes of problems concerning the practical operation of existing library systems. Its concern is toward the immediate, pressing, operational problems that face today's librarians.

Not unlike other organizations such as banks and insurance companies, there are many strictly clerical functions in libraries that are now being mechanized via the digital computer. It seems quite clear that this trend to automate most of the strictly clerical functions within libraries will continue. For a fuller discussion of these issues the reader should see the SDC Technical Report "Technology and Libraries"*.

*C. Cuadra, et al, "Technology and Libraries," TM-3732. Systems Development Corporation, Santa Monica, California. November 15, 1967.

One step "above" the mechanization of strictly clerical functions (such as circulations control) are the problems and opportunities for putting library data (e.g., bibliographic records) in machine form for on-line interrogation and search. Because the Library of Congress has undertaken a massive program to produce and disseminate its current bibliographic records in machine form on magnetic tapes, one can expect that an increasing number of libraries will automate their catalogs. This will allow for machine sorting, reorganizing and updating of files, remote interrogation of the catalog, and if desired, the automatic production of the catalog in book form. Under the influence of the Library of Congress MARC Program, an increasing number of libraries will move in this direction of library automation provided, of course, that the library profession can train an adequate number of people who have the necessary knowledge of systems analysis, computers, and library data processing to deal successfully with this aspect of library operations.

Another category of library automation clusters around the attempts to tie together, through a communications network, currently separate library files. As the trend toward mechanization of library catalogs grows, there will be a growing trend to make the files of one library system available to others; this trend toward library networks will, hopefully, increase coverage and avoid unnecessary duplication of holdings. There are many problems associated with the development and use of very large library networks and centralized systems, but here again a definite picture is emerging. It suggests that networks will play a very large part in libraries of the future. See the review article on this topic by Becker and Olsen.*

2.5 Impact on Theoretical Librarianship

In addition to these three computer applications in practical librarianship, the computer is having an impact on research in librarianship. The computer enables a researcher to formulate and test a variety of indexing methods and search routines that could not be tested and evaluated without mechanical computing aids.

Theoretical librarianship is concerned with the search for optimal tactics and strategies for searching, identifying, associating, clustering, retrieving and disseminating information. Basically the problems of theoretical librarianship are problems of information science because they concern the fundamental commodity of information, its meaning and measures and because

*"Information Networks," Annual Review of Information Science and Technology, Vol. III (1968), Chap. 10, pp. 289-327.

they concern automated information systems and their evaluation. Theoretical librarianship also encompasses all the related conceptual tools and techniques of computer science, cybernetics, and information theory most broadly conceived.

The emphasis in theoretical librarianship is on finding organizational principles that can be formalized so that most of the operations required for information storage, analysis, identification, and transfer can be automated as an on-line system of information interrogation, search and retrieval. The goal, of course, (possibly never to be fully achieved) is to have a computerized library system with the full text of the documents in machine form and so organized that a patron can express his request for information in ordinary English. The automated system, in turn, would interrogate the user further as needed, and then search, identify, and retrieve all and only the desired information.

The problems involved in the design of library information systems --- both for literature searching and for factual question answering --- are enormously complex intellectual problems. In order to solve these problems, one must come to grip with the following questions which belong within the domain of theoretical librarianship:

--- What is the meaning of "information need", "relevance", "about", "subject content"?

--- What is the meaning of "similar in meaning" and "similar in content", and how can the above relationships be measured?

--- By what set of rules and procedures can documents be related (clustered, grouped) according to measures of similarity?

--- What are best measures for the retrieval effectiveness of a library system?

--- By what set of rules and procedures can documents be identified for retrieval purposes?

--- How can problems of encoding, storage, and access be related so that a library file is optimally organized to provide the best response time in the most economical way?

The problem of how requests and documents can be analyzed in order to retrieve all and only the relevant items (and possibly have them ranked by some measure of their degree of relevance) is called the problem of "intellectual access". During the past decade, a growing number of research projects have directed their efforts toward the study of new and improved computer techniques for obtaining deep intellectual access to stored documents. And, as the so-called "library problem" has intensified and as computer memories have been increased in capacity with lowering costs, experiments on the problem of intellectual access have intensified. (For a review of the scope and variety of current activity in this problem area one should consult chapters 4, 6, and 9 of the Annual Review of Information Science and Technology, 1968, Vol. III.)

For example, one of the features of contemporary work in librarianship is the use of quantitative concepts and not classificatory concepts alone. Thus, instead of either assigning an index tag or not, one can assign a tag with a weight which would represent the degree to which the index tag holds for the document in question. One interpretation for weighted indexing is the following: to say that index term I_j holds for document D_i with the weight W_{ij} (where $0 \leq W_{ij} \leq 1$), is to say that there is a probability (whose value is estimated as W_{ij}) that if a patron were to be satisfied with the information contained in D_i, he would be requesting information using the index tag I_j. Given a request for information with a weighted indexing scheme under the above interpretation, the library system must, of course, perform a fair amount of computing in order to determine which document is most probably relevant, next most probably relevant, etc.. Thus, in general, in order to deal with quantitative concepts and quantitative measures of match, we must use a machine because it would be impossible, for all practical purposes, for a human to do the kind of computing necessary in quantitative searches.

However, the problems of intellectual access cannot be solved by technology alone. No amount of computer memory can solve the problems because they are, fundamentally, intellectual problems. These conceptual problems that block the road to full library automation are very deep and very complex. We are only just beginning to see in a clear way what some of the dimensions of these problems are.

2.6 The Future of Libraries and Librarianship

Libraries of the future will be very different kinds of information systems than those in the past. The computing machine, digital communications systems, video display terminals, offer the technology needed to construct highly automated information systems in which large amounts of bibliographic data and full text are stored in digital memories and where a variety of special data banks are interconnected to form centralized repositories accessible to a large number of remote users. The technology is here, but there are complex conceptual problems that must be solved before the technology can be properly exploited in the service of librarianship. Clearly the rate of progress toward the development of such systems hinges greatly on those young people who are now preparing for a career in theoretical librarianship - those who hopefully will be solving the conceptual problems mentioned above. Also, of course, the very image of on-line fully automated library systems will influence the direction of the education and research of current doctoral students. The profession and our libraries are changing, under the impact of many forces, one of the most important of which is the computer. Clearly, librarianship is in a state of transition.

And, finally, one of the things that emerges from an examination of the impact of the computer (and the related concepts from the information sciences) is a change in the very conception of librarianship. Future librarians will talk less about books, and bookshelves, LC numbers, and shelf lists, call numbers and circulation desks, and instead will cast their analyses in the language of information identification and information processing, control, communication, and evaluation of information systems. But we mean more than this. In the past when one thought of a library, one thought of the "Collection" - i.e., the contents (books, manuscripts) of the library. Clearly the documents (or better yet the information they contain) are necessary ingredients of any and all library systems. But, an equally critical element, above and beyond the content of a library are the processes and procedures for identifying and retrieving; i.e., the rules for bibliographic organization and access. A library, in a very fundamental sense, must be conceived of as the collection *plus the processes needed for proper access*.

And, as librarianship moves toward a more scientific state, library scientists will attempt to formulate optimal rules and procedures for on-line interrogating, identifying, relating, searching, selecting, and disseminating information. Library search tactics and strategies for on-line interrogation and search will have been formulated and refined. This suggests the active view of a library as a *process* - not a static view of a library as a collection (plus the catalog scheme). The inner structure of the subject of librarianship is the study of the *processes* for identification and access.

3. THE IMPLICATIONS FOR EDUCATION

3.1 Initial Remarks

Since the digital computer is here to stay, a central question now is how and where the library profession should move to modify and update the education of its students. Where and how should the curriculum be enriched? What kinds of new courses and educational facilities are required?

3.2 Aims of Education in Librarianship

Broadly speaking the aims of education (excluding job training) are the same regardless of the specific field. Education is preparation for the future: teaching a person how to grow intellectually; and how to analyze and evaluate - regardless of the subject matter. As a student advances within a specific field (e.g., in graduate school) the content of his education becomes more specific and the emphasis becomes more directed, although the basic aims should remain the same.

If education is intelligent preparation for the future, and if the future of librarianship includes, among other things, a larger role for computers and information processing, where and how must education for librarianship be modified to account for this? Meeting the aims of education in librarianship increasingly means teaching students to be able to analyze and evaluate the field in terms of the information sciences and of the application of computer technology. With this kind of education, they will be able to read the growing literature in the field and intelligently apply information science techniques. This applies to doctoral students who will eventually become researchers and educators as well as to master's students who will soon become practitioners in the field.

The M.L.S. program within many library schools is designed to train people with an emphasis in applied librarianship, while at the Ph.D. level the educational emphasis shifts to problems of research and to emphasis on an understanding of the basis underlying logical structure of the discipline. Let us look more closely at the impact of the computer for education in applied librarianship and for education in theoretical librarianship.

3.3 Implications for Applied Librarianship

In applied librarianship the emphasis is on the immediate, pressing, operational problems that currently plague our libraries. What kind of education will be most relevant for the library practitioner in order to prepare him for the extended mechanization of the library in the future? What topics need to be mastered, and at what level of depth and detail? And, very importantly, how can these subjects best be taught in library school? The amount of time that is available for formal education on the part of students is limited. Thus, it is not practical to demand or require all of the course work that it would be "nice" for the student to have. What are the relative priorities?

Needless to say, we do not have any final, uncontested answers. Our purpose here is to indicate what we feel is important in the education for librarianship and why we have moved in the directions that we have in the design of the Information Processing Laboratory.

It seems quite clear that library school students should have some reasonably adequate background in elementary logic, mathematics, and statistics. (And, incidentally what is adequate today will surely need to be strengthened in order to be adequate in the near future.) In addition to a knowledge of these formal tools of analysis (elementary logic, mathematics, and statistics), library students should have training in the techniques and methods of systems analysis and operations research. These new methods for analyzing, synthesizing, and evaluating complex information systems, such as libraries, are becoming increasingly important for a career in applied librarianship.

The next group of topics that these students should know are those that cluster around the computer, the principles of programming, and library applications of information technology. Surely, no college graduate today can claim to have a broad education if he does not understand the elements of computer organization, programming, and some kinds of applications. And, in the case of education for librarianship, it is absolutely essential that students know about the digital computer, principles of its logical organization, and principles of programming. Again, the computer (and its related technology) will be one of the essential ingredients of future library systems and future library practitioners must have a strong grasp of this very relevant technology. Moving, once again, from the general to the more specific, it would seem that some background in library automation should be required. The student could be taken in some detail through the logical steps leading from an initial system analysis of some library subsystem (e.g., serials control), through problems of encoding, file conversion and file organization, to design and coding of the computer routines, to testing, evaluation, etc. Thus, the student would learn about the entire process of admittedly a small piece of library automation.

3.4 Implications for Theoretical Librarianship

In theoretical librarianship the emphasis is on long range research problems - the search for fundamental clarification of key library science concepts and the search for principles of information identification, organization, association, search and retrieval. What does the impact of the computer imply for education for theoretical librarianship?

In the specific case of those Ph.D students who wish to specialize in the information sciences (or theoretical) aspects of librarianship, what are the special needs and educational requirements? In addition to a strong background in logic and mathematics (the formal tools for analysis), they need a strong background in the information sciences, e.g., information theory, computer organization and programming, properties of on-line operating systems, and principles of file organization. And although the primary emphasis of the Ph.D students will be toward theoretical librarianship, they must have a good grasp on the problems and solutions of applied librarianship. Hence, there is the educational requirement for the study of systems analysis and its application to the mechanization of clerical functions in libraries. However, a major part of the problems of theoretical librarianship are the problems of intellectual access, i.e., the study of formal methods for analyzing and retrieving stored information in response to a request for information.

3.4.1 The Meaning of Intellectual Access

The problems, techniques, and methods of intellectual access are crucial in the study of theoretical librarianship. The computer and its potential in libraries of the future lays a demand on the profession that the important subject of intellectual access be taught and taught as effectively as possible.

The problem of intellectual access denotes the problem of how to analyze, identify, search, relate, and retrieve documents that are relevant to a library patron's information need, relative to the request that he submits. Formal techniques for intellectual access are those techniques related to the analysis and processing of linguistic expressions that can be described solely in terms of the location and form of the expressions and that make no reference to their meanings. Thus formal techniques are those that are translatable, in principle, into a computer program; i.e., they can be described completely, unambiguously, and without reference to their semantic content.

Automatic indexing is an example of a technique belonging to the topic of intellectual access. Automatic indexing is the method of deciding in a formal way (i.e., based on the occurrence, frequency and grammatical form of the words [and other clues] in a document) the subject heading(s) under which a document should be indexed.

Another class of techniques belonging to the domain of intellectual access is that concerned with statistical measures of association, closeness and distance as applied to literature searching and retrieval. A wide variety of formal rules for deciding about closeness between index tags and about distance between document representations have been described in the recent literature in the field of librarianship. These techniques are central to the problem of intellectual access.

3.4.2 Research Requirements

A critically important part of the education of the Ph.D student, centers around research; i.e., preparing the student to conduct independent, original research by having him produce a doctoral dissertation. An experimental corpus - at least partially in machine form - will be required for those students who choose to do empirical research involving, perhaps, the design, testing, and evaluation of some new technique for obtaining access to stored literature (as for example, use of a new technique for automatic indexing, weighted indexing, associative searching, etc.). In addition to an experimental corpus in machine form, such a student will need certain computer facilities and appropriate software designed to handle his class of library data processing problem.

One of the major educational needs in research librarianship, then, is a new kind of research facility giving students easy access to the information processing tools they need to conduct empirical research in this field. This facility would in a sense be a counterpart to the empirical research facility that, say, a linear accelerator provides for an advanced physics student.

4. THE NATURE AND SCOPE OF AN INFORMATION PROCESSING LABORATORY

4.1 Initial Remarks

We conceive of an Information Processing Laboratory as a new kind of educational and research facility designed specifically to extend, enhance, and provide an innovative mechanism for education of future library scientists. We interpret this Laboratory as a set of remote terminals connected to a central high speed, general purpose, digital computer. The system is designed so that students may sit individually at a terminal and proceed to call up a variety of different kinds of library related procedures. The student can "exercise" the procedures and evaluate the consequences, introduce modifications, make comparative studies and, thus, gain a new kind of insight into the problems and techniques by actually controlling and observing their behavior. In addition the Laboratory will provide the means to teach some topics utilizing the computer in the instruction. And the Labortory will provide the facilities, programs, experimental data bases, etc., needed by advanced students for empirical research.

4.2 Possibilities and Priorities

In an Information Processing Laboratory, one might want to teach: principles of file organization; on-line use of citation indexing; study of circulation flow via simulation models; computer search techniques; library data processing programming; techniques of query formation; measures of retrieval effectiveness; and so on.

Needless to say, it would be impossible to do everything in such a Laboratory because the resources are fixed. And even if the resources were unlimited, one would have to proceed in a serial fashion - a step at a time. Thus we face the question of what is the best first step. What should we start with and why?

The fulfillment of the rosy predictions about on-line, fully automatic text searching systems hinges on successful research on the problems of intellectual access. And, this is the area that we see needs to be most emphasized for the Ph.D. students because educating the Ph.D. student in librarianship is synonymous with educating the future leaders of the profession and the teachers. It is fcr primarily this reason that we have laid great emphasis on teaching the principles and methods intellectual access in the Laboratory. (See Section 4.4)

4.3 Library Systems and the Instructional Role of the Computer

One of the notions that emerges in thinking about future library systems concerns the problem of teaching a patron how to use such a system. Assume that, in some not too far distant future we have a large scale, on-line interrogation, search, and retrieval system. In order to justify the heavy costs in the development and operation of such a system, its retrieval effectiveness will have to be reasonably good. But we know that no simple information processing of a patron's request and the document file will be adequate because the problems of analyzing requests and documents are exceedingly complex.

In order to get at relevant documents in response to a request, the library system and the patron will have to engage in an iiterative process of communication. Given the request, the system makes a first "move" (i.e., an initial identification and access) and then provides the patron with some feedback (perhaps it displays a sample of the kinds of documents it has selected after this first access). The patron must now decide which of the next possible group of alternative search "moves" would be the best. He must participate in the decision of how the search should proceed. This two-way interaction continues to direct and modify the search until the patron feels satisfied (or else too frustrated to continue).

The point of this is to say that before we ever arrive at the fully automated on-line library search system (if we ever do), we will have developed rather effective semi-automated systems in which the system and the patron interact in order to jointly guide the search. Now in order for the patron to work effectively with such a complex system, he must understand how it operates - what the search tactics and methods of intellectual access are. The ordinary patron cannot be expected to know the full complement of possible search "moves" or what exactly is implied by the use of them. Therefore, periodically he may have to interrupt the search and call upon the system to "explain" the meaning and use of certain alternative search options. Thus, we see in library systems of the future the need of rules, procedures, and methods whereby the system can actively instruct the patron about how to guide a search. It is for this basic reason that we have emphasized our work on computer instruction (See Section 4.6).

4.4 Intellectual Access

4.4.1 Statistical Techniques

Of all the logical tools and techniques for intellectual access that have so far been described in the professional literature of librarianship, which should be selected for study? Again, we cannot implement all of them nor would we want to. Some techniques are more important and potentially more valuable than others. Some techniques represent only minor variations of others. Some may be logically prior to the study and use of others. Thus a first problem that faced us in the design of this Laboratory was the evaluation of the class of tools for intellectual access and the selection of one, or a few, to be developed initially.

One might want to teach:

1) Formal methods for thesaurus and dictionary construction

2) Techniques for automatic indexing, automatic extracting, or automatic classification

3) Techniques of relevance feedback for request modification

4) Techniques for computing degrees of relevance depending on variety of matching procedures

5) Techniques for measuring the retrieval effectiveness of a literature searching system

6) Techniques of clustering, clumping, grouping and their use in library situations

7) Statistical techniques for computing degrees or correlation and closeness between two properties

These measures of statistical closeness have a variety of applications in the general problem area of automatic literature search and retrieval.

As a field of study matures there is a constant attempt to replace classificatory (two-valued) concepts with comparative concepts. In the case of literature searching systems this means, among other things, the attempt to compute degrees of relevance, to compute degrees of closeness between index terms, and to compute degrees of closeness between documents - i.e., statistical measures. In order to implement such systems two requirements must be satisfied: First, of course, comparative concepts must be developed so that one can at least say what it means (and how to compute) degrees of relevance, closeness, etc. And second, there must be some kind of mechanical device to do the computing since it would be unthinkable for a human to do the necessary computing in order to deal with requests for information. Because the computer is exactly the device that will compute any function at high speeds and because of the wide range of applications and the potential power of statistical measures, we decided to start with computer routines for teaching the meaning and use of these techniques.

4.4.2 Associative Retrieval

One of the potentially most valuable uses of a statistical measure of closeness is for the process of *associative retrieval*. Associative retrieval can take a number of different forms, but basically it is a formal method for retrieving documents whose index tags are only *close* to a given specification. We emphasize the notion of *close* to distinguish associative retrieval from a conventional retrieval system that requires a *direct match* between a document index tag and a search specification. Thus, for example, if a user makes a request for all documents on the subject, say I_j, then all and only those documents actually indexed under I_j would be selected for retrieval. However, there might be other documents *not* indexed under I_j but under a different term I_k (one which is close, in meaning, to I_j), and some (or all) of these documents might be relevant, relative to the user's information need. If the library system could compute measures of *closeness* between all pairs of index terms, then it might automatically "elaborate" on that initial request and thus retrieve relevant documents which would not have been obtained by the initial request. Furthermore, the system could assign weights to those additional documents retrieved and could rank the final output of documents according to the numerical value of these weights.

All sorts of complex and important questions can be raised about how best to perform associative retrieval. By providing the capability to do associative retrieval, students can investigate such questions as the following: Under what conditions does any one such measure produce better results than any other measure? How can one decide which measures are best for certain kinds of requests and why they perform best? Given a retrieval system that allows associative retrieval, how should this capability influence the way a user should *formulate* a request? Associative retrieval always broadens a search - it does not narrow. And it can sequentially broaden a search in different ways. What are the implications of this broadening influence? At what stages in the search should elaboration take place - and then which of the different directions of elaborating should best be pursued? We are building an Information Processing Laboratory so that advanced students in librarianship may learn about these kinds of questions and begin to get good answers for them.

In section 5 we discuss our associative retrieval work in detail.

4.5 Intellectual Access; Complexity and Understanding

Here we want to emphasize why formal principles of intellectual access should be taught in an *on-line* Laboratory. Formal techniques for intellectual access can be thought of as tools for enabling one to identify and retrieve documents relevant to a request. They can be interpreted as tools for performing *logical work* on stored data. These are complex tools and a complex intellectual process is required on the part of students in order to fully grasp and understand the nature of these tools. We believe that the standard, traditional way of teaching students about this class of logical tools is inadequate and that tools of intellectual access can best be taught, learned, and comprehended via an on-line Laboratory.

The understanding of logical tools for intellectual access comes only with an understanding of what happens when these tools are used. To understand is to understand actual as well as potential consequences of use. And this kind of understanding of "logical behavior" can come only with exposure to the use of such tools under a proper variety of representative circumstances. These tools must be exercised - put through their paces - by students, and the best vehicle for doing so is the digital computer.

All of the logical tools of intellectual access can be programmed as part of a system so that they can be "called up" and used on a variable set of documents. The results of such use can be studied by still other comparison techniques with the assistance of the computer. Thus we see the computer, in an on-line mode, as the ideal vehicle to study tools for intellectual access.

4.6 Machine Tutorial Mode

In planning this Laboratory, we decided to automate, to whatever extent was reasonable, the presentation and instruction in the use of the materials on the subject of intellectual access. However, we could not implement a "course" on methods of intellectual access until we had developed the necessary course content, i.e., until we had the search languages, data bases, association files, evaluation routines, etc. But we realized that we could proceed in parallel to develop tools for the on-line study of intellectual access and also computer-assisted-instruction techniques, if we selected for the content of the latter a topic in traditional librarianship. We selected the topic of subject cataloging. (See Part 6 for details concerning our decision to start with this particular topic within traditional librarianship.)

Our justification for the decision to present course material in the machine tutorial mode was the following: In addition to the possibility of a more efficient introduction to the subject, the student could get at the same time an introduction to on-line information processing, a hands-on experience in interacting with a system via a remote terminal. This experience would positively influence his thinking about on-line terminal processing in other facets of librarianship. We felt that this machine tutorial approach could be especially valuable for those students who were entering the Ph.D. program without having first completed a standard MLS degree. Also we felt that whether or not such a computer-assisted course in subject cataloging were ever adopted, the process of analyzing, synthesizing, and restructuring existing course material for machine presentation could be beneficial. The results of this kind of analysis when fed back could make the course better when presented again by the regular instructor.

5. ASSOCIATIVE RETRIEVAL

5.1 Introduction and Summary

5.1.1 Initial Remarks

As we stated earlier in this report, one of the major objectives of the laboratory is to offer facilities for teaching, demonstrating, and experimenting with techniques of intellectual access. There are many of these techniques that lend themselves to demonstration in the laboratory; we chose associative retrieval as the first technique for which we would develop laboratory tools.

5.1.2 Types of Word Association

Associative retrieval methods are based on the relationships which exist between terms used as content identifiers for a collection of documents. These relationships may be semantic, syntactic, or statistical. Semantic associations of terms depend upon the meaning of terms as governed by common usage - the meaning which is found in the dictionary. Thus "sofa" and "couch" are semantically associated because of their equivalence of meaning. Semantic word associations have a universal validity in that they rely upon the relationships of meaning which exist in language itself.

Syntactic associations between terms take into account the context in which terms appear - usually a single sentence or phrase. These associations depend on the positions of terms in a string of text as determined by the rules of grammar. Thus, in the grammatical expression "freedom of speech," "freedom" and "speech" are syntactically associated. Syntactic associations are not universally valid, but depend on the structure of the sentence under examination.

Statistical associations treat terms as discrete, isolated entities having no semantic or syntactic connection with other terms. This kind of association between terms is dependent upon the statistics of term usage within a given document collection. Terms which are statistically highly correlated with one another in actual usage are considered to be associated. For example, in a computer sciences collection, "data" and "processing" may be highly associated statistically since they are found together frequently in the literature of computing. Statistical word associations do not have a universal validity; they depend entirely on the way terms are used in a particular collection, and may be applicable only to the collection from which they are drawn.

In summary, semantic word associations reflect the absolute meaning of words within the context of language as a whole. Syntactic word associations reflect the proximity of words in grammatical structures within the context of a single sentence or phrase. Statistical word associations reflect the usage of words within the context of a given document collection. This is a larger context than a single sentence, but a smaller context than language as a whole, and at this level word association is believed to reflect the association of words which is made in the discussion of concepts or topics - associations not of synonymy or of proximity, as with semantic or syntactic associations, but associations based on the fact that words are used together to express complex concepts. To this end, it is best applied to a homogeneous document collection of specific and well-defined subject scope.

5.1.3 Determination of Statistical Associations

The associative retrieval tools developed thus far in the laboratory are based upon statistical associations. The attempt to determine semantic and syntactic word associations is an involved task requiring either human judgment or rather complex computer programs. The determination of statistical word associations, on the other hand, is a straightforward task which is very well suited to the capabilities of the digital computer, since it requires only the ability to compare words and to count. Statistical association techniques are ideally suited to a file of indexed documents wherein the content of each document is represented by a set of discrete index terms; this permits very simple matching and counting of index terms.

Various methods have been developed to measure the statistical correlations between terms. Basically they all rely upon the counting of the number of occurrences of single terms and the number of co-occurrences of all possible pairs of terms. Using this data, quantitative measures of association between two terms can be computed. Generally, these association values are some function of:

1) The number of documents to which each of the two terms under consideration have been assigned.

2) The number of documents to which both terms have been assigned.

3) The total number of documents in the collection.

There are several measures based on these parameters which have been proposed in the literature. The underlying idea in many of these measures is to compare the number of documents actually indexed with both terms A and B against the number of documents one would expect to be indexed with both terms A and B based on the frequency of occurrence of term A in the file and the frequency of occurrence of term B in the file with the assumption that terms A and B occur independently of one another. This may be clarified by an example: Suppose a file of 100 documents dealing with information science contains 30 documents indexed with the term SEARCHING, i.e., 30% of the file is indexed with SEARCHING. Assume 20 documents in the entire file are indexed with SEMANTIC. If SEARCHING and SEMANTIC occur independently in the file then, on the basis of their individual occurrences, we could expect 30% of 20 or six documents in the file to be indexed with both SEARCHING and SEMANTIC. However, if we observe that 19 documents in the file are indexed with both these terms then we may reasonably conclude that there is a strong relationship or "association" in this particular file between these two terms.

Numeric measures of association may be computed by applying various statistical formulas, each corresponding to a specific measure, to the term occurrence and co-occurrence data observed in the file under consideration. From these computations, tables of term association data may be formed that show the degree to which each index term used in the file is associated with other index terms. It is not feasible or desirable to create tables that contain all this word-association data; instead, such tables generally list for each term only those few index terms that have been found to be most highly associated with it. Again, these terms are associated because of their frequency of co-occurrence and not because of their meaning; hence, two terms may be highly associated in the tables and yet have no apparent similarity of meaning.

5.1.4 Associative Retrieval

Associative retrieval techniques attempt to overcome the difficulties which result from imperfect indexing. Using index terms to represent the content of a document is very common practice. However, it is a widely recognized fact that indexing is often inadequate for various reasons: applicable terms may not be assigned through oversight, the indexing may be too shallow to represent all the concepts in a document, and the terms assigned may be ambiguous.

In a system which relies on a "direct match" between the terms given in a question and the index terms preassigned to a document, relevant documents may be missed because of imperfect indexing.. The probability of retrieving documents which are relevant to the question but which are not indexed with the exact terms of the question is increased by adding associated terms to the question. For example, consider the case where "co-occurrence" is statistically highly associated with "word frequency". A requestor may ask for documents about "word association" based on "word frequency". He may fail to retrieve relevant documents which are not indexed with "word frequency" but which are indexed with "co-occurrence". In associative retrieval, "co-occurrence" would be added to the request thereby permitting the retrieval of additional relevant documents.

The method is not foolproof since it is based on probability rather than certainty. Because of this, the method has the disadvantage of causing the retrieval of some non-relevant documents and thereby decreasing precision (i.e., the ratio of relevant retrieved documents to total retrieved documents). However, as indicated above, the method does improve recall (i.e., the ratio of relevant retrieved documents to total relevant documents in the file), and it has the further advantage of being fully automatic. The association tables are generated automatically by the computer, and the query is automatically elaborated in a direction which has a high probability of retrieving additional relevant documents.

In addition to improving recall and being fully automatic, statistical word association techniques permit the computation of degrees of relevance of the documents retrieved to the questions asked. The association between a pair of terms is usually expressed as a "correlation coefficient". Using these coefficients of association between terms, the machine can compute the degree of match between the terms of the original request and the terms assigned to a document. This number will reflect the "closeness" between a document and the request, and hence may be considered the "computed relevance number" for the document. When several documents are retrieved in response to a request they may be ranked according to their "computed relevance numbers."

The technique of document retrieval based on statistical term associations is well suited to on-line presentation in an information processing laboratory. The method displays the potential of the digital computer in mechanized literature searching. It displays the power of word association in elaborating on requests and thereby enhancing retrieval. Several different measures of association can be implemented for demonstration and comparison. Search parameters and association files can be varied to produce different search results. Analysis of these results and the reasons behind them can be a rich field of study for users of the Laboratory.

The specific tools relating to associative retrieval that have been developed during Phase I are discussed in sections 5.2 and 5.3.

5.1.5 Teaching Associative Retrieval On-Line

Although the details of the procedures for teaching associative retrieval on-line have not yet been fully worked out, we can indicate some of their general features. First of all, of course, through appropriate lecture and independent reading, the student will have learned about the basic notions of associative retrieval and the various measures of statistical correlation that are used to measure degrees of closeness between index tags. They will have been introduced to the Laboratory, to terminal processing, and to the experimental corpus of stored documents upon which the various methods of intellectual access will be "exercised". They will have been introduced to the indexing scheme used to identify the content of the corpus documents and they will have been introduced to the various ways of posing requests to the system.

Armed with this background and initial understanding, the student is given a topic, expressed in some detail in ordinary language, which describes an information need. His problem is to "translate" the description of the information need into a computer search specification, which in turn, will select all and only those corpus documents that will be relevant to the stated need. In response to his search query the system will display the titles, or an abstract, or possibly the list of cited papers as well, for each document that is selected. After some iteration during which time the student has modified his query, he is satisfied with the set of items that the system has delivered; i.e., he has analyzed the output and he is satisfied that he has selected the best class of stored items. (Perhaps he has already been told exactly how many relevant documents are in the file relative to his search problem.) We should emphasize that during this process of interaction the computer not only can select and display selected documents immediately in response to a request, but it can be programmed to display only the differences in retrieval that a modification in a query makes. The student can see immediately how different query formulations "cause" different classes of documents to be selected.

The student is now ready for associative retrieval. He may initiate associative searching using each of the many different measures of statistical correlation that will be available. For each particular measure he will see the consequences that follow from the use of that measure. He will see how associative retrieval widens a search and why therefore the initial query should start in a more narrow formulation. The student begins to develop an understanding of the meaning of various measures of statistical association and of their use in associative retrieval by observing the behavior of the on-line search system, i.e., by studying the immediate consequences of their use.

5.2 Current Status

5.2.1 Components of the Laboratory

The major components of the Information Processing Laboratory upon which the on-line teaching of associative retrieval rests consist of the following:

1) A collection of documents

2) Files of bibliographic records stored on disc that correspond to the document collection

3) Files of term association data on disc

4) Computer search programs employing methods of associative retrieval. (See Section 5.3)

These also provide the basic framework within which researchers may conduct on-line experiments in associative retrieval in the Laboratory. The sections that follow describe these major components in detail.

To create and maintain the various data files on disc, certain utility programs were written. A useful by-product of many of these programs are printed listings of various data that have helped both students and staff in studying different elements of the system. The utility programs developed in support of the primary associative retrieval software are listed in Appendix 6.

5.2.2 Document File

As a data base for illustrating associative retrieval, we selected a collection of documents that had already been assembled for another project at the Institute. This is a collection of 284 journal articles published since 1957 that deal with various aspects of information science. There were several reasons why this file was attractive. First of all, it already existed; a lot of effort would be saved by using it. Secondly, the fact that it is a file on information science offered two major advantages over other files that might have been considered. One advantage lay in the fact that the content of the documents would be of great interest to the students. In addition, this file would be easier for Institute staff to index in depth than would be a file covering subject matter with which the staff was generally unfamiliar. Finally, this collection of 284 documents seemed to be of suitable size. On the one hand, it seemed large enough to adequately display the principles of associative retrieval, while, on the other hand, it was small enough that it could be indexed and later processed by computer at a reasonable cost.

At the time we selected this file much bibliographic information about each document had already been gathered and keypunched. This included title, author, publisher, journal name, editor, year of publication, and many other items. (For a complete list of these items of bibliographic information see Appendix 7.) However, this collection had not been indexed in depth. To be of use in illustrating associative retrieval the documents would have to be indexed.

While indexing may be carried out using the natural language of a document, we decided to use a controlled list of terms from which all indexing would be done. There were several reasons for this. It would reduce confusion arising from synonyms. It would help users formulate search requests in that the user would have a way of knowing in advance what terms were valid in the system and, therefore, to what set of index terms he was limited in posing requests. Finally, the use of a controlled list would concentrate the assignment of index tags in such a way that the co-occurrence data for pairs of terms would be more favorably distributed for the purposes of illustrating associative retrieval than would be the case with uncontrolled assignment of terms. To state this another way, indexing using the natural language of the document would probably result in a large number of terms, a high portion of which would be very lightly posted, thus yielding very low co-occurrence counts for nearly all pairs of terms. However, with a controlled list of terms, a certain portion of them would be expected to be rather heavily posted, resulting in rather large co-occurrence counts for certain term pairs. This distribution of co-occurrence data increases the likelihood of obtaining coefficients of association of such magnitude as are required for the effective demonstration of associative retrieval.

5.2.3 Subject Authority List

At the time we decided to index from a controlled list of terms, no such list of information science terms was available. We had to form our own list. We did this by examining a substantial body of material in the subject field, selecting candidate terms, identifying synonyms and related terms, and creating an authority list accordingly. Our collection of 284 documents was sufficiently comprehensive to contain candidate descriptors from which a rather complete information science authority list could be formed. Therefore, we went through the text of all 284 documents selecting those words and phrases that we considered to be likely candidates for the authority list. Clearly a list created from these candidate terms would be adequate for indexing the collection itself; it could also be expected to be applied satisfactorily to other collections of information science documents.

In choosing candidate descriptors, no effort was made to curb the selection of the same term or phrase repetitively. As a result some 20,000 candidate descriptors were selected initially. Keypunching these and using the computer to weed out duplicates, the list was reduced to about 10,000 descriptors. The reason for this seemingly very large number of unique candidate terms was that there had been a high proportion of phrases chosen as opposed to single words. This list of 10,000 terms was examined carefully, with close attention paid to synonyms and the relationships between candidate descriptors, and from it a final list of approximately 350 terms was formed. Generally speaking, each single word or phrase descriptor expresses a single concept. These were included because it was felt that it would be easier for users of the Laboratory to deal with a limited number of common topics as single terms rather than as combinations of terms.

The list we formed to guide us in indexing has SEE and SEE ALSO entries as well as some amplifying SCOPE NOTES. For this reason we refer to the list as a "subject authority list" rather than simply an "index term list." It is shown in Appendix 4. Because of certain computing considerations the maximum length of each index term is 16 characters.

5.2.4 Indexing the Collection

Using our subject authority list, we indexed the 284 documents in our collection, assigning an average of 15 terms to each document, with a few documents being assigned as many as 50 terms. Each document was indexed to cover its primary and secondary topics. While the index terms were assigned primarily on the basis of the content of the article, the words appearing in the title were also considered during the indexing operation.

Several people participated in indexing the document collection. Recognizing that one of the weaknesses of manual indexing is the inconsistency that arises not only from one indexer to another but even within the work of one individual indexer, we held frequent meetings to discuss the indexing. To promote consistency some documents were indexed by two people independently and the results compared and discussed with all the indexers. While we cannot vouch for the consistency and accuracy of our indexing, we believe it to be adequate for the present needs of the Laboratory. Appendices 4b and 4c contain index terms sorted on frequency of assignment and on alphabetical order, respectively.

5.2.5 Master Bibliographic Files

The information about the document collection is stored in two distinct files on disc. One file, BIBLIO, contains all bibliographic items except the index terms. The other file, MASTERI, contains the index terms. Both of these files are (in the terminology of the IBM-360 Operating System) indexed sequential files of 80-column card images. Each document is represented by several card images. In each file the "key" of each 80-character record is the 5-character document accession number.

The reasons for storing and maintaining this information in two separate files have to do with speed of searching and speed of display at the remote terminal. The associative retrieval search programs examine only the index terms assigned to a document to determine whether or not the document is to be retrieved in response to a user's query. If all bibliographic information about the document collection were maintained in a single file of card images, the search programs would spend a substantial amount of time reading and passing over card images that had no bearing on the search itself. With the volume of information we have and with the card format we use, we are able to search MASTERI serially from start to finish in one fourth the time that would be required if MASTERI and BIBLIO were merged into a single file. In addition, when using comparatively slow mechanical terminals (a display rate of 15 characters per second is typical), it is desirable to limit the output to a few characters per retrieved document. Otherwise many searches will result in such long output times that the user-system interaction is seriously impaired. Our search programs generally list only the retrieved document's accession number and "computed relevance number." No information from the BIBLIO file is needed to provide this limited output. One of the Laboratory search programs enables one to search for articles by author name. When this mode of search is called for, BIBLIO is read from start to finish and MASTERI is ignored. The following pages show samples of MASTERI and BIBLIO.

FIG 1: SAMPLE PRINTOUT OF BIBLIO FILE

```
B052601AU1 SWANSON, D.R.
B052602JO1 LIBR Q VOL.35,NO.1
B052603JA1 THE EVIDENCE UNDERLYING THE CRANFIELD RESULTS
B052604MD  3
B052605CO1 JAN
B052606YR1 1965
B052607PP1 1-20
B052608CI1 A72
B052609BI1 A87
B053301AU  DYSON, G.M.
B053302AU2 COSSUM, W.E.
B053303AU3 LYNCH, M.F.
B053304AU4 MORGAN, H.L.
B053305JA  MECHANICAL MANIPULATION OF CHEMICAL STRUCTURE
B053306JO  INFORM STOR RETRIEV VOL.1,NOS.2-3
B053307MD  5
B053308CO  JULY
B053309YR  1963
B053310PP  69-99
B056701AU1 KLEMPNER, I.M.
B056702JO1 AMDOC VOL.15,NO.3
B056703JA1 METHODOLOGY FOR THE COMPARATIVE ANALYSIS OF INFORMATION
B056704JA2 STORAGE AND RETRIEVAL SYSTEMS..  A CRITICAL REVIEW
B056705MD  3
B056706CO1 JULY
B056707YR1 1964
B056708PP1 210-6
B056709RE1 A107
B056710RE2 B526
B058101AU  FELS, E.M.
B058102JA  EVALUATION OF THE PERFORMANCE OF AN INFORMATION-RETRIEVAL
B058103JA2 SYSTEM BY MODIFIED MOOERS PLAN
B058104JO  AMDOC VOL.14,NO.1
B058105MD  3
B058106CO  JAN
B058107YR  1963
B058108PP  28-34
B058109RE  A106
B058110RE2 A111
B060101AU  JOHNSON, L.R.
B060102JA  AN INDIRECT CHAINING METHOD FOR ADDRESSING ON SECONDARY KEYS
B060103JO  CACM VOL.4,NO.5
B060104MD  19
B060105CO  MAY
B060106YR  1961
B060107PP  218-22
B060108RE  A80
B063801AU1 BROWNSON, H.L.
B063802JO1 SCIENCE VOL.132,NO.3444
B063803JA1 RESEARCH ON HANDLING SCIENTIFIC INFORMATION
B063804MD  1
B063805CO1 DEC
B063806YR1 1960
B063807PP1 1922-31
```

Columns 7 and 8 are the code identifiers of the document attributes. See Appendix 7.

FIG 2 : SAMPLE PRINTOUT OF MASTERI FILE

```
A013101LDACCESS                ALGORITHM        ASSIGNED          COST
A013102LDDATA                  FILE             INFORMATION       LANGUAGE
A013103LDLIST                  NOTATION         OPERATION         PROCEDURE
A013104LDPROG. LANGUAGE        PROGRAM          SETS              STORAGE
A013105LDSTRING                STRUCTURE        SYNTAX            SYSTEM            *
A013106LDVARIABLE

A013201LDACCFSS                ALGORITHM        COMMUNICATION     COMPUTER
A013202LDCONTEXT               DATA             FLOW OF INFO.     GENERATION
A013203LDGRAMMAR               INFO. RETRIEVAL  INFORMATION       INTERPRET
A013204LDNATURAL LANGUAGE      OUTPUT           PARSE             QUESTION-ANSWER
A013205LDRFLEVANT              SEMANTIC         STORAGE           SURVEY
A013206LDSYNTACTIC ANAL.       SYNTAX           SYSTEM            TIME-SHARING      *
A013207LDTRANSFORMATION

A013301LDABSTRACTING           ALGORITHM        ANALYSIS          COMP LINGUISTICS
A013302LDCONFFRENCE            EDITING          EVALUATION        INFO. RETRIEVAL
A013303LDLINGUISTIC            LOGIC            MATCH             NATURAL LANGUAGE
A013304LDPARSE                 PROG. LANGUAGE   PROGRAM           QUESTION-ANSWER
A013305LDSYMBOLIC LOGIC        TFCHNICAL        TIME-SHARING      TRANSLATION       *

A013401LDALGORITHM             COMPUTER         CONFERENCE        ERROR
A013402LDINTERPRET             MAN-MACHINE      MATHEMATICAL      NATURAL LANGUAGE
A013403LDNOISF                 NOTATION         PROG. LANGUAGE    PROGRAM
A013404LDREDUNDANCY            SEMANTIC         SOFTWARE          SYNTAX
A013405LDSYSTEM                TRANSLATION      USER              WORD              *

A013501LDACCESS                BIBLIOGRAPHY     CENTERS           CIRCULATION
A013502LDDOCUMENT              FLOW OF INFO.    GENERAL           INFO. RETRIEVAL
A013503LDLIBRARIAN             LIBRARY          MECHANIZATION     REMOTE TERMINAL
A013504LDRESEARCH              SCIENTIFIC       SEARCHING         SERVICE           *
A013505LDTECHNOLOGY            TRANSMISSION

A013601LDACQUISITION           ANALYSIS         CIRCULATION       COMMUNICATION
A013602LDLIBRARY               MEASURF          MEETING           PATTERN           *
A013603LDRETRIEVAL             SERVICE          SYSTEM

A013701LDACCESSION NUMBER      BOOK             CLASSIFICATION    LIBRARY           *
A013702LDRETRIEVAL             SIZE             SUBJECT

B001201LDAUTO ABSTRACTING      BIBLIOGRAPHIC    COMMUNIC  ION     LANGUAGE
B001202LDLINGUISTIC            NATURAL          STORAGE           SYSTEM            *
B001203LDTRANSLATION

B001301LDABSTRACTING           ASSOCIATION      CLASSIFICATION    DATA
B001302LDDICTIONARY            FREQUENCY        INDEX             INFORMATION
B001303LDLIBRARY               LITERATURE       MICROFILM         NETWORK           *

B001401LDDOCUMENT              INDEXING         INFO. RETRIEVAL   MICROFILM
B001402LDSCANNING              STORAGE          TERM              TRANSLATION       *

B001501LDAUTOMATION            COMMUNICATION    DISSEMINATION     DOCUMENT
B001502LDINFO. RETRIFVAL       INFORMATION      INPUT             OUTPUT
B001503LDQUESTICN              RETRIEVAL        SIGNIFICANCE      THESAURUS         *
```

5.2.6 Term Association Files

As discussed earlier in this report, associative retrieval based upon statistical term associations involves the formation of tables of association data, one table for each measure represented. During Phase I we have created three files of word association data. The three measures we used were suggested in a paper by J. L. Kuhns entitled: "The Continuum of Coefficients of Association."[1] Without going into a detailed mathematical development, let us discuss the underlying notion behind these measures.

If two terms used to index a collection of documents occur independently in the collection (i.e., if they are both randomly distributed throughout the file), one may calculate, based on their individual frequencies of occurrence, the number of documents in which one might expect the two terms to co-occur. This expected number of co-occurrences may be called the "independence value" since it is based upon the assumption that the two terms occur independently in the file. To test the validity of this assumption, one may observe the actual number of documents to which both descriptors have been assigned and compare this with the independence value. This difference may be called the "excess over the independence value." This quantity, depending upon the statistics of a particular collection, may range from large positive integers to large negative integers. To provide a common basis for comparison and calculation involving different pairs of index terms, some sort of normalizing factor is needed to bring the coefficient of association between any pair of terms, regardless of their occurrence patterns, into the range from -1 to +1. The measures proposed by Kuhns which we have used are all of the form:

excess over independence value/normalizing factor

The formulae for computing these coefficients of association are:

$$W = \frac{X - \frac{n_i n_j}{N}}{\min [n_i, n_j]} \qquad G = \frac{X - \frac{n_i n_j}{N}}{\sqrt{n_i \; n_j}} \qquad S = \frac{X - \frac{n_i n_j}{N}}{N/2}$$

where X = the number of documents to which both terms have been assigned

n_i = the number of documents to which term i has been assigned

n_j = the number of documents to which term j has been assigned

N = the total number of documents in the entire collection

[1] In: Statistical Association Methods for Mechanized Documentation, Symposium Proceedings, Washington, D.C., 1964, pp.33-39. In the notation of that paper the measures we chose are designated W, G, and S.

Notice that all these measures are symmetric, i.e., the coefficient of association between terms i and j is the same as the coefficient of association between terms j and i . Also note that all three formulae yield a coefficient of association between a term and itself that is not equal to unity. This seems to contradict a basis notion of matching, i.e., that a term coincides in meaning with itself and is not merely "close" to itself in meaning. In building the tables of word association data used with the search programs we have overridden this characteristic of the formulae. In our tables the coefficient of association of every index term with itself equals unity.

Another characteristic of these formulae is this. The denominator of measure S is constant and equal to half the number of documents in the collection. The denominators of the other measures are an order of magnitude smaller than this. Thus coefficients of association based on measures W and G are generally much larger than those of measure S and, therefore, with certain search methods are much better suited for demonstrating associative retrieval.

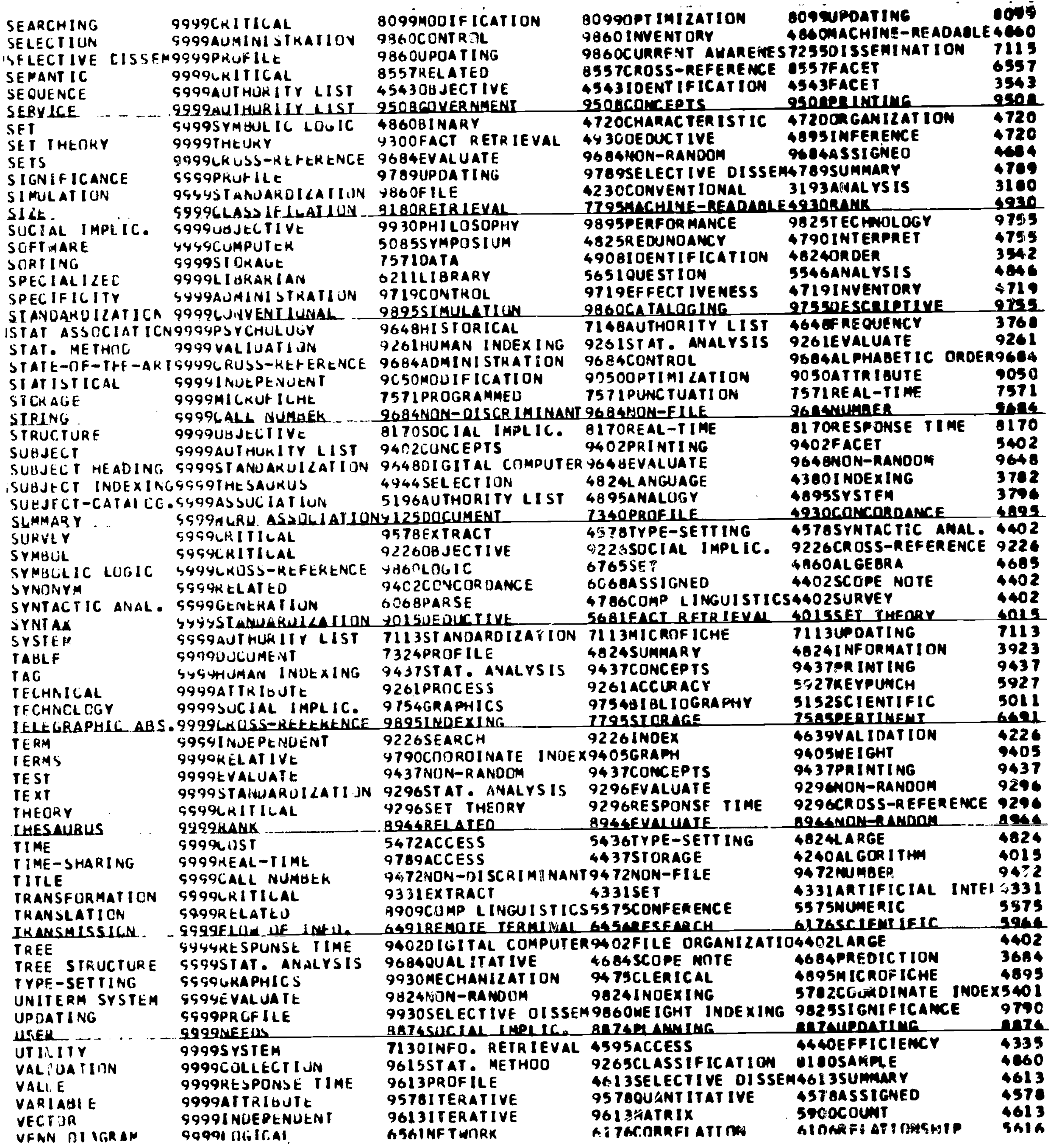
Experimentation using measure S reveals something that might easily be missed when reading a description of various measures. Because the denominator of measure S is a constant, it serves as a normalizing factor only in that it prevents the magnitude of the coefficients from exceeding unity. It does not adjust the coefficients in such a way as to give all pairs of index terms an opportunity of having a high coefficient of association. Since the denominator of measure S is a constant, the relative size of two different coefficients will be determined entirely by the numerator. The numerator is dominated by the absolute number of co-occurrences of two terms. Two heavily posted terms have a much greater probability of co-occurring frequently than do two lightly posted terms. This means that the index terms that measure S finds highly associated with a heavily posted term will, themselves, be heavily posted. Thus, when a user's request involving a frequently assigned term is expanded according to measure S, several other heavily used terms are considered. This results in the retrieval of many more documents than if W or G were used.

Typical use of measure S does produce this result, and this is a good example of how the Laboratory allows a better understanding through active, on-line experimentation. It prompts a student to examine the nature of this measure whereas he might well overlook this analysis without the "hands-on" experience in the Laboratory.

To actually generate term association files, computer programs were written that counted the single occurrences of individual descriptors and the co-occurrences of pairs of terms. Applying specific association measures to this data, the coefficients of association for all possible pairs of terms were calculated. Then a disc file was written corresponding to each measure. These files list for every index term the four other descriptors most highly associated with that term according to the specific formula used. The respective coefficients are also stored in the file. A portion of one of the association files is shown on the following page.

The word association data files are indexed sequential files of logical records of 100 characters each. Each logical record consists of five fields, one for the primary term and each of its four associated terms. Each field is 20 characters long, 16 characters for the name of the term and 4 characters for the coefficient of association. The "key" field is the 16 character name of the term. These logical records are read directly as soon as individual request terms submitted by the user are encountered.

FIG 3: SAMPLE PRINTOUT OF WORD ASSOCIATION FILE (KUHNS W)



```
SEARCHING       9999CRITICAL        8099MODIFICATION    8099OPTIMIZATION    8099UPDATING        8099
SELECTION       9999ADMINISTRATION  9860CONTROL         9860INVENTORY       4860MACHINE-READABLE4860
SELECTIVE DISSEM9999PROFILE         9860UPDATING        9860CURRENT AWARENES7255DISSEMINATION   7115
SEMANTIC        9999CRITICAL        8557RELATED         8557CROSS-REFERENCE 8557FACET           6557
SEQUENCE        9999AUTHORITY LIST  4543OBJECTIVE       4543IDENTIFICATION  4543FACET           3543
SERVICE         9999AUTHORITY LIST  9508GOVERNMENT      9508CONCEPTS        9508PRINTING        9508
SET             9999SYMBOLIC LOGIC  4860BINARY          4720CHARACTERISTIC  4720ORGANIZATION    4720
SET THEORY      9999THEORY          9300FACT RETRIEVAL  4930DEDUCTIVE       4895INFERENCE       4720
SETS            9999CROSS-REFERENCE 9684EVALUATE        9684NON-RANDOM      9684ASSIGNED        4684
SIGNIFICANCE    9999PROFILE         9789UPDATING        9789SELECTIVE DISSEM4789SUMMARY         4789
SIMULATION      9999STANDARDIZATION 9860FILE            4230CONVENTIONAL    3193ANALYSIS        3180
SIZE            9999CLASSIFICATION  8180RETRIEVAL       7795MACHINE-READABLE4930RANK            4930
SOCIAL IMPLIC.  9999OBJECTIVE       9930PHILOSOPHY      9895PERFORMANCE     9825TECHNOLOGY      9755
SOFTWARE        9999COMPUTER        5085SYMPOSIUM       4825REDUNDANCY      4790INTERPRET       4755
SORTING         9999STORAGE         7571DATA            4908IDENTIFICATION  4824ORDER           3542
SPECIALIZED     9999LIBRARIAN       6211LIBRARY         5651QUESTION        5546ANALYSIS        4846
SPECIFICITY     9999ADMINISTRATION  9719CONTROL         9719EFFECTIVENESS   4719INVENTORY       4719
STANDARDIZATION 9999CONVENTIONAL    9895SIMULATION      9860CATALOGING      9755DESCRIPTIVE     9755
STAT ASSOCIATION9999PSYCHOLOGY      9648HISTORICAL      7148AUTHORITY LIST  4648FREQUENCY       3768
STAT. METHOD    9999VALIDATION      9261HUMAN INDEXING  9261STAT. ANALYSIS  9261EVALUATE        9261
STATE-OF-THE-ART9999CROSS-REFERENCE 9684ADMINISTRATION  9684CONTROL         9684ALPHABETIC ORDER9684
STATISTICAL     9999INDEPENDENT     9050MODIFICATION    9050OPTIMIZATION    9050ATTRIBUTE       9050
STORAGE         9999MICROFICHE      7571PROGRAMMED      7571PUNCTUATION     7571REAL-TIME       7571
STRING          9999CALL NUMBER     9684NON-DISCRIMINANT9684NON-FILE        9684NUMBER          9684
STRUCTURE       9999OBJECTIVE       8170SOCIAL IMPLIC.  8170REAL-TIME       8170RESPONSE TIME   8170
SUBJECT         9999AUTHORITY LIST  9402CONCEPTS        9402PRINTING        9402FACET           5402
SUBJECT HEADING 9999STANDARDIZATION 9648DIGITAL COMPUTER9648EVALUATE        9648NON-RANDOM      9648
SUBJECT INDEXING9999THESAURUS       4944SELECTION       4824LANGUAGE        4380INDEXING        3782
SUBJECT-CATALOG.9999ASSOCIATION     5196AUTHORITY LIST  4895ANALOGY         4895SYSTEM          3796
SUMMARY         9999WORD ASSOCIATION9125DOCUMENT        7340PROFILE         4930CONCORDANCE     4895
SURVEY          9999CRITICAL        9578EXTRACT         4578TYPE-SETTING    4578SYNTACTIC ANAL. 4402
SYMBOL          9999CRITICAL        9226OBJECTIVE       9226SOCIAL IMPLIC.  9226CROSS-REFERENCE 9226
SYMBOLIC LOGIC  9999CROSS-REFERENCE 9860LOGIC           6765SET             4860ALGEBRA         4685
SYNONYM         9999RELATED         9402CONCORDANCE     6068ASSIGNED        4402SCOPE NOTE      4402
SYNTACTIC ANAL. 9999GENERATION      6068PARSE           4786COMP LINGUISTICS4402SURVEY          4402
SYNTAX          9999STANDARDIZATION 9015DEDUCTIVE       5681FACT RETRIEVAL  4015SET THEORY      4015
SYSTEM          9999AUTHORITY LIST  7113STANDARDIZATION 7113MICROFICHE      7113UPDATING        7113
TABLE           9999DOCUMENT        7324PROFILE         4824SUMMARY         4824INFORMATION     3923
TAG             9999HUMAN INDEXING  9437STAT. ANALYSIS  9437CONCEPTS        9437PRINTING        9437
TECHNICAL       9999ATTRIBUTE       9261PROCESS         9261ACCURACY        5927KEYPUNCH        5927
TECHNOLOGY      9999SOCIAL IMPLIC.  9754GRAPHICS        9754BIBLIOGRAPHY    5152SCIENTIFIC      5011
TELEGRAPHIC ABS.9999CROSS-REFERENCE 9895INDEXING        7795STORAGE         7585PERTINENT       6491
TERM            9999INDEPENDENT     9226SEARCH          9226INDEX           4639VALIDATION      4226
TERMS           9999RELATIVE        9790COORDINATE INDEX9405GRAPH           9405WEIGHT          9405
TEST            9999EVALUATE        9437NON-RANDOM      9437CONCEPTS        9437PRINTING        9437
TEXT            9999STANDARDIZATION 9296STAT. ANALYSIS  9296EVALUATE        9296NON-RANDOM      9296
THEORY          9999CRITICAL        9296SET THEORY      9296RESPONSE TIME   9296CROSS-REFERENCE 9296
THESAURUS       9999RANK            8944RELATED         8944EVALUATE        8944NON-RANDOM      8944
TIME            9999COST            5472ACCESS          5436TYPE-SETTING    4824LARGE           4824
TIME-SHARING    9999REAL-TIME       9789ACCESS          4437STORAGE         4240ALGORITHM       4015
TITLE           9999CALL NUMBER     9472NON-DISCRIMINANT9472NON-FILE        9472NUMBER          9472
TRANSFORMATION  9999CRITICAL        9331EXTRACT         4331SET             4331ARTIFICIAL INTEL4331
TRANSLATION     9999RELATED         8909COMP LINGUISTICS5575CONFERENCE      5575NUMERIC         5575
TRANSMISSION    9999FLOW OF INFO.   6491REMOTE TERMINAL 6454RESEARCH        6176SCIENTIFIC      5964
TREE            9999RESPONSE TIME   9402DIGITAL COMPUTER9402FILE ORGANIZATIO4402LARGE           4402
TREE STRUCTURE  9999STAT. ANALYSIS  9684QUALITATIVE     4684SCOPE NOTE      4684PREDICTION      3684
TYPE-SETTING    9999GRAPHICS        9930MECHANIZATION   9475CLERICAL        4895MICROFICHE      4895
UNITERM SYSTEM  9999EVALUATE        9824NON-RANDOM      9824INDEXING        5782COORDINATE INDEX5401
UPDATING        9999PROFILE         9930SELECTIVE DISSEM9860WEIGHT INDEXING 9825SIGNIFICANCE    9790
USER            9999NEEDS           8874SOCIAL IMPLIC.  8874PLANNING        8874UPDATING        8874
UTILITY         9999SYSTEM          7130INFO. RETRIEVAL 4595ACCESS          4440EFFICIENCY      4335
VALIDATION      9999COLLECTION      9615STAT. METHOD    9265CLASSIFICATION  8180SAMPLE          4860
VALUE           9999RESPONSE TIME   9613PROFILE         4613SELECTIVE DISSEM4613SUMMARY         4613
VARIABLE        9999ATTRIBUTE       9578ITERATIVE       9578QUANTITATIVE    4578ASSIGNED        4578
VECTOR          9999INDEPENDENT     9613ITERATIVE       9613MATRIX          5900COUNT           4613
VENN DIAGRAM    9999LOGICAL         6561NETWORK         6176CORRELATION     6106RELATIONSHIP    5616
```

5.3 On-Line Retrieval

The task of a document retrieval system is to select from a file those documents, and only those documents, which are relevant to a user's request. A user approaches the system with a question expressed in natural language (the language of ordinary speech). He must first translate his question into a language and form acceptable to the system. This initial step, known as "question analysis" or "query formulation" involves the translation of the concepts of the question into the indexing language used by the system. Indexing and query formulation are symmetrical processes; both translate natural language concepts into a set of terms from the thesaurus. Retrieval is then a matching of term sets. The query is input to the system, the search routines are executed, and output from the system is a list of document numbers purportedly relevant to the original question.

In the Information Processing Laboratory, retrieval takes place on-line. The user types in his request at the remote terminal. A record containing a document number with its index terms is read into the computer from the MASTERI file (on disc) and the terms are compared to the terms of the query. Where the terms match to a certain specified criterion, the document will be retrieved and its number will be typed at the terminal.

To provide a variety of ways to pose requests and thereby provide different approaches to conducting experiments in associative retrieval, three search programs have been developed. In all three programs the user may invoke any of the three word association files, or, if he wishes, he may ignore these files and search in "direct match" mode.

The three different search programs each use a different type of question format as input. These formats represent different levels of complexity in posing questions. By offering this range of choice in question formats, the system provides a flexibility which is highly desirable for its purpose of serving users with varying degrees of sophistication. The user, whether novice or specialist, can choose that method of posing questions which best suits his level of knowledge and his needs. The ability to choose different question formats also provides the opportunity for experimentation; a student may express the same essential question in different formats in order to determine the impact of varying this parameter. While varying the form of the question, he may also vary his search strategy and the association files he uses, and thus discover the optimal combinations of different approaches.

5.3.1 Search Program No. 1

In the first search program, designated LABSRC 1 (for LAB SeaRCh program 1), the user expresses his question in the form of a single list of terms. The terms are logically independent and are not structured into a Boolean expression - the logical operators "and," "or," and "not" are not used. An important parameter of LABSRC 1 is the "minimum value" or "minval" which a document must have in order for it to be retrieved. The user specifies this "minval" when he inputs his query.

In the direct match mode this minimum value is simply the number of query terms that must have been assigned to a document in order for it to be retrieved. For example, if a user made this request,

application
list
structure
natural language
translation
minval = 04.00

he is specifying that at least four of the five terms of the question must have been assigned to any document that is retrieved. He cannot specify which four must be present; any document with less than four will not be retrieved.

When he performs a search using the word association files, the coefficients of association play a role in determining a document's "value" or "weight". Each query term present in the document contributes a weight of 1.0, just as in the direct match mode. When a query term is found to be absent from the document's list of assigned terms, the four terms associated with the missing query term (as found in the specified word association file) are matched against the document. If one of these associated terms is assigned to the document, its coefficient of association will be added to the retrieval weight of that document. If more than one of the four associated terms is assigned to the document, the one having the largest coefficient of association will contribute this value to the retrieval weight of the document. This retrieval weight is the sum of the values of the individual terms, terms present in the original query contributing a value of 1.0, and associated terms contributing the value of their coefficients. After all of the query terms have been matched against the document in this way, the document's retrieval weight is compared to the minimum value specified by the user. If the document's weight is greater than or equal to the minimum value, the document is retrieved; if not, the document is rejected.

For example, suppose a document is indexed with the terms "list," "structure" and "natural language." It would not satisfy the request above if the search was performed in direct match mode, since it has only three of the query terms. However, if we use word association from the Kuhns W file, the document is retrieved because it has the term "program" associated with the query term "application" (with a coefficient of .5616), and the term "comp. linguistics" associated with the query term "translation" (with a coefficient of .5575). By adding the values of the query terms (a total of 3.00) and the associated terms (.5616 + .5575), we reach a total of 4.11, which is above the "minval" of 4.00. This document is therefore retrieved.

When searching in direct match mode, the user specifies a "minval" which is an integer since it represents the minimum number of terms which a document must have for retrieval. When searching in the associative search mode, he specifies a "minval" which can be either an integer or a decimal fraction, such as 3.5 or 4.78. In this mode he is specifying the minimum "weight" which a document must have for retrieval. The number of query terms and the coefficients of association for associated terms determine a document's retrieval weight, and the user's "minval" determines which documents should be retrieved. Thus, the coefficients of association play a role in selecting or rejecting documents in LABSRC 1.

The retrieval weight of a document can be interpreted as its "computed relevance number" since it represents the degree of match between the document and the request.

A sample printout from the 2740 terminal illustrating the operation of LABSRC 1 is given on the next page.

FIG 4: SAMPLE RUN FROM LABSRC 1

```
TERMINAL CLEAR
labsrc1
WILL YOU SEARCH ON INDEX TERMS?
yes
SELECT WORD ASSOCIATION FILE
kuhnsw
DO YOU WANT OPERATING INSTRUCTIONS?
yes
ENTER LIST OF TERMS,ONE PER LINE. ENTER 'MINVAL' AS:
MINVAL=XX.XX OR MINVAL=XX.XX* (*MEANS WORD ASSOCIATION
DATA WILL BE IGNORED).  MINVAL ENTRY LINE MAY BE ANYWHERE.
LAST ENTRY MUST BE 'END'.

application
list
structure
natural language
translation
minval=04.00*
end
FILE IS NOW BEING SEARCHED

B0260: VALUE=04.00

END OF SEARCH

WILL YOU SEARCH ON INDEX TERMS?
yes
SELECT WORD ASSOCIATION FILE
kuhnsw
DO YOU WANT OPERATING INSTRUCTIONS?
no

application
list
structure
natural language
translation
minval=04.00
end
FILE IS NOW BEING SEARCHED

A0004: VALUE=04.11

B0260: VALUE=04.00

B0638: VALUE=04.20

END OF SEARCH
```

5.3.2 Search Program No. 2

A second search program, LABSRC 2 (LAB SeaRCh program 2), provides the user with some degree of logical flexibility in posing questions. He is able to specify certain terms that *must* be present in a document, others that must *not* be present, and groups of alternative terms any one of which must be present. In this program, the user enters his query as a set of term lists which are preceded by one of the logical operators "and", "or", and "not", These lists may be thought of as question fragments. There is an implied "and" operator joining the logical query fragments represented by each list.

LABSRC 2 differs from LABSRC 1 in two major ways. In the form of the query, LABSRC 2 permits the use of several term lists with logical operators rather than a single term list without logical operators. In retrieval LABSRC 2 depends on satisfying the several conditions of the retrieval prescription, rather than depending on the matching of a document's weight with the user's minimum value. A typical query for LABSRC 2 is the following:

```
AND
    info retrieval
    experiment

OR
    recall
    precision

OR
    design
    evaluation

NOT
    cost
```

In other words, the user wants only those documents indexed with both the terms 'info retrieval' and 'experiment', plus either 'recall' or 'precision', plus either 'design' or 'evaluation', and the user does not want documents indexed with 'cost'.

The requestor may submit one AND list, one NOT list, and up to four OR lists. A maximum of 10 terms may appear in the AND list, a maximum of 5 terms in the NOT list, and a maximum of 5 terms in each of the OR lists. He may search in direct match mode, or he may invoke word association from one of the three term association files.

If a document is to be retrieved, it must satisfy for each of the three types of lists either of the two following sets of conditions (a and b) depending on whether the search is performed in direct match mode or in associative mode:

AND a) Direct match: All the terms in the AND list must be present in the document.

b) Associative: If one of the terms of the AND list is absent, one of its associated terms must be present.

OR a) Direct match: At least one of the terms in each OR list must be present.

b) Associative: If none of the terms of an OR list is present in the document, one of the terms associated with one of the terms of the list must be present.

NOT a) Direct match: The document must not have any of the terms of the NOT list.

b) Associative: The association file is not used when processing the NOT list; it is always processed in direct match mode. A document fails to satisfy the NOT list requirement only if one of the terms submitted by the user is present; there is no reference to the word association files. The presence in a document of a term that is highly associated with a term in the NOT list is not deemed sufficient cause for the document to be rejected.

In sum, in using LABSRC 2 in the direct match mode, a document will be rejected if any term in the user's AND list is missing from the document, if all the terms in any of the user's OR lists are missing, or if any term in the user's NOT list is present. In associative retrieval mode, the same rules hold except that a term in the association file may be substituted for a missing term in the AND and OR lists. Unlike LABSRC 1, it is not the "weight" of a document which determines whether or not it will be retrieved but simply the presence or absence of query and associated terms. The user does not set a "minimum value;" all documents which satisfy the retrieval conditions are retrieved.

However, the "retrieval weight" of a document is used in LABSRC 2 as a "computed relevance number." The computed relevance number does not influence which documents will be retrieved; it is merely displayed for the user's information. In associative retrieval mode LABSRC 2 calculates a weight of each list or query fragment in the following way: each term in the AND-list that is present in the document is assigned a weight of 1.0. If the query term is absent but an associated term is present, its coefficient of association is assigned as the weight for that term. The cumulative weight for the AND-list is the product of the weights of the individual terms. If a term in an OR-list query is present in the document, the OR-list is assigned a weight of 1.0. If no OR-list term is present, the OR-list is given a weight equal to the coefficient of association of that term in the document that is most highly associated with one of the terms in the OR-list. The NOT list does not play a part in the calculation. The weights of the AND-list and each OR-list are multiplied to yield a "computed relevance number" for the retrieved document that is displayed to the user along with the document accession number. In direct match mode LABSRC 2 merely lists the accession number of the retrieved documents; there is no "computed relevance number" since all retrieved documents have identical numbers, their retrieval being dependent on the same prescription of required query terms. This page contains a sample question in LABSRC 2, followed by a sample computer run.

SAMPLE LABSRC 2 QUESTION

Assume a document is indexed with the terms:

EDUCATION
PLANNING
LABORATORY
ON-LINE
BIBLIOGRAPHY
SERVICE

Further assume:

RESEARCH is associated with LABORATORY by (.8)
REMOTE TERMINAL is associated with ON-LINE by (.9)
LIBRARY is associated with BIBLIOGRAPHY by (.6)
LIBRARIAN is associated with SERVICE by (.7)

Assume a query composed of these 3 lists:

AND	OR	OR
EDUCATION	COMPUTER	LIBRARY
PLANNING	NETWORK	LIBRARIAN
RESEARCH	REMOTE TERMINAL	

This document's relevance number would be completed as follows:

Weight of AND list:
EDUCATION is present = 1.0
PLANNING is present = 1.0
RESEARCH is absent but LABORATORY is present = .8
AND list = 1.0 x 1.0 x .8 = .8

Weight of first OR list:
No term submitted by user is present but
REMOTE TERMINAL is highly associated with ON-LINE
first OR list = .9

Weight of second OR list:
LIBRARY is absent but BIBLIOGRAPHY is present
LIBRARIAN is absent but SERVICE is present
weight = max (.6, .7) = .7

The relevance number of the document is the product of the weights of the individual lists, i.e., (.8) x (.9) x (.7) = .504

FIG 5: SAMPLE RUN FROM LABSRC 2

```
TERMINAL CLEAR
labsrc2
THIS IS LAB SEARCH PROGRAM 2

DO YOU WANT TO USE WORD ASSOCIATION DATA?
no
PLEASE ENTER YOUR REQUEST

and
info. retrieval
experiment
or
recall
precision
or
design
evaluation
not
cost
end

FILE NOW BEING SEARCHED

DOC. NUMBER=A0072

DOC. NUMBER=A0087

END OF SEARCH

PLEASE SPECIFY RESTART OR EXIT
restart
DO YOU WANT TO USE WORD ASSOCIATION DATA?
yes
WHICH ASSOCIATION FILE DO YOU WANT TO USE?
kuhnsg
PLEASE ENTER YOUR REQUEST

and
info. retrieval
experiment
or
recall
precision
or
design
evaluation
not
cost
end

FILE NOW BEING SEARCHED

DOC. NUMBER=A0003     COEFF=0.00

DOC. NUMBER=A0019     COEFF=0.17

DOC. NUMBER=A0072     COEFF=1.00

DOC. NUMBER=A0073     COEFF=0.19

DOC. NUMBER=A0087     COEFF=1.00

DOC. NUMBER=A0107     COEFF=0.32

END OF SEARCH
```

5.3.3 Search Program No. 3

The final search program based on associative retrieval, LABSRC 3, is by far the most sophisticated of the three. The most distinguishing feature of the program is the dynamic interaction permitted between system and user. It provides the user complete logical flexibility in posing requests. Any logically valid combination of index tags and operators may be used. LABSRC 3 permits any level of parenthetic nesting in phrasing questions. With this program the user may emphasize certain elements of his request by assigning weights to individual terms or to parenthetic fragments of his question. Just as with the other search programs, LABSRC 3 may be used in either the direct match mode or associative retrieval mode. In the latter mode, LABSRC 3 can assign a computed relevance number to each retrieved document that can be used to rank the documents according to the closeness of match between document and request as determined by the system.

The interaction between system and user may take many forms. Among these are the following: the user may modify his request in a variety of ways; he may exercise some control over the extent to which his question will be expanded by the use of term association data; he may choose to display only a selected subset of the retrieved documents; he may alter the normal flow of control through the program. This interactive capability is achieved through the use of a special "command language" within LABSRC 3. A full description of the command language, along with other detailed information about LABSRC 3, is given in Appendix 5.

Without discussing the inner workings of LABSRC 3 in great detail, let us describe in a general fashion the way this search program operates. If the user does not invoke any options of the command language, LABSRC 3 will proceed in a "normal pass" by asking the following questions, each of which may be regarded as a major junction point in the program:

1. DO YOU WANT WORD ASSOCIATION?
2. PLEASE SPECIFY ASSOCIATION FILE.
   (Only applies to associative search)
3. DO YOU WANT SCORING?*
   (only applies to associative search)
4. ENTER BOOLEAN EXPRESSION.
5. DO YOU WANT RESULTS PRINTED?
6. SPECIFY RESTART OR EXIT.

*Scoring = computation of relevance numbers.

In a normal run the only options offered by LABSRC 3 that are not available in the other search programs are the options to suppress scoring and to choose how much output is to be displayed. Of course, the format used in entering the question is much different. A question posed to LABSRC 3 might take this form:

'RETRIEVAL' AND 'EFFECTIVENESS' AND ('RECALL' OR 'PRECISION')
  AND NOT 'COST'

Searching in direct match mode proceeds as in the other search programs; terms required by the question must be assigned to a document for it to be retrieved. Scoring does not apply when using the direct match mode. When scoring is called for in associative retrieval mode, relevance numbers are computed in the same manner as in LABSRC 2.

LABSRC 3 permits the user to assign weights to individual terms or parenthetic fragments of his question. In this way he can emphasize the importance of certain elements of the query. The weights are entered as decimal fractions in the range .0000 to .9999. In scoring, the assigned weights become multiplying factors in determining the computed relevance number. It is in this way that the relevance measure of a retrieved document reflects the closeness of the document to the user's weighted request.

In the example question above no weights are assigned. In this case every term and parenthetic expression bears an implied weight of 1.0.

The power of LABSRC 3 lies in its capacity for interaction with the user. At any point other than where the user is asked to enter his request, one may reply with a command that alters the normal flow of the program. The command may be submitted in natural language; the Laboratory user need not learn a great number of special keywords and formats to use the command language of LABSRC 3. LABSRC 3 has a text analyzer that examines commands entered by the user and identifies certain verbs and keywords to determine what command is to be initiated.

Appendix 5 lists all commands available within LABSRC 3. We cite only a few of them here to illustrate how a sequence of interactive operations initiated by the user might proceed.

Suppose after processing the above question ('RETRIEVAL' AND 'EFFECTIVENESS' AND ('RECALL' OR 'PRECISION') AND NOT 'COST') in a normal pass the user wishes to de-emphasize the term 'EFFECTIVENESS'. He may do this by responding to SPECIFY RESTART OR EXIT with the command:

ASSIGN .7000 TO 'EFFECTIVENESS'

In scoring, the weight of EFFECTIVENESS (or an associated term) will be multiplied by .7. This will result in a lower ranking for those documents that formerly had a high relevance ranking due to the presence of EFFECTIVENESS.

If the user wants to analyze why certain results are obtained, he may display the terms associated with the descriptors in his request as they appear in the term association data file he has chosen. This may be done with the command:

```
DISPLAY ALL ASSOCIATION DATA
```

If the user wishes to display just those terms associated with EFFECTIVENESS, he could issue the command:

```
DISPLAY TERMS ASSOCIATED WITH 'EFFECTIVENESS'
```

Other commands enable one to display associated terms to any desired level, e.g., the most highly associated term only or the two or three most highly associated terms. If the user wishes to modify the search in such a way that, for example, only the two most highly associated terms for each request term participate in the search, then he may enter the command:

```
MODIFY TO USE TWO MOST HIGHLY ASSOCIATED TERMS
```

As another example, if the user wanted to expand the search in the normal way according to a specified word association file with the exception that no descriptor associated with RETRIEVAL should be considered unless its coefficient of association is greater than .8, he could use this command:

```
SEARCH USING ONLY THE TERMS RELATED TO 'RETRIEVAL' THAT ARE *GT*
  .8000
```

Unlike the other two search programs, LABSRC 3 does not automatically display the accession number of each document satisfying the user's request as soon as the document is examined. Instead it counts the number of documents that satisfy the question and reports this number to the user. The purpose of this is to provide the user with control over the volume of the output. The user may then call for the display of all of them or only some of them as he desires. When associative retrieval mode is used, LABSRC 3 will, if the user wishes, sort the retrieved documents into either ascending or descending order according to the values of the computed relevance numbers. This feature is optional rather than fixed because there are certain experiments (such as comparative studies) in which a student may want the listing in accession number order.

Another powerful feature of LABSRC 3 is the parser used to process the Boolean expressions that constitute search requests. The parser produces directly executable code; there is no intermediate form generated. The logic of the original question is embodied in this code. There is no need to invoke an interpretive technique each time a new document is examined. This results in very swift searching of the document colle ,ion. This is quite important in a man-machine dialogue where system effectiveness depends upon good system response time.

A sample run using LABSRC 3 is shown on the following page.

2

FIG 6: SAMPLE RUN FROM LABSRC 3

```
TERMINAL CLEAR
labsrc3
LABSRC3 AT YOUR SERVICE

Q01 - DO YOU WANT WORD ASSOCIATION?
yes
Q02 - PLEASE SPECIFY ASSOCIATION FILE
kuhnsw
Q03 - DO YOU WANT SCORING?
yes
Q04 - ENTER BOOLEAN EXPRESSION
('acquisition' or 'circulation' or 'cataloging')@
and 'library' and ('computer' or 'automation'@
or 'mechanization') and not 'indexing'
FILE IS NOW BEING SEARCHED

007 DOCUMENTS SATISFY EXPRESSION

Q05 - DO YOU WANT RESULTS PRINTED
yes

A0022  .802

A0029  .802

A0030  .787

A0051  .802

A0135  .857

B0054  .591

B0468  .878

Q06 - SPECIFY RESTART OR EXIT
search using no association

FILE,IS NOW BEING SEARCHED

002 DOCUMENTS SATISFY EXPRESSION

Q05 - DO YOU WANT RESULTS PRINTED
yes

A0135

B0468

Q06 - SPECIFY RESTART OR EXIT
exit
TERMINAL CLEAR
```

6. MACHINE TUTORIAL MODE

6.1 Introduction and Summary

6.1.1 Initial Remarks

As explained in section 4.3 we decided to parallel the development of the on-line study of intellectual access with investigation of computer-assisted instruction techniques, since the latter are expected to play an important part in the interactive process between future library systems and patrons. As such, the techniques themselves are appropriate objects of study within the laboratory setting. Moreover, work in this area could fittingly deal with segments of instruction in traditional librarianship, thus serving not only to demonstrate the medium but to put it to immediate use as well.

The term "CAI" has become the generally accepted designation for computer-"assisted" or "administered" or "augmented" instruction at all levels and in a variety of basic types. In order to identify the kind of CAI we had in mind, we adopted the phrase "Machine Tutorial Mode (MTM)". This term is intended to connote fairly free and flexible dialogue between a teacher and an adult student, as compared with modes better suited to elementary teaching, to computational problem-solving, and to the imparting of simple skills.

One of our first concerns was the selection of the subject matter for which MTM would be a suitable vehicle. The highest priorities of "applied librarianship" which relate directly to problems of intellectual access are cataloging, classification, indexing, reference, and bibliography. From among these we chose to deal first with cataloging, specifically alphabetical subject cataloging as practiced in U.S. libraries.

We were aware that the then existing techniques in CAI had not yet been extended to apply to graduate instruction except in a very few instances, and that it would be necessary for us to do a certain amount of pioneering. In particular we had to try to alter, if possible, certain constraints which might have been quite acceptable in a system intended for other uses. For example it would not have been appropriate to restrict students to rigidly formatted responses, especially since some of the concepts they would be dealing with would admit a variety of interpretations.

6.1.2 Why Subject Cataloging?

Subject cataloging was selected as the first course to be implemented in MTM because:

1) It related directly to the Laboratory's ongoing inquiry into nature of intellectual access.

2) It could be presented without resorting to graphics (diagrams, photographs, etc.,) which were outside the range of the initial hardware configuration of the Laboratory.

3) Subject cataloging deals with an entity with which almost everyone - as a library user - has had some experience.

4) It has general applicability in schools offering the Master's Degree in Library Science.

5) It is closely related to other subjects in the traditional curriculum, such as descriptive cataloging and classification.

6) It is a laboratory-type course and its implementation could possibly reduce the student-hour workload to a degree not attainable with a non-lab course (e.g., university library administration, or history of libraries.)

Also subject cataloging posed special teaching problems which we wanted to confront. Far from being a set of rote procedures, as sometimes imagined by the beginner, it involves interpretation of principles and policies which are often divergent or seem to be. In the teaching of subject cataloging it is frequently necessary, for example, to distinguish between terms which represent existing practice based on historical precedent and those terms which for the moment appear to be more rational. (That is, a difficult point to put across is that consistency is important to the user. Thus the fact that an otherwise rational term is not within existing practice tends to deprive it of some of its rationality.)

We postulated that if MTM succeeded - even partially - in the area of subject cataloging, it could be expected to succeed in others.

6.1.3 Machine Tutorial Mode; Potentials, Constraints and Background

Potentials and Constraints. We were not at all sure that CAI techniques could be molded to the needs of graduate students, because of the difficulty of attaining meaningful interaction with the students at the desired intellectual levels. However, the potential of the approach is such that we felt it to be well worth the attempt.

The potentials of the machine tutorial mode include those generally attributed to "high-level" CAI:

1) The ability to engage the student privately in an active conversational interchange composed of elements that can be predicted. Exposition is accompanied by frequent questions and other opportunities for the student to express himself, even though his answers may not in every case affect the unfolding of the presentation.

2) The ability to branch, i.e., to vary the presentation according to the machine's interpretation of student responses. Program reactions can be varied to suit categories of responses - commending those which contain desired elements and supplying corrective instruction in reply to those which contain elements of error.

3) The ability to record student performance in detail, for both individuals and groups. This ability lends itself not only to evaluation of student performance but to evaluation and refinement of the course itself.

4) The ability to administer instruction with patience and equanimity well beyond typical norms.

5) The ability to administer problem-solving exercises in a real time sense. In a course involving laboratory assignments, the system is able to provide immediate commentary on work submitted.

Counterposed against the above abilities are the constraints of the computer, chiefly its inability to cope with human dialogue for which it has not been amply programmed in advance. It cannot follow the convolutions, inversions, and implied relationships of ordinary conversation. In order to compensate for this inability - that is, to provide a modicum of flexibility in dealing with student input - a considerable programming effort is required.

Background. In trying to find models adaptable to our work, we discovered that, although many researchers in various parts of the country were exploring forms and applications of CAI systems, their efforts in this broad new field were extremely diverse. At the time work on the Information Processing Laboratory was begun, very little had been done to compare and combine their findings. There was also an almost universal tendency on the part of those active in CAI development to deal with readily definable course content in order to be able to concentrate on the behavior of the medium vis-à-vis the student. In contrast, the model we hoped to develop for the Information Processing Laboratory had to give equal attention to corpus and to the mode of transfer.

6.2 Policy Decisions

6.2.1 Course Boundaries

The first important policy decision was that of delimiting the scope of the course. Library schools do not give equal emphasis to subject cataloging, nor do they all introduce it in the same sequence as does the School of Librarianship at Berkeley. However, since our intention was to design a directly implementable segment for incorporation in an on-going curriculum, we based out approach on the Berkeley model. Briefly, this encompasses:

1) In the first quarter, an historical review of cataloging in general, consideration of subject cataloging according to LC practice and the Sears abridgement of LC practice, LC classification, Dewey Decimal classification, and a survey of other classification systems;

2) In the second quarter, descriptive cataloging according to the Anglo-American code, and filing rules;

3) In the third quarter, cataloging of special collections, plus consideration of classified systems.

In order to extract and deal with subject cataloging as a distinct element yet retain its function as an organic unit within the curriculum, it was necessary to prescribe its interfaces (particularly with descriptive cataloging and classification) with great care. We assumed that it would be administered at the very outset of the student's year, with little or no preparation other than general background information.

It was decided very early to make no attempt to link subject cataloging with classification, for the reason that except on theoretical grounds neither contributes markedly to the understanding of the other.

A brief discussion of main entries and added entries was included in order to give the student an insight into the overall catalog structure, even though the practice at Berkeley is to defer this material until dealing with descriptive cataloging in the second quarter.

6.2.2 Subject Authority

Probably no two specialists in the field of subject cataloging would agree on the ideal corpus for a text on the subject, or would agree on matters of relative emphasis, even if they could agree on corpus. This is to be expected, because subject cataloging is itself an imperfect methodology, prone to a variety of interpretations.

We decided to base the course largely on the interpretation and exposition of Library of Congress practice provided by the late David Judson Haykin in his Subject Headings, a Practical Guide, GPO, 1951.

Haykin offers reasonable statements in support of this practice without straying into apologetics for past inconsistencies in its application to the LC catalog itself, and without bemusing his readers with controversial theoretics.

This basic material is supplemented by comparisons between the LC and Sears lists of subject headings. In addition, numerous other sources, such as Needham, Coates, and Mann, were consulted in order to derive the benefit of their particular approaches to the subject cataloging problem.

6.2.3 Choice of CAI Programming Language

Some CAI programming is done using assembly level computer languages. But for the most part so-called "CAI languages" are developed using existing higher level languages such as FORTRAN or PL/1. PILOT (Programmed Instruction Learning Or Teaching), written in a "selected character string matching" (i.e., keyword matching) language or PL/1, is an example. This new and powerful high-level CAI language was under intensive development at the University of California Medical Center, San Francisco, at the time we commenced work on MTM. After reviewing other high level languages such as COURSE-WRITER, PLANIT, LYRIC, MENTOR, and CAL(UCI), we decided that PILOT, as designed, would provide the best combination of features desired for the MTM program. The fact that it was developmental during most of the period covered by this report has afforded an opportunity for us to participate in refining some of its specifications through the continual process of reporting back our operational experience in working with it. (See R. Karpinski, et al. "PILOT; a Conversational Language - User's Guide." Office of Information Systems, University of California Medical Center, San Francisco, California. December, 1968.)

6.3 Current Status

6.3.1 Components of the MTM Program

The MTM program within the Information Processing Laboratory contains the following three components:

1) a machine-administered course in subject cataloging;

2) a machine-administered laboratory in subject cataloging to accompany the above course;

3) an empirical methodology for preparing MTM courses in other subjects.

The first item is complete except for final editing and refinement of scoring mechanisms. The second item is half completed: the overall design and organization and the specification of standard subroutines has been accomplished while implementation of unique book problems is one-third complete. Once this is finished, the laboratory supplement must be studied and revised in an operational context. The third item has been formulated and defined.

Each of the three components is discussed separately in the following pages.

6.3.2 Equipment

The MTM program uses two mechanical terminals located in the Information Processing Laboratory: an IBM 2741 and a Teletype Model 35, both driven by an IBM 360/50 computer with a 256K memory located at the University of California Medical Center in San Francisco. Linkage is through acoustic couplers and commercial grade telephone lines. Other system hardware is described in Appendix 3.

From the beginning, we have regarded these mechanical terminals as an interim installation only, since we envisaged the ultimate installation as consisting of cathode-ray tube consoles which are preferred for their high display rates and quiet operation. Since it has been necessary to present text in fairly large blocks in order to deal with the conceptual material in the subject cataloging course, it has become even more evident that CRTs have a commanding advantage over mechanical terminals. A three-console CRT system is slated for installation in January, 1969.

6.4 MTM Course in Subject Cataloging (201x)

6.4.1 Presentation

The material is presented in 253 frames of varying length and intricacy, joined by logical connectives - cause/effect, elaboration, example, contrast, exceptions, ramifications, etc. - intended to lead the student's attention continually onward toward the end of the course. Some of these connectives are stipulated in the text, while others are readily inferred. An example of a connective stipulated in the text is the following statement and question: "Subdivided headings received a great deal of attention in Sears' Suggestions for the Beginner. Do you remember the four basic types we discussed earlier?"

Each frame begins with a statement, followed by a question or requirement, followed by the student response (or, input), followed by the program reaction. If the student answered correctly, the program will pass the student onto another frame. If the student has answered incorrectly the program will give the student some cues and then go back to student input mode.

In order to accommodate student responses which the course author has failed to anticipate or has deemed too remote or too general to include in his list of answers wanted or not wanted, a 'carry-all response' is contained in each frame. It is worded in such a way as not to be non-sequitur (we hope) and to return the student back to the question by strengthening the basic cue. This carry-all response might be simply "I don't understand you. Please reword your answer." Or it might say "Try again, remembering what we said about ______(the subject being discussed)." A frame usually contains one or more prompting cues for the student who gives wrong answers; if he fails the second or third time, the program gives the correct answer and the student is passed onto another frame. Appendix 8 contains a sample program-student interaction sequence.

The course unfolds in five broad movements, as follows:

1. First, the student is confronted with the problem of indexing theory, i.e., how does one represent a collection of documents in such a way that a literate person may approach it via many different terms?

The first third of the course is taken up with this problem and measures designed to alleviate it.

2. Next, some of the technical problems associated with the above measures are discussed, such as the need to limit certain subject blocks by judicious scattering.

3. This leads to careful consideration of the idea of subdivision as a control device. The four types of subdivision (chronological, geographical, topical and form) are discussed, and students are made aware that their misuse leads to classificatory treatment inconsistent with the idea of dictionary arrangement.

4. Up to this point in the course, LC practice has been frequently cited, and the LC list has been offered as a kind of bellwether to the flock. Now a process of reviewing and extendi^g some of the earlier instruction is initiated, in order to reinforce and enrich the student's understanding. This is done through comparison of the Sears approach with the fuller LC approach to various problems.

5. Finally, special situations are touched upon, such as the effect of ethnic factors on formulation of "literary" headings. This concludes the course.

6.4.2 Teaching Strategy and Tactics

The elements of strategy and tactics we considered in the course are the following: 1) style of presentation; 2) latitude of student responses; 3) continuity (with respect to provisions for reviewing material and for signing in and signing off); 4) student aids. Each is discussed separately below.

1. Style of Presentation. The linearity of the machine-tutorial mode calls for a strong narrative line of presentation in order to capture and retain the student's attention from beginning to end. At the same time, however, one wants the student to have a feeling of real participation in his own instruction. Since the course is to be given to a highly diverse set of students, we followed a median path between 'guidance' and thought-provoking questions.

Elements of the 'conversational mode' are used in the course to give the student the feeling of individual tutoring and to lessen the feeling that he is conversing with a computer rather than a human. We believe that textual statements should be models of clarity and precision and should err, if at all, on the side of saying too little rather than saying too much. Then, if the student fails to grasp the idea of the statement, the fact will be apparent in the ensuing interchange, and the program will supply correction and clarification as necessary. This kind of presentation contrasts sharply with the more diffuse style of a typical classroom lecture which must provide universal correction, elaboration, and reinforcement as it goes along.

The student is usually aware from the start that the computer cannot converse with him in human fashion. He should be brought to appreciate, however, that a human teacher is nevertheless trying to communicate with him personally through the computer medium, and that the progress of the interchange is dynamic and exceedingly variable. (Over 15 million variant paths through 201x are possible, excluding review sequences.) He should be informed, moreover, that many of his responses will be saved and later studied by the human teacher in order to refine and extend the program's interactive capability.

To illustrate, here is part of the introduction to the course:

> YOU ARE NOW ENROLLED IN MTM COURSE 201x: "SUBJECT CATALOGING."'
> MY NAME IS MR. MEREDITH, AND I AM USING THE COMPUTER IN THE SAME WAY YOU ARE: AS A MEANS OF COMMUNICATION, ALMOST LIKE A TELEPHONE (ALBEIT A VERY COMPLICATED ONE.) I HOPE YOU WILL EXCUSE INSTANCES WHEREIN MY SIDE OF THE CONVERSATION STRIKES YOU AS SOMEWHAT CURT OR UNFAIR. THIS MAY OCCUR WHEN I FAIL TO ANTICIPATE A PARTICULAR RESPONSE FROM YOU. FOR ONE THING, IT IS ALMOST IMPOSSIBLE FOR THE PROGRAM TO COPE WITH DOUBLE NEGATIVES.
>
> THE BACKSPACE OPERATES AS AN ERASER. IF YOU MAKE A TYPOGRAPHICAL ERROR IN ANSWERING, JUST BACKSPACE TO THE PART YOU WANT TO CHANGE, AND THEN RETYPE ON THE SAME LINE, OR, IF YOU PREFER, ROLL THE PAPER MANUALLY ONE LINE BEFORE STARTING TO RETYPE. YOU MAY ALSO ERASE A WHOLE LINE BY TOUCHING THE "ATTENTION" KEY.
>
> SOME STUDENTS ARE A LITTLE NERVOUS ABOUT WORKING WITH A TERMINAL FOR THE FIRST TIME. WOULD YOU LIKE SOME PRACTICE BEFORE WE ENTER THE COURSE?

The conversational tone is heightened by using the first-person singular instead of the first-person plural; the student should never be given the feeling that he is dealing with a committee. Calling the student by name, or referring to something he said previously and relating it to the corrective or reinforcing matter to be conveyed are both good devices. The use of a homely phrase now and then does no harm, and students are delighted to discover occasional touches of humor in the programmed responses. Examples should be colorful as well as utilitarian.

Flat rejection of a student's response is avoided; instead, he should if possible be told wherein he has erred. The responses designed to handle unanticipated student entries seek to be equally tactful, for all too often some of these unanticipated responses prove to be quite reasonable! Sometimes it is well to admit something which both students and author already know: "I'm sorry, I just can't seem to understand you. Please re-word your response."

Variety is another important element in every segment of the course. Variations keep the student alert, as long as they are not too obtrusive. Each new frame appears as something new, calling for a new and different type of student input. Factors of textual length, complexity, format, and difficulty of response afford great leeway in this respect.

2. Latitude of Student Responses. Ten different types of student response are entertained in course 201x, in the following proportions:

Multiple choice	27%
Supply word(s), for blank	22%
Yes/no	14%
Formulate subject heading	13%
True/false	6%
Supply word(s), other	5%
Supply statement	5%
Furnish example	4%
Furnish opinion	3%
Matching	1%

The preponderance of multiple-choice and fill-the-blank questions apparent from the list was unplanned and probably reflects the greater ease with which assessment of responses to these types can be programmed.

The question-answer form of MTM differs markedly from that of a quiz or a formal interview because student input does not connote finality. Not every question can be answered by reference to preceding statements, and at times the student is asked questions which can be answered only after a certain amount of discussion or which require him to draw on personal resources quite outside the programmed instruction. The counter-reply then leads him to a correct formulation. The object, of course, is to make the student think rather than to make him strain for clues. The idea that one can discuss certain points without prejudice should in the long run prove very attractive to graduate students. In a subsequent version of the course it may be possible to provide branching routines tailored to the background and individual aims of students better than is now the case. This type of refinement is envisioned as taking place over a long period of time and will entail detailed analysis of operating records in conjunction with student profiles.

Some CAI experts have advocated permitting students to indicate that they know correct answers simply by pressing a key which tells the computer, in effect, to proceed to the next frame. That is, since CAI supposedly allows students to proceed at their own pace, they should not be required to input answers they know to be correct. Although this would be suitable for some types of instruction, the need for discussion and reinforcement did not recommend its use in course 201x.

Another interesting proposal is that the speed with which computer responses to student input are delivered should be delayed to resemble human response time, in order to provide a more comfortable rhythm in the interactive process. We believe that such a device is not needed in this particular course because only the yes/no responses seem instantaneous. By the time a student has read a content-bearing programmed response, the impression of computer speed should have dwindled. The suggestion merits further investigation, however.

3. Continuity. Giving the student the means to review material previously covered provides a sense of continuity to the course. The subject cataloging course contains two types of review: voluntary review which is invoked by the student and involuntary review which is automatically invoked by the computer. Involuntary review is used with great caution for if given overtly the student may feel that he is being treated unfairly and if given covertly he will probably recognize the review material and think the computer has gone awry.

Voluntary review, on the other hand, gives the student a heightened sense of participation in his own instruction. He may think, "What if I had answered 'x' instead of 'y' on that tough question awhile back? I'd like to try it again." An added advantage is gained through displaying a list of available review topics: such a list reassures the student that he has not, in fact, missed any of the main points of the discussion (if this indeed is the case) and helps him formulate his own concept of how the material is logically organized. This last, of course, is very important to the memory process as we understand it.

Unfortunately, invitations to review break the thread of discourse. We judged it best to limit them in number and to plant them at points where interruption would be least damaging.

Of greater concern are the breaks caused by human fatigue and operational time limits. It is impossible to restore the student's stream of attention simply by flipping a switch when he logs back in at the terminal. For this reason it is highly desirable that the system's "restart facility" should resume not at the point of previous sign-off but at an earlier point which will replay the last three or four frames of the preceding session. The task of programming this feature is substantial, but it is well worthwhile for the sake of continuity. Normally the student will not have kept notes, and the availability of hard-copy printout of previous sessions is problematical.

An equally important question concerns sign-off arrangements. We are not sure of the optimum duration of a console session, for even a single student. With some students, we expect that learning efficiency (the instructional transfer rate) will decline markedly after 25-30 minutes at the console. Others might be able to continue at a high level twice that long, depending not only on individual receptivity and stamina but also on where they happen to be in the course itself, parts of which are more difficult than others.

To set the limit of terminal sessions according to some administratively convenient standard would be to relinquish part of the tutorial advantage. The student is a good judge of his own condition, and we feel that he should be allowed to sign off whenever he is ready for closure. But what of the student who is either too proud or too awed by the system to sign off when he really should? In such a case the program itself should make the decision. It can do this by first establishing a norm based on performance during time 0^m-20^m, then comparing current performance in ten-minute blocks against this norm. When the transfer rate declines to about 75% of the norm, a flag is set which causes the program to switch to automatic signoff at the next designated closure point. Certainly automatic signoff should be accompanied by a short statement explaining the action. (Note: System features permitting the above arrangements have not yet been implemented in PILOT.)

4. Student Aids. Students are not encouraged to resort to the LC List or any other source to help them formulate answers. We try to tell them enough so that they can draw logical conclusions and enter into intelligent discussions without this sort of side activity, which can become fairly expensive in terms of terminal time.

6.4.3 Scoring, Measurement, and Control

Introduction. The ability to evaluate individual student performance currently and cumulatively is often advanced as giving CAI an overwhelming advantage over traditional means of instruction.

evaluation or scoring can be used as the determinant of a student's path through a course. The student's achievement can be expressed in many ways: the number of times the program in interaction with the student has achieved desired match, undesired match, or no match at all, separately organized according to the following four categories:

1. Sequential Blocks - the student's experience with a single frame, a series of frames, a terminal session, two or more sessions, the entire course.

2. Topical Blocks - the student's experience with a particular concept, a facet of that concept, or a combination of related concepts.

3. Tactical Blocks - the student's experience with certain types of presentation or forms of question.

4. Condition Indicators - reflections of the student's apparent efficiency, perceptiveness, or attitude.

None of these is exclusive of the others as there is nothing to prevent a measure from being used again and again for different analytical purposes.

Fig. 7 gives sample tallies which illustrate the machine potentials of scoring and measurement. The "subjects" refer to topical areas within subject cataloging and to areas of student performance, e.g., specificity, direct access, use of natural language, use of key terms, synonyms, etc. The first type of tally shows the student's level of performance throughout the course and also shows his overall performance in a given subject area. The second type of tally shows student performance in various (but not all possible) types of response.

Another possible tally is to determine the variance in student performance at different times during the terminal session. Or, one can tally the areas in which the student asked for a review and determine whether he or she did better in that subject after the review. If the tally shows that the student did not ask for any reviews, this may indicate that he is overconfident.

Fig. 7: SAMPLE MACHINE MEASUREMENTS

TALLY 1

		¼ Course	½ Course	¾ Course	Overall Performance in Subject Area	
Subj A	-3	-2	-1	0	-6	
Subj B	-2	-3	-4	-6	-15	(Student's weakest subject)
Subj C	-1	-1	-1	-1	-4	
Subj D	0	0	0	0	0	(Student's best subject)
Subj E	-1	0	-1	0	-2	
Constancy of student performance	-7	-6	-7	-7		

TALLY 2

	Subj A	Subj B	Subj C	Subj D	Overall Performance in Type of Response	
Multiple Choice	-1	-2	-4	-1	-8	
Matching	-1	-1	-2	-1	-5	
Yes/No	-3	-4	-2	-1	-10	(Student's weakest response)
True/False	-2	-3	-2	-2	-9	
Furnish example	0	0	0	0	0	(Student's best response)
Overall Performance in Subject Area	-7	-10	-10	-5		

Value of Analysis. Initially, the analysis of student records tells us much more about the course itself than it does about the students. Until editing and re-editing on the basis of operational experience has been accomplished and the course has thereby attained stability, a student's record can reflect his personal achievement only in very general terms. It can tell us if he managed to get through the course at all, the ideas he seemed to have greatest difficulty with, and (by means of the above-mentioned condition indicators) whether his input has been willfully contrary, mischievous, or downright stupid - in which case we would consider it invalid for course evaluative purposes.

One should think of scoring as part of a larger system of measurement and control. In most CAI languages, such a system can make real time determinations governing program execution, i.e., the unfolding of pre-planned statements in a conditional sequence which depends on the matching of student input with a specified string of characters, or on the values stored in a file of variables, or on the two in combination. The file of variables is very useful. Singly or in combination they can be used to override explicit branching, but this is only part of their utility. They can be referred to at any time and their contents altered, transferred, combined, or displayed according to the wishes of the program designer - for the purpose of furnishing to the human teacher (author, proctor, teaching assistant) a concise record of the student's achievement.

When the system permits automatic recording of elapsed time, the record is amplified and individual and mean rates of progress can be established. Conceivably through comparison with a current rate, these rates could be used dynamically to influence program execution. However, we expect their greatest value will lie in their contribution to the analysis of student records.

201x Measurements and Controls. The measurements and controls of course 201x go somewhat beyond the bounds of necessity because of its innovative nature and the fact that it has been implemented in a language in which investigation of new types of control had research value. The measurement and control structure of the course includes provision for the following:

a. Typical sign-in and registration procedure.

b. An optional console-familiarization routine.

c. Numerical variables governing mechanistic intra-frame operations (these are either cleared or initialized after each frame).

d. Numerical variables for storing student's scores as they accumulate.

There are 35 categories of scoring topics, as listed in Fig. 8 (which also shows their distribution throughout the course). (All scores are negative or "bad-tallies," these being much easier to plan and manipulate than pre-set "good-tallies" subject to reduction.) In order to prevent excessive downgrading in a frame containing numerous undesired-match or no match possibilities, an arbitrary maximum score, based on content, was set for each frame. This requires temporary storage of bad-tallies until the student has completed the frame.

e. Two voluntary review points, listing a total of 15 topics from which the student recursively selects those he wishes to review. When review is requested, all existing bad-tallies in the review block are transferred to new addresses.

f. Three involuntary reviews (overt) invoked when certain negative scores reach a set value.

g. Scattered commendatory statements based on performance in a series of frames (e.g., "I see that you made only two mistakes in the preceding drill section. Good for you!" In this way the variable does double duty - it triggers the statement which in turn quotes it.)

h. When requested by the teaching assistant or proctor, the 201x program outputs the student record in detail:

1) The last numbered frame completed.

2) All negative scores, listed according to the various scoring categories, for which see Fig. 8.

3) A list of review topics invoked by the student.

4) The negative scores attained on review.

5) (If a certain condition indicator is present): a statement to the effect that the student seems to be somewhat overconfident.

6) (If a certain variable exceeds a set value): a statement to the effect that the student seems to be sabotaging the instruction.

7) A summary of negative scores combined in five groups (roughly coinciding with the five main "movements" comprising the instruction).

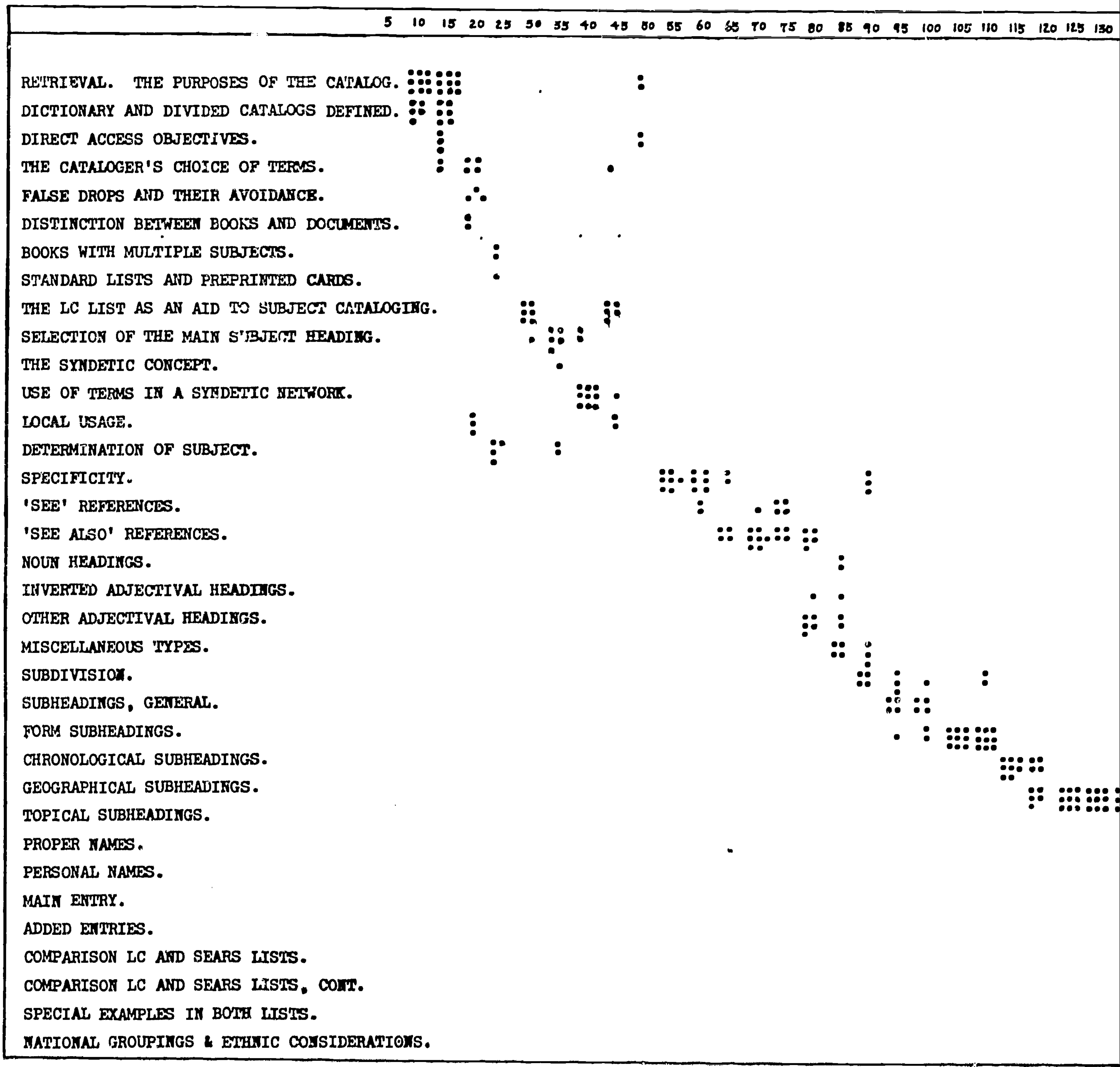
8) A net positive score is computed by subtracting the sum of the preceding item 7 from 1000.

Although the student may assume that his progress is being recorded, there is no point in advertising to him the range and intricacy of the process. This would be almost as unnerving for the student as a session with a human tutor in which the latter took copious notes, used a tape recorder, and kept clattering away at an adding machine during the interchange.

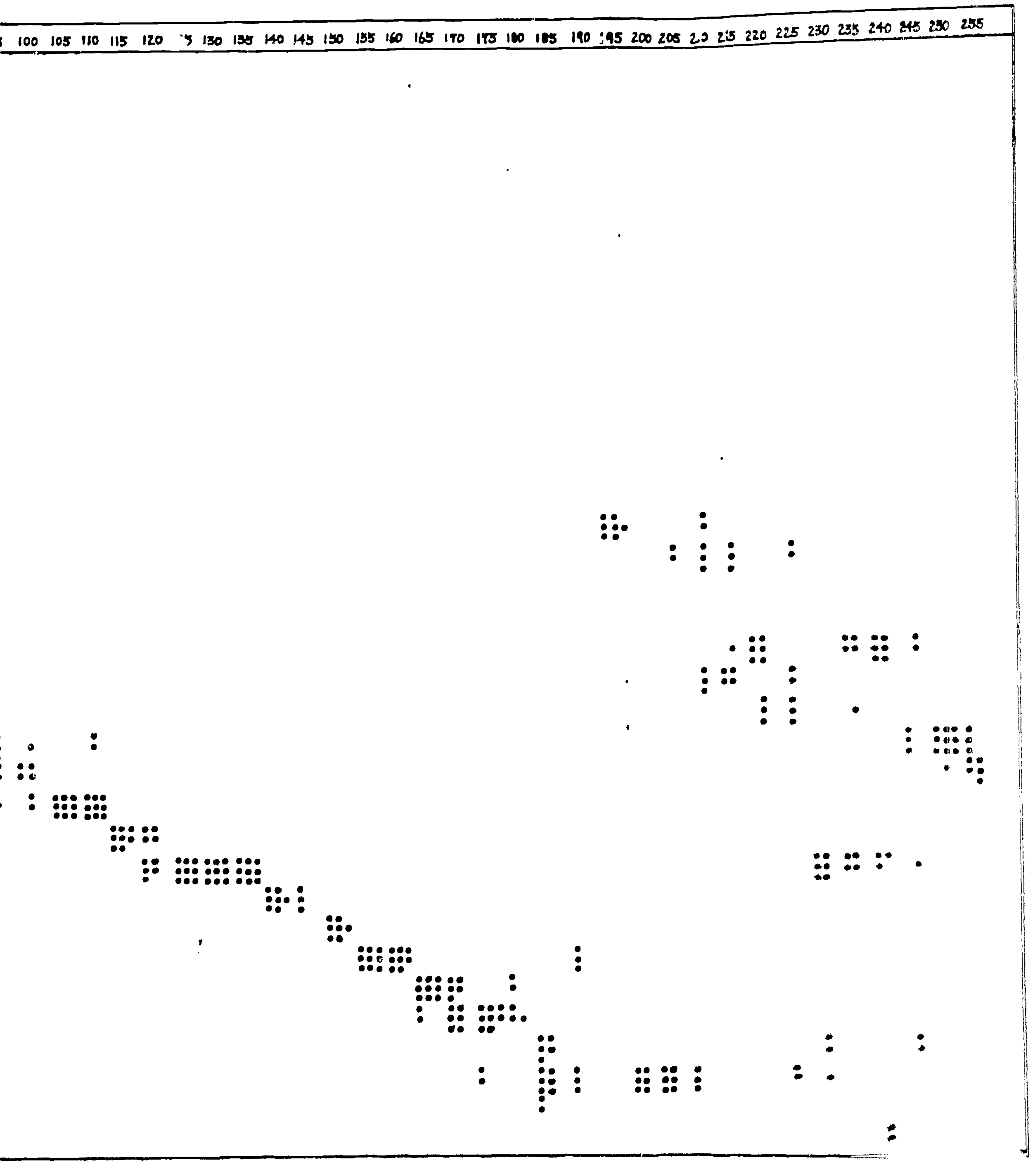
Problems in Development. All this machinery takes a great deal of time to put together and get into working order. It exacts a price in computer core memory. It represents an intellectual investment that must be patched and shored up each time a frame is revised, and accordingly tends to ossify the course. In other words, the system of measurement and control can become so intricate that it overshadows the subject matter of the course itself. (The situation is alleviated somewhat, fortunately, whenever a control or measurement function is taken over by the CAI language system in use.)

Our experience with 201x indicates that the course author should not allow himself to become so intrigued with the various control and measurement devices that the vitality of the instruction suffers. (This kind of situation can be forestalled in the design phase by specifying only those measurement functions which will contribute either to the instructional plan or to necessary evaluations.)

FIG. 8: DISTRIBUTION OF SCORIN



OF SCORING TALLIES



6.5 MTM Laboratory Supplement in Subject Cataloging (201xL)

6.5.1 Introduction

A laboratory conducted in conjunction with a course in cataloging is typically structured around an increasingly complex collection of books to which students are required to assign subject headings. The tasks students are required to perform correspond roughly to current lecture material. If the MTM course in subject cataloging is adopted in the Berkeley M.L.S. program, this would require extensive realignment of the laboratory supplement which accompanies the present subject cataloging course. We felt that since the laboratory part of the course would have to be revised anyway, it would be appropriate to do so in the same mode as the main course, i.e., the MTM mode.

The most striking advantage to be gained from an MTM laboratory supplement is the same as that for an MTM basic course, i.e., the computer's ability to respond immediately to input from a student terminal. In a conventional laboratory environment, the student obtains no final evaluation of his work until after an assignment has been handed in, manually annotated and graded by a laboratory assistant, and returned. The delays inherent in such a procedure tend to blur the student's original reasons for making certain choices.

The computer's ability to respond immediately to student input promises a better situation wherein students can receive immediate feedback on their submissions, can quickly revise these, and resubmit them until they achieve the desired result. They can carry on a discussion of various choices without the feeling of finality which builds up around a written laboratory assignment.

6.5.2 Structure and Scope

As designed, the laboratory supplement will consist of seven units aimed at giving the student experience in cataloging books of gradually increasing variety and complexity. The books chosen for processing were grouped as follows:

Lab Unit 1: Books lending themselves to simple, straightforward subject headings, illustrating: a) the virtues of simplicity and directness; b) the use of natural language; c) the decision process in choice of inclusive headings vs. multiple headings; and d) the use of "see" and "see also" references and their "x" and "xx" counterparts.

Lab Unit 2: Books illustrating: a) use of adjectival headings; b) use of phrase headings; c) use of parenthetical qualifiers.

Lab Unit 3-5: Books illustrating: a) use of form subdivision; b) use of geographical subdivision; c) use of chronological subdivision; d) use of topical subdivision.

Lab Unit 6: Books illustrating: a) use of proper name headings; and b) headings appropriate for belles lettres.

Lab Unit 7: Books illustrating technical differences in use of the Sears List as compared with the LC List.

Physical materials will consist of some 92 books selected and grouped to suit the above categories. This collection can be added to and modified at any time. From among the 12-18 books in a group, the student will choose 5 and will assign one or more subject headings for each, submitting each in turn to MTM for evaluation and comment. (Sometimes the comment will amount to something like "You have overlooked the fact that.....[statement]......Please sign off, reconsider, and resubmit.")

The 201x Lab supplement is expected to bulk considerably larger than the basic course in terms of storage requirement, as indicated in the following comparison of source decks:

	Course 201x	
	Total	8000 cards

	Course 201xL	
Introduction		250 cards
Lab.1		2750
Lab.2		2500
Lab.3		2500*
Lab.4		2500*
Lab.5		2500*
Lab.6		2500*
Lab.7		2500*
	Total	18,000*

(* = projected)

The specifications for the subject cataloging lab supplement were formulated in mid-1968, after the profile of the basic course 201x had been well established and after PILOT was selected as the language of implementation.

6.5.3 Mode of Presentation

The mode of presentation is quite different from that which has been provided in the basic course. Each book is dealt with as a separate entity, with no logical progression joining it to the others of its group. Everything pertinent which can be said about the book and the problem of assigning to it a subject heading must be available during execution of the routines designed around it. This represents a task of some magnitude even though common errors of construction, punctuation, hyphenation, etc., can be handled through standard callable subroutines valid for all cases.

It might be possible to avoid some of this reiteration by prescribing one set of five books, in given order, for each laboratory unit. The reason for not doing so was that we felt students would work better independently of each other, free of pre-ordained sequence. The arrangement has the added advantages of reducing mutual interference between students and in giving the student a chance to go on and process more than the required number of books, if he wishes, either for added practice or in hope of clearing up some bothersome question.

6.5.4 Conserving Terminal Time

The use of reference books (the LC and Sears lists) in connection with MTM terminal operation raises the question of how to avoid wasting terminal time on-line while the student is engaged in considering the books themselves and deciding how he should treat them. The solution is to restrict on-line operation to periods of active interchange.

We have no data on which to base a reliable ratio, and the actual running time of lab units on-line will fluctuate radically with student achievement rates. We expect that it might be on the order of four minutes of study and decision to one minute of terminal time. For planning purposes, we consider that the typical student will spend an average of 12 minutes on-line per book, plus 36 minutes off-line, or 48 minutes per book and 4 hours per unit, on and off-line. (These estimates are scaled to selectric terminal speeds.) This means that three students should rotate on a single terminal, with an allowance of 20% for lost motion. Extremely rapid sign-on and break routines and rapid file call-up present no great technical problems. It is expected that during a lab session a student should be able to restart by means of a single input.

6.5.5 Linkage with Course 201x

No cross-monitoring between 201x and 201xL is provided. It is clearly in the student's interest not to undertake a lab unit before he has been adequately prepared for it. On the other hand, it would be a mistake to withhold instruction simply because a lab unit had not yet been accomplished. The degree of separation between the basic course and the laboratory supplement avoids imposing, on the operational mode of either, constraints which might be dictated by the other. It also permits them to be separately implemented, an action which might be occasioned by phasing-in-problems, equipment limitations, etc.

However, there is coordination between the laboratory supplement and the basic course, and it is arranged in the following way: within 201x, the student will be "cleared" for Lab. 1 after he has passed through Frame # N, for Lab. 2 after Frame # NN, for Lab. 3 after Frame # NNN, and so on. Instead of interrupting instruction to notify the student of clearance, the information is withheld until signoff. (Not yet implemented in PILOT). The student then or later signs on with a call for the lab unit for which he has been cleared or for any of the preceding labs. Or, he may wait until he has completed the basic course before starting any of the lab units.

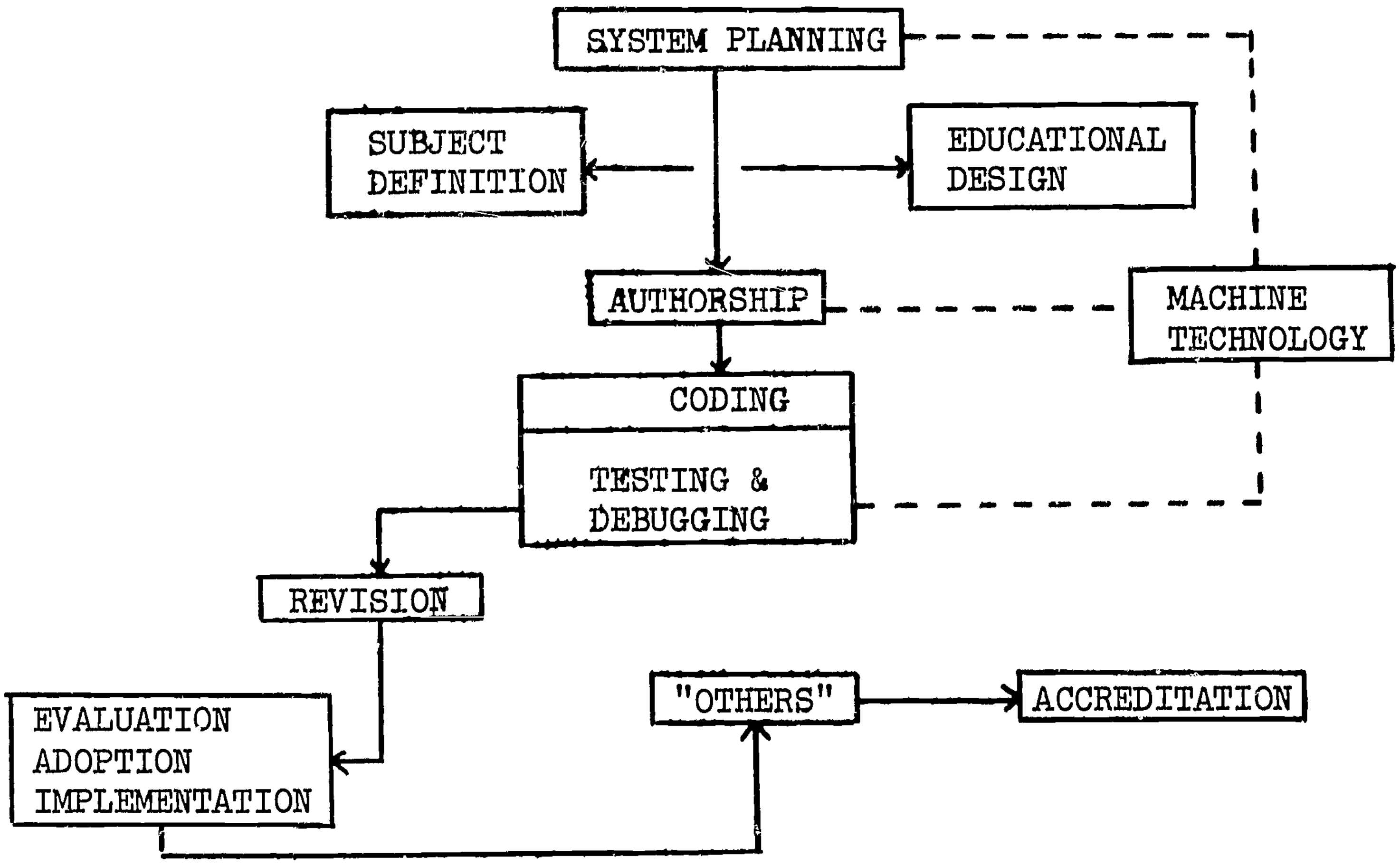
6.6 MTM Methodology

6.6.1 Introduction

The MTM course in subject cataloging and its laboratory supplement constitute readily identifiable products of research and development on the Information Processing Laboratory. A less tangible result, but one which could prove equally useful in another way, is the methodology which has emerged as a direct result of the work on the course itself.

The term "methodology" may be somewhat ambitious in this instance because it has been tested on such a limited scale and is, in fact, still evolutionary. However, we feel that the rules and procedures developed to date are valid guidelines for creation of computer-assisted or computer-augmented courses of instruction involving: a) interactive student-computer communication; b) conceptual material and pedagogical strategies too complex for the program to be generated automatically or even semi-automatically; c) a high degree of academic responsibility at the graduate level; and d) use of a high-level CAI language. The tasks and subtasks which must be performed to produce an acceptable MTM unit are shown in the diagram below, and each task is discussed separately thereafter.

FIG. 9: TASKS AND SUBTASKS OF AN MTM UNIT



6.6.2 System Planning

This presupposes awareness of an academic goal and of the problems which beset its attainment. System planning entails recognizing a range of existing or potential solution-contributory elements and from these selecting a promising set which is then fused into a solution for coping with the problem. In the case of an MTM course, effective system planning requires a multidisciplinary approach establishing 1) what is to be taught; 2) the depth, intensity, and duration of the teaching; and 3) the means to be used. Either the education specialist or the computer specialist may take the initiative in this, but a joint effort will be called for.

6.6.3 Subject Definition

This consists of establishing the boundaries of the material to be transferred, the identification of each item within those boundaries, and a clear distinction between items to be mastered and items to which the student need only be exposed. The subject specialist must present the material in a readily understandable form to the educational designer and the writer. The subject specialist cannot assume and should not even try to assume all of the other subtasks. The subject specialist is usually a member of the faculty, whose degree of authority depends in part on single-minded devotion to his specialty and for whom excursions into unrelated activities are sometimes wasteful. The subject specialist's role is integral with his status as a member of the faculty, and he may be thought of as the executive of a curriculum development committee of the faculty. (An exception could be argued for the subject specialist whose specialty is tutorial psychology and who proposes to teach that subject by machine, but such an exception only confirms the general rule.)

There is much communication between the subject specialist and the education specialist, of course, and between the subject specialist and the planning body (if there is one). But regardless of how well defined or how diffuse this interface may be, the subject specialist should be held responsible only for producing and monitoring the substantive material around which a course of instruction is designed.

6.6.4 Educational Design

The educational designer, who is a professional educator, considers the nature of the subject material in the light of what is known or surmised of educational psychology, determines what portions of it may be amenable to MTM, outlines a sequential strategy for its presentation, and establishes ground rules governing format and interchange. It is this last item which is the most difficult to prescribe in advance of the actual writing because there is little procedural guidance for working in tutorial mode either with the help of machines or without it. One suspects that the old-fashioned tutor was less of a conversational textbook than an ambulatory syllabus (Mark Hopkins and his log notwithstanding).

A technique of saying things and then posing questions about them is standard with educators. The result is hardly "conversational" but it is instructive. The educator must decide how far to venture into more discursive interchange and where to insert lacunae which he hopes the student will be able to fill in either spontaneously or on cue. The educational design function also includes the making of decisions on how extensive and elaborate a control and measurement structure to incorporate in the MTM course.

These and a host of technical questions bearing on the proposed course must at least be considered before the actual writing process is initiated, even though in many cases only tentative answers can be assigned.

6.6.5 Authorship

The technical educational writer operates within the substantive domain established in the subject definition phase, according to guidelines formulated in the educational design phase. His or her task consists of transmuting corpus into a stream of controlled statements and questions against which is matched an uncontrollable stream of student reply. The anticipation of variance in the latter is the most brain-racking task of the author. He hopes to make the dialog meaningful, and to do so he must deal with the meaning of as many different student replies as possible. The rest he can only acknowledge and try to pull together with carry-all responses.

We advise the author to rough out a basic textual line, using the perfect student as his "straight man," in order to discern the pattern of frames and families of frames into which the material seems to fall. He can then go back and start covering contingencies. During the course of this operation he can use his own informal notation to indicate text, question, anticipated answer and corresponding response, and the stepped carry-all responses for use against unanticipated answers. It is not necessary that this notation coincide with existing CAI language operation codes or even that each item be in workable sequence according to a particular language. Authorship is complicated enough without injecting mechanistic rules into the creative process.

We further advise the author to adopt a consistent style of address suitable for dialog with an intellectual equal, avoiding sterile phrases, stuffiness, and dogmatic assertions of authority, yet retaining sufficient didactic control to convince the student that he knows what he is talking about. He should strive to endow the course with the color of a single personality, regardless of how many people have participated in its formulation. In this sense the authorship function includes responsibility for a factor which can make or break the final result: the general tone of the course.

6.6.6 Coding

Coding should be carried on separately from the process of authorship, even if one is dealing with a so-called "user-oriented" language. Such languages do indeed permit people with little or no knowledge of programming to formulate acceptable machine input. The general characteristics of the language may indeed be readily explainable, the opcodes mnemonically apt, the reserved symbols few, and the fields free. But until one has used the language for an extended period of time, any coding requirement seems to get in the way of authorship, and conversely the logic and accuracy required to do a good job of coding suffer in the presence of the creative muse. Coding may be done by research assistants or by coders. (In situations involving uncomplicated exchanges and standard routines, the foregoing strictures can be relaxed for experienced author/coders, but even in such cases we have not noticed any particular advantage in doing so.)

A form similar to that furnished in Fig. 10 is recommended for both authorship and coding. If the author's text is liberally spaced, the necessary coding can often be accomplished directly thereon. Various command and control statements are entered in the right hand column opposite the text strings to which they apply. The material is then ready to be punched, the key-punch operator proceeding straight across the page, line by line. This form has proven most satisfactory among several we have tried, and it has been used exclusively for the past eight months.

For some writing and encoding operations (such as that of the laboratory routines in 201xL) flowcharting is very helpful. Flowcharts are also an excellent means of guarding against oversight, even though they require extra time to prepare and check.

FIG 10: CAI WORKSHEET

Course_____ Draft_____ Page_____

Labels	Opc	T, A, G, B, F Statements	C	Control Statements

Size: 13"long, 8"wide

Initial _____ ILR:6/68

6.6.7 Testing and Debugging

This requires a great deal of time and effort, in spite of the careful coding and keypunching. This is especially true if the system does not provide on-line edit capability, because a single flaw can knock out one or more frames which must then await recompilation before they can be further tested. The recommended procedure under these circumstances is to go through the entire course marking up the formatted listing (if a formatted listing is available) from the daily terminal printout, with a separate list of line numbers affected. A keypuncher then follows along, revising the cards as necessary. Finally the whole deck is recompiled and the process begins over again. Everyone on the project should participate in the testing and debugging.

Two different kinds of debugging should be recognized: 1) mechanistic debugging, which is concerned with discrepancies which contaminate or spoil the running of the program, and 2) pedagogical debugging, which is concerned with conversational aberrations, non-sequiturs, etc., which in the main represent oversights of authorship and educational design. Simple pedagogical debugging can be carried on concurrently with mechanistic debugging, and in practice the two are not distinguished.

6.6.8 Revision

This is usually referred back to the author, educator or subject specialist. It is occasioned by patently justifiable criticism of the content or of the form, sequence, or tone of presentation. The participation of unbiased volunteers is valuable in detecting the necessity for this, but their comments must be carefully weighed in the context of their expertise and the effects of their usual random entry into the course.

The decision to revise even a single frame cannot be taken lightly because almost invariably such revision will have repercussions in other frames. The advantage of a slight improvement in one statement may be lost if it results in loss of tactical effectiveness of associated statements. Review sequences may be affected as well, and numerous adjustments in the measurement and control apparatus may be required. This is unfortunate, but it seems to be a characteristic of CAI that it is so highly integrated and it represents such a large investment by the time a course is compiled and tested that revision becomes very costly and time-consuming. We feel that correct scheduling of sub-tasks can help reduce the need for deep revision. Also the danger of embarking on excessive revision because of isolated criticisms can be reduced by maintaining a log of test runs in which the proto-students record their comment. These log entries can then be combined and analyzed systematically.

6.6.9 Evaluation, Adoption, and Implementation

We have not yet examined these functions with sufficient thoroughness to warrant making any recommendations. These functions fall into the domain of the systems planner, the subject specialist and the educational specialist, as well as others outside the range of immediate discussion.

6.6.10 Machine Technology

This is not precisely a function (as are systems planning, educational design, etc.), but it is rather a contributory element affecting all of the functions or subtasks involved in creating an MTM course. Machine technology tends to prescribe the shape of the course and the operational conditions under which it will be utilized but it should be fairly obvious that the tendency should be resisted when it threatens course objectives (assuming that the original concept is sound).

Machine technology is an ever-present factor for everyone involved in formulating an MTM course, with the clear exception of the subject specialist whom we have defined elsewhere as a person who should be solely and exclusively responsible for corpus. It follows that acquaintance with machine techrology is indispensable to anyone performing functions of system planning, educational design, authorship, encoding, debugging, or revision.

6.6.11 Overlapping Functions

In the functional scheme advanced above, any phase of activity may be participated in by more than one person, and one person may take part in two or more phases. A likely instance would be for the educational specialist to chair a system planning committee, to assume responsibility for educational design, and to do some of the course writing. He might have a co-author who would also perform part or all of the encoding. A keypuncher might learn enough about coding to participate in this function, as well as in testing, debugging, and revision.

APPENDIX 1

CHRONOLOGICAL REVIEW OF MAJOR DEVELOPMENTS DURING PHASE 1.

CHRONOLOGICAL REVIEW OF MAJOR DEVELOPMENTS DURING PHASE I.

1967

June 15	Start of project.
June-Sept.	Basic planning of the Laboratory.
June-Nov.	Evaluation of central computing facilities and negotiation with Berkeley campus Computer Center.
Sept.	Selection of fields for initial development; start of MTM course in subject cataloging; IBM-2740 remote terminal delivered.
Oct.	Start of development of ILR monitor program.
Nov.	Decision made to use 360/40 as central processor; start of work on associative retrieval; interim draft of MTM course in subject cataloging completed.

1968

Jan.	Authority list formed for indexing document file; program written to convert author names to canonical form; ILR monitor went into operation serving one terminal; final draft of subject cataloging course completed.
Feb.	Information science documents indexed.
March	Creation of first word association file; LABSRC1 put into operation.
April	IBM-2314 storage device installed.
May	Machine debugging of subject cataloging course began; additional word association files created.

June Work on cataloging lab supplement course began; ILR monitor program now able to serve multiple terminals simultaneously.

July Acoustic coupler obtained; start of daily MTM debugging (remotely: Berkeley-San Francisco).

August Hypothetical system performance study completed.

Sept. LABSRC2 became operational, LABSRC3 became partially operational; acquired IBM-2741 remote terminal and a second acoustic coupler.

Oct. Machine debugging of basic cataloging course completed. (First version, of course.)

Nov. Abstracts obtained for all documents in information science master file. Decision made to purchase Sanders 720 CRT units plus related hardware. Work started on developing a suitable MTM language to be used on Berkeley campus.

APPENDIX 2

PHYSICAL ARRANGEMENT OF THE INFORMATION PROCESSING LABORATORY

PHYSICAL ARRANGEMENT OF THE INFORMATION PROCESSING LABORATORY

Location. The Information Processing Laboratory is located centrally within the School of Librarianship, on the 4th floor of the Doe Library Building of the University. It is expected that expansion through an increase in the number of student terminals will be accommodated in an adjacent room of the same dimensions.

Equipment. The equipment shown in the diagram following this page is identified as follows:

TTY - Teletype Model 35-ASR with punched tape output attachment.

IBM 2741 - IBM 2741 Selectric Terminal.

Acoustic Couplers - Anderson Jacobsen Acoustic Coupler Model ADC 260.

CRTs - Sanders 720 remote terminal stations consisting of: Sanders 708H Terminals, with 722A-3 Keyboards and 7284 modification for 84-character line. One unit is equipped with Sanders Photo Pen and Amplifier 7220-1.

Control Unit - Sanders 701, with 7215A Synchronous I/O, 1705A Memory (3), 7221-3 Peripheral Control Module., and 706-2 Basic Hard Copy Adapter.

Data Set - (for TTY) Western Electric 103F.*

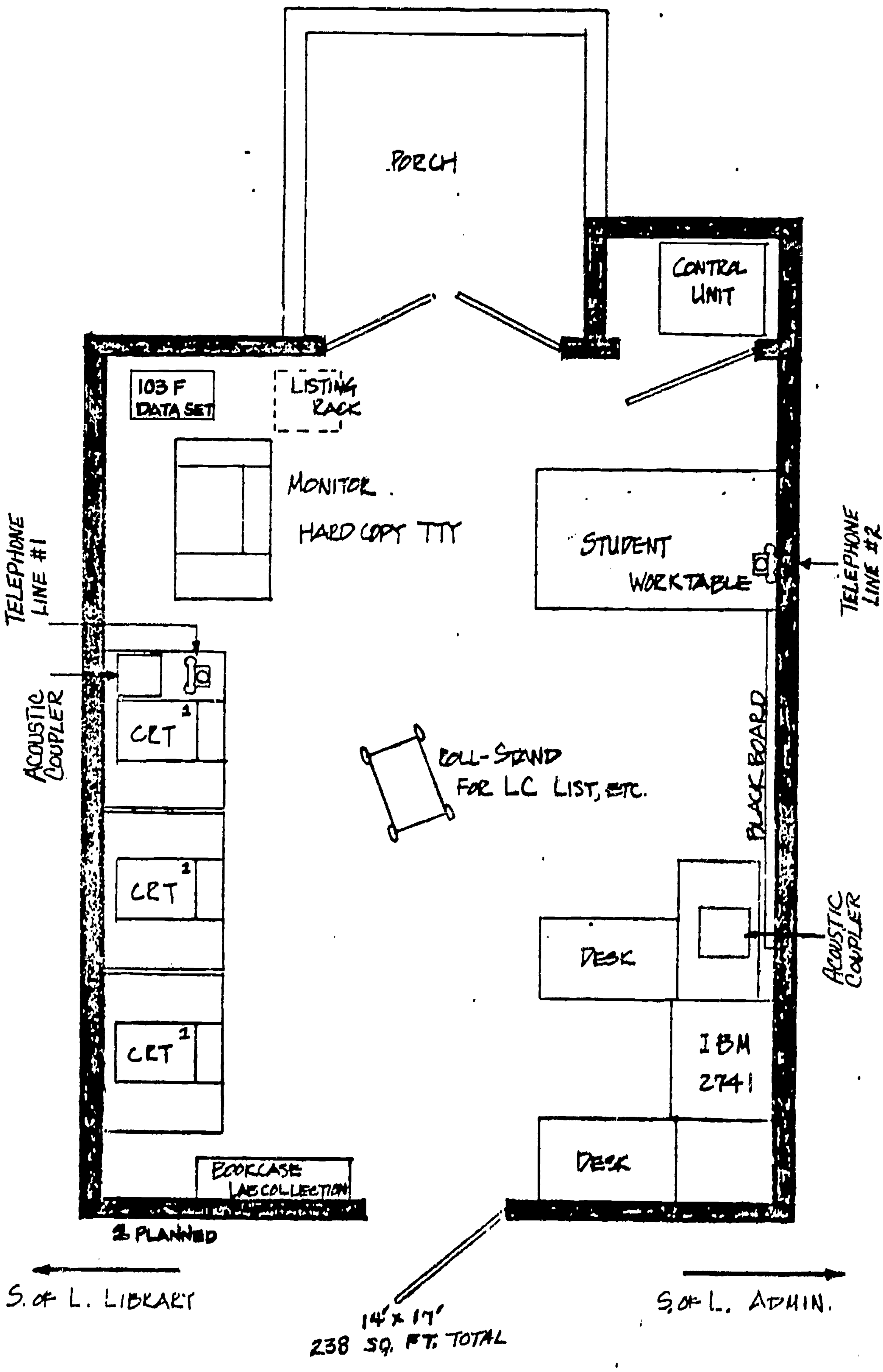
Data Set - (for CRT Control Unit)* General Electric TDM-220 D20 Modem. (functionally equivalent to Western Electric 201B1 modem)

*(Plus one each located at the other end of the transmission lines. The mechanical terminals use a commercial voice-grade Schedule 4 2-wire line. The CRT terminals use a 4-wire Schedule 4 voice-grade line. The central computer is about 1 cable-mile from the laboratory room.)

Serial Data Communications Buffer (not shown) - Sanders 731/1 (located at the Computer Center), serving the CRT system in the manner of an IBM 2701. Mechanical terminals are served by an IBM 2701 at the central computer.

Computer Link-up. Communication with the Berkeley campus Computing Center will be entirely via fixed circuit and data sets. The acoustic couplers are used for tying the teletype and the selectric terminal with off-campus computers such as the IBM 360/50, located at the University of California Medical Center, San Francisco.

ROOM 430, SCHOOL OF LIBRARIANSHIP, UCB



APPENDIX 3

HARDWARE

HARDWARE

INITIAL CONSIDERATIONS

Central Computer. At the time this project started, the Computer Center on the Berkeley campus had three distinct computing systems: 1) a directly-coupled IBM 7040-7094 system that had carried the brunt of the computing load for several years; 2) a CDC-6400 which had been installed shortly before for the purpose of taking most of the load off the 7040-7094 system; 3) the IBM-360/40 which was used primarily to support operations on the other systems. No major computer user on the campus used the 360 as a primary computer. The 7040-7094 system was scheduled to be removed (and indeed was removed) in early 1968. Therefore, we had to decide between the CDC-6400 and the IBM-360/40 as to which machine would become the central computer for the Laboratory.

There were several factors to consider in making this choice: speed, memory size, size and type of auxiliary storage devices, machine organization, availability, supporting software, cost, and suitability of supporting a network of remote terminals. After careful consideration of these factors, we chose the 360/40. There were two factors in favor of the 360/40 that influenced us strongly. The first was that the 360/40, being used by the campus Computer Center basically as a secondary machine, offered the greater promise of being available for long periods of time each day. This was of prime importance to the Laboratory as its facilities must be available to users many hours each day.

The second major factor in favor of the 360/40 was its internal organization. This machine devotes eight binary bits to the representation of each character. This means that the 360/40 may distinguish internally between 256 unique characters. The CDC-6400, on the other hand, is so organized that just six bits are devoted to each character. Thus, it is able to represent only 64 unique characters internally. This difference is important to the Information Processing Laboratory. In handling library data one must often deal with characters that are not present in the Roman alphabet. Also, it is desirable to be able to process other nonstandard characters such as diacritical marks. We felt that the 64-character limitation of the CDC-6400 would restrict the flexibility of the Laboratory unduly.

Remote Facilities. With respect to remote terminal equipment, both mechanical terminals (remote typewriters) and cathode ray tube (CRT) equipment were considered. A vital aspect of any terminal network is the communication facilities that link the network to the central computer. With mechanical terminals, one may use a dedicated communication line in conjunction with data sets, or one may transmit over normal telephone lines via a "dial-up" operation using acoustic couplers. When using CRT equipment a private leased line communication link is needed to achieve best performance by the remote terminals.

At the beginning of this project there was already on hand a Teletype (Model 35) unit. Shortly thereafter, an IBM-2740 typewriter terminal was obtained. These units were purchased by the School of Librarianship. We decided that these mechanical terminals, though comparatively slow in displaying output, would be adequate to meet the goals of Phase I. It did not seem appropriate to plan to install CRT terminals until development was well along in several areas. Since we would be using the on-campus 360/40, we chose to dedicate two local telephone lines to serving these terminals rather than acquiring acoustic couplers and using a "dial-up" procedure.

CURRENT FACILITIES

Central Facility. Much of our development work has been done using the IBM-360/40 at the Berkeley campus Computer Center. This is the machine that will support the Laboratory once it becomes operational. It has a 128-K memory, four 7-track magnetic tape drives, two card readers, one card punch, an 1100-line/minute line printer, and an operator's 1052 typewriter. Early in Phase I this 360/40 system had four 2311 disc storage units with a combined capacity of 29 million characters. As mentioned earlier this machine is run under control of the IBM Operating System. Two of the 2311 disc storage devices were devoted to the exclusive use of OS itself. In April, 1968, a large 2314 disc storage device was installed. This unit has a capacity of over 200 million characters. Until October 14, 1968, the entire storage capacity of the 2314 was dedicated to the exclusive use of the Institute of Library Research. Since that date we now share the 2314 with other campus users, ILR retaining exclusive use of half its storage capacity. The four 2311 disc units were removed in mid-October, 1968. The area of the Laboratory project where this large auxiliary storage unit will be most needed is that of processing large files. The Laboratory would be quite restricted if we were limited to the 2311 devices.

As discussed in section 6, our MTM development work has been carried on using the facilities of the U.C. Medical Center in San Francisco. The computer we use there is an IBM-360/50 with a 256-K memory and the usual complement of peripheral equipment, plus an IBM-2314 storage device.

Remote Terminals. In 1967, two mechanical terminal devices were purchased with funds provided by the U.C. School of Librarianship. One of these is an IBM-2740 typewriter, the other a

Teletype Model 35. The 2740 is linked to the 360/40 via an A.T.&T. Model 103-F data set and a dedicated voice-grade telephone line. Similar communication equipment allows us to link the Teletype Terminal to the 360/40 as well. At the 360, data sets and a 2701 data adapter unit provide the required interface.

In the early development stages of the MTM programs, project staff travelled to the U.C. Medical Center to carry on their machine work. However, in July, 1968 we obtained an Anderson-Jacobson Mcdel 260 acoustic coupler that we now use in a "dial-up" procedure to link the Teletype to the 360/50 in San Francisco. Effort on the MTM work increased to the point that in early September, 1968 we obtained an IBM-2741 terminal and a second acoustic coupler. We now routinely have two remote terminals being used in developing MTM programs communicating with San Francisco simultaneously.

APPENDIX 4

INDEX TERM LISTS

APPENDIX 4a: SUBJECT AUTHORITY LIST

INFORMATION PROCESSING LABORATORY PROJECT
JANUARY 31, 1968
REVISED DECEMBER 16, 1968

ABBREVIATIONS

S	= SEE
SA	= SEE ALSO
SN	= IN THE SENSE OF (I.E. SCOPE NOTE)
*	= NO DOCUMENTS YET INDEXED WITH THIS TERM
+	= TERM NOT ALLOWED, RELATED TERM TO BE USED

*ABBREVIATION
ABSTRACT
ABSTRACTING
ACCESS
ACCESSION NUMBER
ACCURACY
ACQUISITION
ADDRESS
ADMINISTRATION
ALGEBRA
+ALGOL
S PROG. LANGUAGE
ALGORITHM
ALPHABETIC
ALPHABETIC ORDER
ALPHANUMERIC
*ALTERNATIVES
AMBIGUITY
ANALOGY
ANALYSIS
ANSWER
*ANTHOLOGY
SA BIBLIOGRAPHY
APPLICATION
+ARITHMETIC
S MATHEMATICS
ARRAY
+ARTICLE
S DOCUMENT
ARTIFICIAL INTEL
ASSIGNED
ASSOCIATION
ASSOCIATIVE

+ATTRIBUTE
S CHARACTERISTIC
AUTHOR
AUTHORITY LIST
SA THESAURUS
AUTO ABSTRACTING
AUTO. INDEXING
AUTOMATIC
AUTOMATION
SA MECHANIZATION

BATCH PROCESSING
BIBLIOGRAPHIC
BIBLIOGRAPHY
SA ANTHOLOGY
BINARY
BOOK
BOOLEAN
SA LOGICAL

CALL NUMBER
CANONICAL
SA NORMALIZED
CARD
CARD CATALOG
CATALOG
CATALOGING
CATEGORIES
CENTERS
CENTRALIZED
CHARACTERISTIC

CHEMICAL
CIRCULATION
CITATION
CITATION INDEX
*CLAIM
SA COPYRIGHT
SA PATENT
CLASSIF. SCHEME
CLASSIFICATION
CLERICAL
+CLUE WORD
S KEYWORD
CLUMP
CLUSTER
CO-OCCURRENCE
+COBOL
S PROG. LANGUAGE
CODE
SN MEDIA DESIGNATION
CODING
SN COMPUTER CODING
COEFFECIENT
COLLECTION
*COLLOQUIUM
SA CONFERENCE
SA MEETING
SA SYMPOSIUM
COMBINATIONS
+COMIT
S PROG. LANGUAGE
COMMUNICATION
COMP LINGUISTICS
COMPARISON
COMPUTER
CONCEPT
CONCORCANCE
CONDITIONAL PROB
CONFERENCE
SA COLLOQUIUM
SA MEETING
SA SYMPOSIUM
CONNECTION
+CONSECUTIVE
S ORDER
+CONSOLE
S REMOTE TERMINAL
CONTENT
CONTENT ANALYSIS
CONTEXT
CONTROL
CONTROLLED
CONVENTIONAL
CONVERSION
COORDINATE
COORDINATE INDEX
SA UNITERM SYSTEM

*COPYRIGHT
SA CLAIM
SA PATENT
+CORE
S STORAGE
CORRELATION
COST
COUNT
COUPLING
CRANFIELD
CRITERIA
CRITICAL
SN REVIEWING, NOT VITAL
CROSS REFERENCE
CURRENT AWARENES
CURRICULUM
+CUSTOMER
S USER

DATA
*DECENTRALIZATION
DECISION THEORY
DEDUCTIVE
DEGREE
DEPTH OF INDEXIN
DESCRIPTIVE
DESCRIPTOR
SA KEYWORD
SA TAG
SA TERM
DESIGN
SA PLANNING
DICTIONARY
+DIFFERENCE
S COMPARISON
+DIGITAL COMPUTER
S COMPUTER
DISCRIMINANT
+DISPLAY
S REMOTE TERMINAL
DISSEMINATION
*DISSERTATION
DOCUMENT
SA JOURNAL
DOCUMENTATION
DUAL DICTIONARY

+ECONOMICS
S COST
EDITING
EDUCATION
EFFECTIVENESS
SA EFFICIENCY

```
 EFFICIENCY
      SA  EFFECTIVENESS
+ELECTRONIC COMPUTER
      S   COMPUTER
+EMPIRICAL
      S   EXPERIMENT
+ENCODING
      S   CODING
 ENTROPY
 ENTRY
      SN  ACCESS POINT
 ERROR
 EVALUATION
      SA  TEST
      SA  UTILITY
      SA  VALUE
 EXPERIMENT
 EXTRACT


 FACET
 FACETED CLASSIF.
 FACT RETRIEVAL
+FACTOR ANALYSIS
      S   STAT. METHOD
 FALSE DROP
 FEEDBACK
 FILE
      SA  LIST
      SA  STRING
 FILE ORGANIZATIO
 FLOW OF INFO.
 FORMAT
+FORTRAN
      S   PROG. LANGUAGE
 FREQUENCY
 FUNCTION
      SN  OPERATIONAL, NOT
          MATHMATICAL


 GENERAL
 GENERATION
      SN  PRODUCTION
 GENERIC
+GOAL
      S   OBJECTIVE
 GOVERNMENT
 GRAMMAR
 GRAPH
      SN  MATHEMATICAL GRAPH
      SA  TABLE
 GRAPHICS
      SN  GRAPHIC MATERIALS E.G.
          PHOTOS.
```

```
+GROUP
      S   CLUMP

 HARDWARE
      SN  COMPUTERS, MICROFILM
          EQUIPMENT, ETC.
      SA  MECHANICAL
+HEADINGS
      S   SUBJECT HEADING
 HIERARCHY
 HISTORICAL
+HUMAN
      S   MANUAL
+HUMAN INDEXING
      S   MANUAL INDEXING

*IDENTICAL
 IDENTIFICATION
 ILLUSTRATION
*IMPLEMENTATION
 INDEPENDENT
 INDEX
 INDEXING
 INFERENCE
 INFO. RETRIEVAL
 INFO. SCIENCE
 INFORMATION
 INPUT
+INQUIRER
      S   USER
+INQUIRY
      S   QUESTION
+INSTRUCTION
      S   EDUCATION
 INTELLECTUAL
 INTERDISCIPLINAR
 INTERFACE
 INTERPRET
+INTERROGATE
      S   QUESTION
+INTERSECTION
      S   VENN DIAGRAM
 INTRODUCTORY
 INTUITIVE
 INVENTORY
*INVERTED
 IRRELEVANT
+ITEM
      S   DOCUMENT
 ITERATIVE

      SA  RECURSIVE
 JOURNAL
      SA   DOCUMENT


 KEYPUNCH
 KEYWORD
      SA   DESCRIPTCR
      SA   TAG
      SA   TERM
 KWIC


 LANGUAGE
 LARGE
 LATTICE
 LAW
+LEVEL
      S    DEGREE
+LEXICAL
      S    ALPHABETIC
+LEXICON
      S    DICTIONARY
 LIBRARIAN
 LIBRARY
 LINGUISTIC
 LINK
 LIST
      SA   FILE
      SA   STRING
 LITERATURE
 LOGIC
 LOGICAL
      SA   BOOLEAN


+MACHINE
      S    HARDWARE
 MACHINE-READABLE
+MAGNETIC TAPE
      S    STORAGE
 MAN-MACHINE
 MANUAL
 MANUAL INDEXING
 MATCH
 MATHEMATICAL
 MATHEMATICS
      SA   PROBABILITY
 MATRIX
 MEANING
 MEASURE
 MECHANICAL
      SA   HARDWARE
 MECHANIZATION
      SA   AUTOMATION
 MEDIUM
```

```
 MEETING
      SA   COLLOQUIUM
      SA   CONFERENCE
      SA   SYMPCSIUM
+MEMORY
      S    STORAGE
 METHODOLOGY
+METRIC
      S    MEASURE
 MICROFICHE
 MICROFILM
 MODEL
      SA   SIMULATION
 MODIFICATION
 MULTIPLE


 NATIONAL
 NATURAL
 NATURAL LANGUAGE
 NEEDS
 NETWORK
      SN   ORGANIZATIONAL STRUCTURE
      SA   ORGANIZATION
 NOISE
+NOMENCLATURE
      S    NOTATION
 NON-CONVENTIONAL
 NON-DISCRIMINANT
 NON-FILE
 NON-RANDOM
 NON-RELEVENT
*NORMALIZED
      SA   CANCNICAL
 NOTATION
      SA   TERMINOLOGY
 NUMBER
 NUMERIC


 OBJECTIVE
      SN   GOAL, NOT AS OPPOSED
           TO SUBJECTIVE
*OCCURRENCE
 OFF-LINE
 ON-LINE
 OPERATION
 OPTIMIZATION
 ORDER
 ORGANIZATION
      SA   NETWCRK
 OUTPUT


+PAIR
      S    WORD ASSOCIATION
```

```
+PAPER
      S    DOCUMENT
 PARAMETER
      SA   VARIABLE
 PARSE
 PATENT
      SA   CLAIM
      SA   COPYRIGHT
 PATTERN
 PERFORMANCE
+PERIODICAL
      S    JOURNAL
 PERMUTED
 PERTINENT
      SA   RELEVANT
 PHILOSOPHY
      SA   POLICY
+PHOTO
      S    GRAPHICS
 PLANNING
      SA   DESIGN
+PLOT
      S    GRAPH
+POLICY
      SA   PHILOSOPHY
+POPULATION
      S    COLLECTION
 PRECISION
 PREDICTION
*PRINCIPLE
+PRINT-OUT
      S    OUTPUT
 PRINTING
+PRIVACY
      S    SECRECY
 PROBABILITY
      SA   MATHEMATICS
 PROCEDURE
 PROCEEDINGS
 PROCESSING
 PROFILE
 PROG. LANGUAGE
 PROGRAM
      SN   COMPUTER PROGRAM
      SA   ROUTINE
      SA   SOFTWARE
      SA   SUBROUTINE
 PROGRAMMED
+PROPERTY
      S    CHARACTERISTIC
 PSYCHOLOGY
+PUBLICATION
      S    DOCUMENT
 PUNCHED
+PUNCHED-CARD
      S    STORAGE
```

```
 PUNCTUATION
+PURPOSE
      S    OBJECTIVE


 QUALITATIVE
      SA   SUBJECTIVE
 QUANTITATIVE
+QUERY
      S    QUESTION
 QUESTION
      SN   BOTH NOUN AND VERB
 QUESTION-ANSWER


 RANDOM
 RANDOM-ACCESS
 RANK
 READING
 REAL-TIME
 RECALL
 RECOGNITION
 RECORD
+RECORDED INFO.
      S    RECORD
 RECURSIVE
      SA   ITERATIVE
 REDUNDANCY
 REFERENCE
*REJECTION
 RELATED
 RELATIONSHIP
 RELATIVE
 RELEVANCE
 RELEVANT
      SA   PERTINENT
+REMOTE TELETYPES
      S    REMOTE TERMINAL
 REMOTE TERMINAL
      SA   VISUAL DIS. CON.
+REPORT
      S    DOCUMENT
+REQUEST
      S    QUESTION
 RESEARCH
+RESPONSE
      S    ANSWER
 RESPONSE TIME
 RETRIEVAL
 RETRIEVAL SYSTEM
 REVIEW
      SA   SUMMARY
      SA   SURVEY
 ROLE
```

```
ROUTINE
     SN  COMPUTER ROUTINE
     SA  PROGRAM
     SA  SOFTWARE
     SA  SUBROUTINE
RULE


SAMPLE
SCANNING
SCIENTIFIC
SCOPE NOTE
SEARCH CRITERIA
SEARCH STRATEGY
SEARCHING
*SECRECY
SEE ALSO
     SN  AS USED IN CATALOGING
SEE-REFERENCE
SELECTION
SELECTIVE DISSEM
SEMANTIC
     SA  SYNTAX
SEQUENCE
+SERIAL
     S   JOURNAL
SERVICE
SET THEORY
SETS
SHELFLIST
SIGNIFICANCE
SIMULATION
     SA  MODEL
SIZE
SMALL
SOCIAL IMPLIC.
SOFTWARE
     SA  PROGRAM
     SA  ROUTINE
     SA  SUBROUTINE
SORTING
SOURCE
SPECIALIZED
SPECIFICITY
STANDARDIZATION
STAT ASSOCIATION
STAT. ANALYSIS
     SA  STAT. METHOD
STAT. METHOD
     SA  STAT. ANALYSIS
STATE-OF-THE-ART
STATISTICAL
+STOCHASTIC
     S   RANDOM
STORAGE
```

```
STRING
     SA  FILE
     SA  LIST
STRUCTURE
SUBJECT
SUBJECT HEADING
SUBJECT INDEXING
SUBJECT-CATALOG.
+SUBJECTIVE
     SA  QUALITATIVE
SUBROUTINE
     SA  PROGRAM
     SA  ROUTINE
     SA  SOFTWARE
SUMMARY
     SA  REVIEW
     SA  SURVEY
SURVEY
     SA  REVIEW
     SA  SUMMARY
SYMBOL
SYMBOLIC LOGIC
SYMPOSIUM
     SA   COLLOQUIUM
     SA   CONFERENCE
     SA   MEETING
SYNONYM
SYNTACTIC ANAL.
SYNTAX
     SA  SEMANTIC
SYSTEM


TABLE
     SA  GRAPH
TAG
     SA  DESCRIPTOR
     SA  KEYWORD
     SA  TERM
+TAPE
     S   STORAGE
+TEACHING
     S   EDUCATION
TECHNICAL
TECHNICAL REPORT
TECHNOLOGY
TELEGRAPHIC ABS.
TERM
     SA  DESCRIPTOR
     SA  KEYWORD
     SA  TAG
+TERMINAL
     S   REMOTE TERMINAL
TERMINOLOGY
     SA  NOTATION
```

```
 TEST
      SA  EVALUATION
      SA  UTILITY
      SA  VALUE
 TEXT
 THEORY
 THESAURUS
      SA  AUTHORITY LIST
 TIME
 TIME-SHARING
 TITLE
+TOPIC
      S   SUBJECT
 TRANSFORMATION
 TRANSLATION
*TRANSLITERATION
 TRANSMISSION
 TREE
 TREE STRUCTURE
 TRUNCATION
*TYPE STYLE
 TYPE-SETTING
*TYPOGRAPHICAL


+UNION
      SN  SET THEORY UNION
      S   VENN DIAGRAM
*UNION CATALOG
+UNITERM
      S   DESCRIPTOR
 UNITERM SYSTEM
      SA  COORDINATE INDEX
 UPDATING
 USER
 UTILITY
      SA  EVALUATION
      SA  TEST
      SA  VALUE


 VALIDATION
 VALUE
      SA  EVAUATION
      SA  TEST
      SA  UTILITY
 VARIABLE
      SA  PARAMETER
 VECTOR
 VENN DIAGRAM
*VISUAL DIS. CON.
      SA  REMOTE TERMINAL
 VOCABULARY


 WEIGHT
 WEIGHT INDEXING
 WORD
 WORD ASSOCIATION
 WORD FREQUENCY
+WORD PAIRS
      S   WORD ASSOCIATION
```

APPENDIX 4b: INDEX TERM LIST SORTED ON FREQUENCY OF ASSIGNMENT

INDEX TERM	NO. OF REFS.
INFO. RETRIEVAL	84
SYSTEM	84
DOCUMENT	78
COMPUTER	69
STORAGE	69
INDEXING	64
RETRIEVAL	63
INFORMATION	59
SEARCHING	58
ANALYSIS	53
CLASSIFICATION	52
STRUCTURE	52
INDEX	49
RELEVANCE	49
LANGUAGE	46
EVALUATION	44
EXPERIMENT	44
ASSOCIATION	42
SEMANTIC	41
MATRIX	39
NATURAL LANGUAGE	38
WORD	36
FREQUENCY	35
DESCRIPTOR	34
QUESTION	33
DICTIONARY	32
PROGRAM	32
USER	32
DATA	31
MEASURE	31
TRANSLATION	31
LIBRARY	30
RELATIONSHIP	30
THESAURUS	30
HIERARCHY	29
ALGORITHM	28
AUTOMATIC	28
COMMUNICATION	28
INPUT	28
LINGUISTIC	28
STATISTICAL	28
SYNTAX	28
PROBABILITY	27
GRAMMAR	26
OUTPUT	26
QUESTION-ANSWER	26
REFERENCE	26
WORD ASSOCIATION	25
LITERATURE	24
FILE	22
LOGIC	22
MATCH	22
PROCESSING	22
RELEVANT	22
SEARCH STRATEGY	22
SYMBOL	22
TECHNICAL	22
AUTO. INDEXING	21
BIBLIOGRAPHIC	21
SCIENTIFIC	21
STAT. METHOD	21
CONCEPT	20
EFFICIENCY	20
RECALL	20
TEXT	20
THEORY	20
ABSTRACT	19
CO-OCCURRENCE	19
CODING	19
KEYWORD	19
TRANSFORMATION	19
WEIGHT	19
GRAPH	18
VOCABULARY	18
CLUMP	17
HARDWARE	17
MODEL	17
SUBJECT	17
SYNONYM	17
SYNTACTIC ANAL.	17
TREE	17
COMPARISON	16
COORDINATE INDEX	16
CORRELATION	16
MECHANIZATION	16
TAG	16
TEST	16
ACCESS	15
BIBLIOGRAPHY	15
CLASSIF. SCHEME	15
CONTENT	15
COST	15
EDUCATION	15
LATTICE	15
LINK	15
MATHEMATICAL	15
RETRIEVAL SYSTEM	15
TITLE	15
ASSOCIATIVE	14
MEANING	14
NETWORK	14
RESEARCH	14
SCANNING	14
SERVICE	14
ABSTRACTING	13
BOOLEAN	13
CITATION INDEX	13

CLUSTER	13
FUNCTION	13
LIBRARIAN	13
LIST	13
ORDER	13
PARSE	13
RANDOM	13
SEQUENCE	13
CITATION	12
MAN-MACHINE	12
PROG. LANGUAGE	12
SURVEY	12
VALUE	12
VARIABLE	12
AUTO ABSTRACTING	11
CODE	11
COEFFICIENT	11
COLLECTION	11
CONTEXT	11
DISSEMINATION	11
METHODOLOGY	11
NOISE	11
PRECISION	11
PROCEDURE	11
SETS	11
SUBJECT HEADING	11
VECTOR	11
ADDRESS	10
AUTOMATION	10
CHARACTERISTIC	10
CURRICULUM	10
DEGREE	10
DOCUMENTATION	10
ILLUSTRATION	10
MECHANICAL	10
NOTATION	10
STAT ASSOCIATION	10
ALGEBRA	9
CATEGORIES	9
COORDINATE	9
DESIGN	9
ERROR	9
GENERAL	9
GENERIC	9
INFERENCE	9
INFO. SCIENCE	9
KWIC	9
NEEDS	9
PARAMETER	9
ROLE	9
RULE	9
STATE-OF-THE-ART	9
STRING	9
TREE STRUCTURE	9
BOOK	8
CARD	8
CATALOG	8
CONTENT ANALYSIS	8
CRITERIA	8
DEPTH OF INDEXIN	8
EDITING	8
FALSE DROP	8
FEEDBACK	8
FILE ORGANIZATIO	8
INTRODUCTORY	8
JOURNAL	8
MANUAL	8
MATHEMATICS	8
ORGANIZATION	8
PROCEEDINGS	8
ROUTINE	8
SPECIFICITY	8
TECHNOLOGY	8
ANSWER	7
BINARY	7
CATALOGING	7
CONNECTION	7
COUPLING	7
CROSS REFERENCE	7
CURRENT AWARENES	7
DESCRIPTIVE	7
INTERPRET	7
REMOTE TERMINAL	7
ALPHABETIC	6
AMBIGUITY	6
COMP LINGUISTICS	6
CONDITIONAL PROB	6
CONFERENCE	6
ENTRY	6
LAW	6
PATENT	6
PERMUTED	6
PREDICTION	6
RECOGNITION	6
REDUNDANCY	6
RELATIVE	6
SIGNIFICANCE	6
TIME-SHARING	6
WORD FREQUENCY	6
AUTHOR	5
CANONICAL	5
CENTERS	5
CHEMICAL	5
DISCRIMINANT	5
FACET	5
FLOW OF INFO.	5
FORMAT	5
INTERDISCIPLINAR	5
IRRELEVANT	5
MEETING	5
NATIONAL	5
PATTERN	5

PERFORMANCE	5
PERTINENT	5
PUNCHED	5
RANDOM-ACCESS	5
RECURSIVE	5
REVIEW	5
SIMULATION	5
SORTING	5
SUBJECT INDEXING	5
SYMPOSIUM	5
TABLE	5
TIME	5
UNITERM SYSTEM	5
WEIGHT INDEXING	5
ACCESSION NUMBER	4
CIRCULATION	4
CRANFIELD	4
ENTROPY	4
FACETED CLASSIF.	4
HISTORICAL	4
INTELLECTUAL	4
LOGICAL	4
OPERATION	4
RECORD	4
SAMPLE	4
SELECTION	4
SELECTIVE DISSEM	4
SOFTWARE	4
SOURCE	4
SUBJECT-CATALOG.	4
SYMBOLIC LOGIC	4
UTILITY	4
ACCURACY	3
ACQUISITION	3
APPLICATION	3
ARRAY	3
CENTRALIZED	3
CLERICAL	3
COMBINATIONS	3
CONCORDANCE	3
CONTROLLED	3
CONVENTIONAL	3
CONVERSION	3
COUNT	3
DECISION THEORY	3
DEDUCTIVE	3
EXTRACT	3
GENERATION	3
KEYPUNCH	3
MACHINE-READABLE	3
MEDIUM	3
MICROFICHE	3
MICROFILM	3
MULTIPLE	3
NON-RELEVANT	3
NUMERIC	3
ON-LINE	3
PHILOSOPHY	3
PLANNING	3
PROGRAMMED	3
READING	3
SPECIALIZED	3
TELEGRAPHIC ABS.	3
TRANSMISSION	3
TRUNCATION	3
VENN DIAGRAM	3
ALPHANUMERIC	2
ANALOGY	2
ARTIFICIAL INTEL	2
ASSIGNED	2
AUTHORITY LIST	2
CARD CATALOG	2
EFFECTIVENESS	2
FACT RETRIEVAL	2
IDENTIFICATION	2
INTERFACE	2
INTUITIVE	2
INVENTORY	2
LARGE	2
MANUAL INDEXING	2
NATURAL	2
NON-CONVENTIONAL	2
OBJECTIVE	2
OFF-LINE	2
PROFILE	2
QUALITATIVE	2
QUANTITATIVE	2
RANK	2
SCOPE NOTE	2
SEARCH CRITERIA	2
SEE ALSO	2
SET THEORY	2
SIZE	2
SOCIAL IMPLIC.	2
SUBROUTINE	2
SUMMARY	2
TYPE-SETTING	2
VALIDATION	2
ADMINISTRATION	1
ALPHABETIC ORDER	1
BATCH PROCESSING	1
CALL NUMBER	1
CONTROL	1
CRITICAL	1
DUAL DICTIONARY	1
GOVERNMENT	1
GRAPHICS	1
INDEPENDENT	1
ITERATIVE	1
MODIFICATION	1
NON-DISCRIMINANT	1
NON-FILE	1

```
NON-RANDOM               1
NUMBER                   1
OPTIMIZATION             1
PRINTING                 1
PSYCHOLOGY               1
PUNCTUATION              1
REAL-TIME                1
RELATED                  1
RESPONSE TIME            1
SEE-REFERENCE            1
SHELFLIST                1
SMALL                    1
STANDARDIZATION          1
STAT. ANALYSIS           1
TECHNICAL REPORT         1
TERMINOLOGY              1
UPDATING                 1
ABBREVIATION             0
ALTERNATIVES             0
ANTHOLOGY                0
CLAIM                    0
COLLOQUIUM               0
COPYRIGHT                0
DECENTRALIZATION         0
DISSERTATION             0
IDENTICAL                0
IMPLEMENTATION           0
INVERTED                 0
NORMALIZED               0
OCCURRENCE               0
PRINCIPLE                0
REJECTION                0
SECRECY                  0
TRANSLITERATION          0
TYPE STYLE               0
TYPOGRAPHICAL            0
UNION CATALOG            0
VISUAL DIS. CON.         0
```

APPENDIX 4c: INDEX TERM LIST ALPHABETICALLY SORTED

INDEX TERM	NO. OF REFS.
ABBREVIATION	0
ABSTRACT	19
ABSTRACTING	13
ACCESS	15
ACCESSION NUMBER	4
ACCURACY	3
ACQUISITION	3
ADDRESS	10
ADMINISTRATION	1
ALGEBRA	9
ALGORITHM	28
ALPHABETIC	6
ALPHABETIC ORDER	1
ALPHANUMERIC	2
ALTERNATIVES	0
AMBIGUITY	6
ANALOGY	2
ANALYSIS	53
ANSWER	7
ANTHOLOGY	0
APPLICATION	3
ARRAY	3
ARTIFICIAL INTEL	2
ASSIGNED	2
ASSOCIATION	42
ASSOCIATIVE	14
AUTHOR	5
AUTHORITY LIST	2
AUTO ABSTRACTING	11
AUTO. INDEXING	21
AUTOMATIC	28
AUTOMATION	10
BATCH PROCESSING	1
BIBLIOGRAPHIC	21
BIBLIOGRAPHY	15
BINARY	7
BOOK	8
BOOLEAN	13
CALL NUMBER	1
CANONICAL	5
CARD	8
CARD CATALOG	2
CATALOG	8
CATALOGING	7
CATEGORIES	9
CENTERS	5
CENTRALIZED	3
CHARACTERISTIC	10
CHEMICAL	5
CIRCULATION	4
CITATION	12
CITATION INDEX	13
CLAIM	0
CLASSIF. SCHEME	15
CLASSIFICATION	52
CLERICAL	3
CLUMP	17
CLUSTER	13
CO-OCCURRENCE	19
CODE	11
CODING	19
COEFFICIENT	11
COLLECTION	11
COLLOQUIUM	0
COMBINATIONS	3
COMMUNICATION	28
COMP LINGUISTICS	6
COMPARISON	16
COMPUTER	69
CONCEPT	20
CONCORDANCE	3
CONDITIONAL PROB	6
CONFERENCE	6
CONNECTION	7
CONTENT	15
CONTENT ANALYSIS	8
CONTEXT	11
CONTROL	1
CONTROLLED	3
CONVENTIONAL	3
CONVERSION	3
COORDINATE	9
COORDINATE INDEX	16
COPYRIGHT	0
CORRELATION	16
COST	15
COUNT	3
COUPLING	7
CRANFIELD	4
CRITERIA	8
CRITICAL	1
CROSS REFERENCE	7
CURRENT AWARENES	7
CURRICULUM	10
DATA	31
DECENTRALIZATION	0
DECISION THEORY	3
DEDUCTIVE	3
DEGREE	10
DEPTH OF INDEXIN	8
DESCRIPTIVE	7
DESCRIPTOR	34
DESIGN	9
DICTIONARY	32
DISCRIMINANT	5
DISSEMINATION	11
DISSERTATION	0

DOCUMENT	78
DOCUMENTATION	10
DUAL DICTIONARY	1
EDITING	8
EDUCATION	15
EFFECTIVENESS	2
EFFICIENCY	20
ENTROPY	4
ENTRY	6
ERROR	9
EVALUATION	44
EXPERIMENT	44
EXTRACT	3
FACET	5
FACETED CLASSIF.	4
FACT RETRIEVAL	2
FALSE DROP	8
FEEDBACK	8
FILE	22
FILE ORGANIZATIO	8
FLOW OF INFO.	5
FORMAT	5
FREQUENCY	35
FUNCTION	13
GENERAL	9
GENERATION	3
GENERIC	9
GOVERNMENT	1
GRAMMAR	26
GRAPH	18
GRAPHICS	1
HARDWARE	17
HIERARCHY	29
HISTORICAL	4
IDENTICAL	0
IDENTIFICATION	2
ILLUSTRATION	10
IMPLEMENTATION	0
INDEPENDENT	1
INDEX	49
INDEXING	64
INFERENCE	9
INFO. RETRIEVAL	84
INFO. SCIENCE	9
INFORMATION	59
INPUT	28
INTELLECTUAL	4
INTERDISCIPLINAR	5
INTERFACE	2
INTERPRET	7
INTRODUCTORY	8
INTUITIVE	2
INVENTORY	2
INVERTED	0
IRRELEVANT	5
ITERATIVE	1
JOURNAL	8
KEYPUNCH	3
KEYWORD	19
KWIC	9
LANGUAGE	46
LARGE	2
LATTICE	15
LAW	6
LIBRARIAN	13
LIBRARY	30
LINGUISTIC	28
LINK	15
LIST	13
LITERATURE	24
LOGIC	22
LOGICAL	4
MACHINE-READABLE	3
MAN-MACHINE	12
MANUAL	8
MANUAL INDEXING	2
MATCH	22
MATHEMATICAL	15
MATHEMATICS	8
MATRIX	39
MEANING	14
MEASURE	31
MECHANICAL	10
MECHANIZATION	16
MEDIUM	3
MEETING	5
METHODOLOGY	11
MICROFICHE	3
MICROFILM	3
MODEL	17
MODIFICATION	1
MULTIPLE	3
NATIONAL	5
NATURAL	2
NATURAL LANGUAGE	38
NEEDS	9
NETWORK	14
NOISE	11
NON-CONVENTIONAL	2
NON-DISCRIMINANT	1
NON-FILE	1
NON-RANDOM	1
NON-RELEVANT	3
NORMALIZED	0
NOTATION	10
NUMBER	1
NUMERIC	3
OBJECTIVE	2
OCCURRENCE	0
OFF-LINE	2
ON-LINE	3
OPERATION	4

OPTIMIZATION	1
ORDER	13
ORGANIZATION	8
OUTPUT	26
PARAMETER	9
PARSE	13
PATENT	6
PATTERN	5
PERFORMANCE	5
PERMUTED	6
PERTINENT	5
PHILOSOPHY	3
PLANNING	3
PRECISION	11
PREDICTION	6
PRINCIPLE	0
PRINTING	1
PROBABILITY	27
PROCEDURE	11
PROCEEDINGS	8
PROCESSING	22
PROFILE	2
PROG. LANGUAGE	12
PROGRAM	32
PROGRAMMED	3
PSYCHOLOGY	1
PUNCHED	5
PUNCTUATION	1
QUALITATIVE	2
QUANTITATIVE	2
QUESTION	33
QUESTION-ANSWER	26
RANDOM	13
RANDOM-ACCESS	5
RANK	2
READING	3
REAL-TIME	1
RECALL	20
RECOGNITION	6
RECORD	4
RECURSIVE	5
REDUNDANCY	6
REFERENCE	26
REJECTION	0
RELATED	1
RELATIONSHIP	30
RELATIVE	6
RELEVANCE	49
RELEVANT	22
REMOTE TERMINAL	7
RESEARCH	14
RESPONSE TIME	1
RETRIEVAL	63
RETRIEVAL SYSTEM	15
REVIEW	5
ROLE	9
ROUTINE	8
RULE	9
SAMPLE	4
SCANNING	14
SCIENTIFIC	21
SCOPE NOTE	2
SEARCH CRITERIA	2
SEARCH STRATEGY	22
SEARCHING	58
SECRECY	0
SEE ALSO	2
SEE-REFERENCE	1
SELECTION	4
SELECTIVE DISSEM	4
SEMANTIC	41
SEQUENCE	13
SERVICE	14
SET THEORY	2
SETS	11
SHELFLIST	1
SIGNIFICANCE	6
SIMULATION	5
SIZE	2
SMALL	1
SOCIAL IMPLIC.	2
SOFTWARE	4
SORTING	5
SOURCE	4
SPECIALIZED	3
SPECIFICITY	8
STANDARDIZATION	1
STAT ASSOCIATION	10
STAT. ANALYSIS	1
STAT. METHOD	21
STATE-OF-THE-ART	9
STATISTICAL	28
STORAGE	69
STRING	9
STRUCTURE	52
SUBJECT	17
SUBJECT HEADING	11
SUBJECT INDEXING	5
SUBJECT-CATALOG.	4
SUBROUTINE	2
SUMMARY	2
SURVEY	12
SYMBOL	22
SYMBOLIC LOGIC	4
SYMPOSIUM	5
SYNONYM	17
SYNTACTIC ANAL.	17
SYNTAX	28
SYSTEM	84
TABLE	5
TAG	16
TECHNICAL	22

TECHNICAL REPORT	1
TECHNOLOGY	8
TELEGRAPHIC ABS.	3
TERMINOLOGY	1
TEST	16
TEXT	20
THEORY	20
THESAURUS	30
TIME	5
TIME-SHARING	6
TITLE	15
TRANSFORMATION	19
TRANSLATION	31
TRANSLITERATION	0
TRANSMISSION	3
TREE	17
TREE STRUCTURE	9
TRUNCATION	3
TYPE STYLE	0
TYPE-SETTING	2
TYPOGRAPHICAL	0
UNION CATALOG	0
UNITERM SYSTEM	5
UPDATING	1
USER	32
UTILITY	4
VALIDATION	2
VALUE	12
VARIABLE	12
VECTOR	11
VENN DIAGRAM	3
VISUAL DIS. CON.	0
VOCABULARY	18
WEIGHT	19
WEIGHT INDEXING	5
WORD	36
WORD ASSOCIATION	25
WORD FREQUENCY	6

APPENDIX 5

LABSRC 3: A DETAILED DESCRIPTION

LABSRC 3: A DETAILED DESCRIPTION

1. Introduction

LABSRC 3 is one of three search programs designed to teach and demonstrate information retrieval techniques to students and faculty of the Library School. Of the three programs, LABSRC 3 is by far the most sophisticated because it allows true interaction between the user and the program and also permits the user to submit requests in the form of Boolean expressions. This appendix describes LABSRC 3 with particular emphasis on those features that distinguish it from the other two search programs.

2. General

The program allows requests in the form of Boolean expressions that consist of valid index terms joined together with the usual logical connectives. Weights can be assigned by the user to particular index terms and to parenthetic subexpressions, so that the relative importance of different parts of the query can be indicated. LABSRC 3 also provides options so that the search can be performed in direct-match mode or associative-retrieval mode and so that relevance numbers for the retrieved documents can be computed if the user so desires. When the options have been specified and the query submitted, LABSRC 3 searches the MASTER I file.* The program utilizes reasonably advanced techniques to minimize search time. The query, for instance, is analyzed by a parser that puts out directly executable code in the form of a subroutine that embodies the logic of the query. Since the logic is to be analyzed once for each document, this technique is superior to interpretive methods.

LABSRC 3 asks six questions during a normal pass through the program so that the options and the query can be input by the user. These questions are:

Q01 Do you want word association?

Q02 Specify association file

Q03 Do you want scoring?

Q04 Enter Boolean expression

Q05 Do you want results printed?

Q06 Specify restart or exit.

The questions are self-explanatory and the answers given by the user are straightforward. However, at any time, instead of answering the question with a relevant answer, the user can input a command in the command language and essentially take over control to exploit the

*If the user so desires, the query may be expanded using terms drawn from any one of three files of term association data.

program fully. This is what makes LABSRC 3 truly interactive. Entering commands is therefore quite easy and natural since LABSRC 3 figures out whether the reply is an answer to the question or a command.

3. Boolean Expressions and Request Formulation

The syntax of Boolean Expressions accepted by LABSRC 3 as legal requests is given below in Backus-Naur Form.

<Index Term> = any legal term that belongs to the thesaurus

<Decimal no.> = any 4 digit decimal number n, where $0 \leq n \leq .9999$

<Primary> = ' <Index Term> ' | (<Boolean Expression>)

<Primary Exp.> = <Primary> | NOT <Primary>

<Secondary> = <Primary Exp.> | <Decimal no.> * <Primary>

<AND-Exp.> = <Secondary> | <Secondary> AND <AND-exp.>

<Boolean Expression> = <AND Expr.> | <Boolean Expression> OR <AND-Exp.>

<Request> = <Boolean Expression>

Note that any index term or parenthetic sub-expression can be weighted down by multiplying it by a decimal number.

Ex. without weights:

('Language' OR 'syntax') AND NOT 'grammar'

Ex. with weights:

0.5670* ('Language' OR 0.5000* 'syntax') AND NOT 'grammar'

It is suggested that weights for individual index terms be assigned through the ASSIGN command rather than explicitly typing them in.

When Boolean expressions that are longer than one line are to be typed in, the last character in any incomplete line must be @. LABSRC 3 concatenates the lines together. The carriage is returned to indicate that an incomplete line has been input.

4. Associative Retrieval and Scoring

As pointed out earlier, LABSRC 3 is capable of searching the MASTERI file in either direct match or associative mode. In the former mode the user's request is used as submitted, while in the latter mode the request is extended to include more terms from the association file selected by the user. Each index term can be extended to include up to a maximum of 4 associated terms. Some or all of these terms can be later eliminated from the search by the user with the help of the command language.

When scoring is specified, LABSRC 3 calculates relevance numbers that reflect the closeness between the request and the document. When direct match searching is asked for, scoring is obviously unnecessary but can be specified.

Since association values reflect the degree of correlation between any two terms, these values are used to obtain some measure of relevance for the documents found.

When scoring is utilized, an AND between any two terms or subexpressions results in their values being multiplied. An OR results in the term or subexpression with the higher association value being chosen.

The NOT in LABSRC 3 is a unary operator and, when scoring is utilized, is treated as follows:

NOT a1 where a1 is a term or a subexpression

If the effective value of a1 ≠ 0, it is made 0. If the effective value of a1 = 0, it is made .9999.

When weights have been specified, the program uses simple multiplication to incorporate the effects of weights in the relevance number.

The following example assumes that the user has requested scoring and word association using the KUHNS W file. Let us also assume that the input expression is:

.500* ('LANGUAGE' AND 'GRAMMAR') AND NOT 'SYNTAX'

From the KUHNS W file the input terms will be expanded as follows:

Term	Association Value	Term	Association Value
Language	.9999	Grammar	.9999
Objective	.8381	Parse	.4469
Social Implic.	.8381	Syntactic. Anal.	.4379
Standardization	.8381	Fact Retrieval	.4085
Related	.8381	Set Theory	.4085

Even though word association has been requested, 'Syntax' will not be expanded since it is a negated term.

Now, during the search let us assume that we are looking at document A0103, which is indexed under the following terms:

Algorithm
Analysis
Automatic
Classification
Clump
Cluster
Computer
Dictionary
Document
Documentation
Evaluation
Indexing
Information
Keyword
Language
Natural Language
Output
Processing
Prog. Language
Relevance
Retrieval
Retrieval System
State-of-the-Art
Statistical
Survey
Syntactic Anal.
Thesaurus
Time Sharing
Translation

Since 'Language' occurs in the document, it will be represented by a value of .999. Although 'grammar' does not occur, 'Syntactic anal.' does. 'Grammar' will be represented by a value of .4379. 'Syntax' does not occur. Therefore, A0103 satisfies the input expression and its relevance value is .500 x (.999 x .437) x .999 which works out to .216.

If weights had been assigned to individual terms, these weights would have been multiplied into the effective values for each term before the evaluation of the expression.

5. The Command Language

Design criteria. The command language was designed with three main objectives in mind. First, the language should be easy to use and should be equally amenable to novice and sophistocated users. Second, it should allow the user to interact with the program effectively and should familiarize the user with all aspects of LABSRC 3. Lastly, the language should be easy to implement.

The first objective was met by allowing commands to be input in pseudo-natural language. A text analyser was written that analyses the given command and transforms it into an internal canonical form containing the relevant parts of the command. The second was met by permitting a large variety of commands. The last objective was met by specifying that the commands be in the form of a verb and a predicate.

Forms of commands. As indicated above a command consists of a verb and a predicate. The verb can be one of the following:

1) Display
2) Count
3) Modify
4) Search
5) Proceed
6) Replace
7) Replace op. wt.
8) Replace str. wt.
9) Execute
10) Initialize
11) Assign
12) Go to
13) Sorta
14) Sortd
15) Exit

The predicate consists of any sentence in natural language containing keywords or a cryptic form consisting of one of the same keywords (with the exception of the predicate for EXECUTE).

For example, the DISPLAY and COUNT commands result in output to the terminal. Since mechanical terminals and CRT terminals over a slow speed communications line are fairly slow, the COUNT command can be used to find out how much information is going to be output. The DISPLAY command can then be used to output the data or parts of the data; for example:

COUNT the number of documents found

DISPLAY all documents with a relevance value *GT* .1234

DISPLAY all the most highly associated terms

DISPLAY the terms associated with 'grammar'

These commands, as shown, are perfectly valid. As can be seen, the predicate can be in natural language. The keywords have been underlined.

One should note that keywords and numbers in the predicate result in the outputting of parts of lists in memory. A sentence without keywords will generally output the whole list. Therefore, a naive user will generally get more than he wants. For example, the command DISPLAY association data (no keywords) will result in the whole association table being displayed. The COMMAND analyzer checks for the order of the keywords, among other things, and issues diagnostic messages when there is ambiguity.

It is also possible for the user to enter his request in a cryptic form containing just the keywords, for brevity. The analyzer converts the forms below into a 24 byte 'instruction' which is then interpreted by various routines. An advanced programmer may prefer an assembler-like language. This is done by using the EXECUTE commands:

```
DISPLAY all terms associated with 'grammar' with an
        association value *LT*.5678

DISPLAY 'grammar' *LT*.5678

EXECUTE D4 grammar,,.5678,L
```

The above three commands are equivalent. The predicate of the EXECUTE consists of the subfields of the internal canonical form produced by the analyzer.

Use of Commands. The questions asked by LABSRC 3 during a normal pass* are identified by numbers Q01-Q06 as shown in Section 2.

a) Replies to Q01, Q02, Q03, Q05, Q06 can be commands.

*A pass that has not been interrupted with a command.

b) A reply to Q04 must not be a command.

c) Once a command has been entered, the normal predefined program flow is no longer in operation and any number of commands may be given.

d) It is suggested that during the first pass through the program or after a "restart" reply to Q06, no commands are typed in as answers to Q01-Q04 (other than a forward default branch - explained later).

6. The Commands

This section describes the actual commands in detail. They are described according to their functions as shown below.

6.1 Branching Commands

GO TO Q-- or GO TO Q--S cause the program to branch to the question number indicated, i.e. normal program flow is interrupted. There are two kinds of branches - forward and backward. If the GO TO Q--(S) command refers to a question number less than or equal to the present question number where the command has been typed in, this branch is referred to as a backward branch. Otherwise it is a forward branch.

Backward branches do not require any further explanation.

If the branch is a forward branch, a GO TO Q-- causes a branch using default answers (see table below) for all intermediate questions skipped over. A GO TO Q--S forward branch causes a branch using the most current answers (specified by the user) for all intermediate questions skipped over.

Question No.	Default Options
Q01	yes
Q02	DOYLE
Q03	yes
Q04	previous Boolean expression
Q05	no
Q06	exit

The PROCEED command returns control to normal program flow. If at any time flow has been interrupted by commands, PROCEED causes the program to branch to the next question.

EXIT causes the program to exit and control is returned to.the Terminal Monitor System.

6.2 Commands that Assign and Edit Weights for Boolean Expressions

a) Weights may be assigned explicitly in the Boolean expression with the * operator. These weights are referred to below as string weights. String weights, of course, may be weights for individual index terms or for parenthetic subexpressions, i.e.,* must be followed by an index term or a left parenthesis.

b) Weights may be assigned to individual operands (index terms) in the Boolean expression at any time by the ASSIGN .---- to '-----' command. These weights are called operand weights.

c) String weights may be changed with the REPLACE STR. WT. •----- by •----- command. <u>All</u> string weights equal to the first weight in the command will be replaced by the second.

d) Operand weights may be changed similar to c) by the REPLACE OP. WT. •----- by •----- command.

e) A REPLACE •----- by •----- command results in <u>both</u> string and operand weights being changed.

<u>Notes to a-e above</u>:

a-e: The weights must be 4 digit decimal numbers between 0 and 1.

a-e: When editing has been done, remember that the weights assigned are the most current set of weights.

a-e: Diagnostics are provided if the weights in a command are greater than 1 or if the index term in the ASSIGN command does not exist.

b: An index term may be eliminated from the search by assigning a weight of .0000.

a-b: Note that weights for index terms can be string weights or can be entered as operand weights with the ASSIGN command.

c-e: If the first weight in any replace command does not exist, no modification is done. No diagnostic is provided.

6.3 Commands that Edit Association Data

This is primarily done by the SEARCH and/or MODIFY commands. If n terms exist in the Boolean expression one can visualize the association data as a table of n rows and 5 columns. Each row corresponds to the original term and its four associated terms along with their association values. The above commands permit the user to selectively eliminate terms from the search. (Terms can only be eliminated. A term previously eliminated cannot be reinstated unless an INITIALIZE is executed. This command has no predicate.)

The commands allow the user to operate on one selected row at a time or on all rows simultaneously. The SEARCH and MODIFY commands can be classified into two types - those that eliminate a certain number of columns or those that allow terms to be eliminated on the basis that these association values are greater than, equal to, or less than some specified threshold. The difference between SEARCH and MODIFY commands is that while SEARCH automatically initiates a search after the prescribed modification, MODIFY does not initiate a search.

Exs: SEARCH using only the most highly associated terms

SEARCH using terms *EQ*.9999

General Notes:

a) Ex: MODIFY to use association to a depth of 2 terms implies that only the two most highly associated terms should be used.

b) An INITIALIZE restores the table to the state corresponding to the most current answers to Q01 and Q04.

c) A change in the answer to Q02 is not reflected in the table until a search has been made.

d) Note that the original terms themselves can be eliminated from the search with the SEARCH or MODIFY commands.

Ex: SEARCH using terms *LT*.9999

6.4 Commands Relating to the Display of Documents

a) SORTD and SORTA sort the documents in descending or ascending order or relevance respectively. These make sense, of course, only when scoring has been asked for. These commands have no predicate.

b) The COUNT commands pertaining to documents count the number of documents (relative to a threshold if specified).

Ex: COUNT documents with relevance values *GT*.5000

c) The DISPLAY commands pertaining to documents display a specified number of documents (relative to a threshold, if specified).

Ex: DISPLAY 7 documents

Notes to a-c above:

a,c: With the combination of the SORT(A/D) and DISPLAY commands, selected portions of the list of documents found can be output.

c: The commands allow a certain amount of guessing. Ex: Assume DISPLAY 8 documents *GT*.5000 is input. If only 5 documents have relevance values greater than .5000, only 5 will be output. If 15 such documents exist, only 8 will be output.

6.5 Commands Pertaining to the Display of Association Data

a) The COUNT commands pertaining to terms allow the user to count the number of terms relative to a specified association value.

Ex: COUNT terms *EQ*.8000

b) The DISPLAY commands pertaining to terms allow the user to display parts of the association table. Here a selected row or all rows can be displayed to any depth (maximum of 4).

Ex: DISPLAY the most highly associated terms.

Notes to a-b above:

b: DISPLAYing terms relative to a threshold is not permitted.

b: *'s are printed after terms that are not to be included in the search. These, of course, are terms eliminated by SEARCH/MODIFY commands and/or terms eliminated due to their appearance in negated substrings in the Boolean Expression.

Table 1 indicates the forms of legal commands.

TABLE 1: FORMS OF LEGAL COMMANDS

	COMMAND	
VERB	PREDICATE KEYWORDS	LEGEND
GO TO	QO-	- = 1,2,3,4,5, or 6
GO TO	QO-s	
PROCEED	no keywords	
EXIT	no keywords	
ASSIGN	•----, 'index term'	index term = any operand in Boolean expression
REPLACE OP. WT.	•---- , •----	•---- = 4 digit decimal no.
REPLACE STR. WT.	•---- , •----	
REPLACE	•---- , •----	
SEARCH	no keywords	
SEARCH/MODIFY	no	(Note: here "no" means no association as in SEARCH USING NO ASSOC.)
SEARCH/MODIFY	no, 'index term'	
SEARCH/MODIFY	x_1, 'index term'	x_1 = 1,2,3, or 4
SEARCH/MODIFY	x_1	
SEARCH/MODIFY	'index term', $*x_2*$, •----	x_2 = GT, LT, or EQ
SEARCH/MODIFY	$*x_2*$, •----	
SEARCH/MODIFY	most, 'index term'	
SEARCH/MODIFY	most	
INITIALISE/INITIALIZE	no keywords	
SORTD	no keywords	
SORTA	no keywords	
COUNT	documents	

TABLE 1 (Continued)

	COMMAND	
VERB	PREDICATE KEYWORDS	LEGEND
COUNT	documents, *x_2*, ·-----	
DISPLAY	x_3, documents	x_3 is optional - may be absent or be a number
DISPLAY	x_3, documents, *x_2*, ·-----	x_3 is optional
COUNT	'index term', *x_2*, ·-----	
COUNT	*x_2*, ·-----	
DISPLAY	no keywords	
DISPLAY	'index term'	
DISPLAY	x_1, 'index term'	
DISPLAY	x_1	
DISPLAY	most	
DISPLAY	most, 'index term'	

APPENDIX 6

UTILITY PROGRAMS

UTILITY PROGRAMS

MASTER I Generator. Generates from cards an indexed file and corresponding printed output. The file is indexed on document accession number and each entry contains the index terms for that document.

INVERT I Generator. Generates from the MASTER I file an inverted file (with respect to index terms) and corresponding printed output. INVERT I is a sequential file and each entry contains the accession numbers of documents assigned a particular index term and the total number of documents using that term.

CO-OCCURRENCE Generator. Generates from INVERT I a sequential file and printed output. Each record corresponds to an index term and is one row of the co-occurrence matrix (i.e., the matrix whose elements are the number of times a particular pair of index terms co-occur in the MASTER I file). The printout from this program permits one to readily note the number of co-occurrences of any pair of terms in the file.

ASSOCIATION COEFFICIENT Generator. Generates an indexed file (indexed on index terms) and printed output. Each entry contains the four index terms most highly associated with the header term together with their coefficients of association. The program is set up so that different associative measures may be generated by using different subroutines.

MASTER B Generator. Generates a sequential file and produces an equivalent printout. Each entry, defined by a document accession number, contains the bibliographic information (author, title, publisher, etc.) for that document.

INVERT B Generator. Inverts the MASTER B file with respect to author and generates a printed output. Documents written by more than one author are listed under each author and "see also" notes refer to the other author(s).

APPENDIX 7

DOCUMENT ATTRIBUTE CODES

DOCUMENT ATTRIBUTE CODES

(Reprinted from An Experimental Inquiry into Context Information Processing by Laura Gould et al, Institute of Library Research, Berkeley, January 1969, pp. 93-95. Funded by National Science Foundation Grant No. GN643.)

A. MAJOR CODES

Code	Meaning	Example
AD	Author-assigned descriptor	110
AU	Author (personal) of corpus or cited doc.	BAR-HILLEL, Y.
BS	Book series plus series number	NBS MISC PUBL*NO.269
CA	Corporate author	RAND
ED	Editor of a book	LIVINGSTON, H.H.
ID	Indexer-assigned descriptor	3
JO	Journal name plus issue number	JACM*VOL.1,NO.2
LD	Lab descriptor	279
PU	Publisher	SPARTAN
RA	Review author	OPLER, A.
RD	Reviewer-assigned descriptor	7
RJ	Review journal name plus issue number	COMPUT REV*VOL.6,NO.5
RS	Report series plus series number	ASTIA AD SERIES*AD NO.231606
XC	Computing Reviews index tag	3.7
XE	IEEE index tag	PROGRAMMING
XM	Math Reviews index tag	1
YR	Year of publication	65

A. MAJOR CODES (cont.)

Code	Meaning	Example	Explanation
BI	Bibliography	B2 BI A1	Indicates that B2 is part of the document file because it was found in a bibliographic list at the end of A1
CI	Citation	B2 CI A1	Indicates that B2 is part of the document file because its content was specifically discussed (i.e. cited) in the body of A1
CM	Comment	B2 CM A1	Indicates that the content of B2 constitutes a comment upon or an answer to A1
RE	Reference	B2 RE A1	Indicates that B2 is part of the document file because it was mentioned without specific discussion of its content, in the body of A1

B. MINOR CODES

Code	Meaning	Example(s)	Explanation
CO	Collation	TH.,JUNE PR. FALL	Contains notation of thesis (TH.), preprint (PR.), months and other notes
PP	Pagination	176-9	Page numbers of document represented with a minimum number of digits
PR	Presented at	ANNUAL MEETING OF THE AMERICAN SOCIETY FOR ENGINEERING	Name of congress, meeting, etc. where document was presented, ususally including date and place

B. MINOR CODES (cont.)

Code	Meaning	Example	Explanation
DU	Duplicate	B2 DU A1	Indicates that B2 is a duplicate of A1, in the sense that it is identical to it, but published elsewhere
RL	Related	B2 RL A1	Indicates that B2 and A1 are related in the sense that they are parts of a sequence of documents (e.g. PART 1, PART 2, or VOL. 1, VOL. 2)
VE	Version	B2 VE A1	Indicates that B2 constitutes another version of A1, which is modified, expanded, etc., but not identical with it (see DU above)

C. TITLE CODES

Code	Meaning	Example	Explanation
BO	Book	ICSI PROC 1958	Title of a book
JT	Journal title (theme)	PARAMETERS OF INFORMATION	Title or theme of a journal issue
RP	Report	AUTOMATIC INDEXING	Title of a report bound alone
BA	Book article	THE BASIC TYPES OF INFORMATION	Chapter of a book
JA	Journal article	SYNTACTIC STRUCTURE AND AMBIGUITY IN ENGLISH	Title of a journal article
RR	Report	DOD USER STUDY, PHASE I	Individual report bound within a volume containing several reports

APPENDIX 8

PROGRAM-STUDENT INTERACTION: TYPICAL SEQUENCE

PROGRAM - STUDENT INTERACTION: TYPICAL SEQUENCE

```
THERE IS A CODE IN EFFECT FOR TRACINGS:

THOSE NUMBERED WITH  ARABIC NUMERALS ARE FOR SUBJECT ADDED ENTRIES.

THOSE NUMBERED WITH  ROMAN NUMERALS ARE FOR OTHER ELEMENTS, SUCH AS
TITLES.

FROM WHAT  I HAVE JUST SAID, WHAT MIGHT THIS BE?

   1.   ROME--HISTORY--REPUBLIC, 265-30 B.C.

.ANOTHER ELEMENT                         PERIOD  INDICATES  STUDENT  INPUT

TRY AGAIN.  THE COMPLETE TERM.

.SUBJECT ENTRY

WHAT KIND OF ENTRY AND WHAT KIND OF DEVICE?  TRY AGAIN.

.SUBJECT ADDED ENTRY

A SUBJECT ADDED ENTRY TRACING.  YOU COULD TELL BY THE  ARABIC NUMERAL
EVEN IF IT WEREN'T CLEAR FROM THE HEADING ITSELF.


I WANT YOU TO TRY IDENTIFYING THE TYPES OF TRACING LISTED BELOW, FROM
EXAMPLES WHICH  I WILL GIVE YOU IN A MOMENT

          TITLE ADDED ENTRY
          JOINT AUTHOR ADDED ENTRY
          SERIES ADDED ENTRY
          SUBJECT ADDED ENTRY
          TRANSLATOR ADDED ENTRY

READY?  TAKE YOUR TIME.  WHENEVER YOU WANT TO LOOK AT THE LIST, JUST
TYPE THE WORDS " REPEAT LIST."

     SUPPOSE WE HAVE A BOOK ENTITLED " THE  REPUBLIC OF  ROME," BY
JOINT AUTHORS  RENE  L.  PAGE AND  JEAN  LAVAL.  THE FORMER HAS BEEN
SELECTED AS THE PRINCIPAL AUTHOR FOR PURPOSES OF MAIN ENTRY.  THE BOOK
IS A TRANSLATION FROM THE  FRENCH BY  A. B.  SEA.  IT IS PART OF THE "
TEMPS  PERDUS  SERIES".

WHAT IS THIS TRACING:    1.   ROME--HISTORY--REPUBLIC, 265-30 B.C.

.SUBJECT

TRY AGAIN.  THREE WORDS.

.SUBJECT ENTRY TRACING

THAT WAS A  SUBJECT ADDED ENTRY.

HOW ABOUT:                 2. ROME--HISTORY--CIVIL  WAR, 49-48 B.C.?

.SAME AS ABOVE

GOOD.

HOW ABOUT:        3. ROME--OFFICIALS AND EMPLOYEES

.DITTO

FINE.  NOW WHAT IS THIS?      II. LAVAL,  JEAN, 1930-      , JT. AUTH.

.JOI

+++ TYPE AGAIN +++

.AUTHOR ENTRY
```

```
 TRY AGAIN.  ALL FOUR WORDS, PLEASE.

.JOINT AUTHOR MAIN ENTRY

 THAT ONE WAS A  JOINT AUTHOR ADDED ENTRY.

 TRY                        I.   SEA,  A. B., 1940-     , TR.

.LET ME CHECK THE LIST AGAIN


          TITLE ADDED ENTRY
          JOINT AUTHOR ADDED ENTRY
          SERIES ADDED ENTRY
          SUBJECT ADDED ENTRY
          TRANSLATOR ADDED ENTRY

.OK, A.B. SEA LOOKS LIKE A TRANSLATOR ADDED ENTRY
.

 YOU'RE DOING BEAUTIFULLY.  HERE'S ANOTHER:    IV.   THE  REPUBLIC OF
ROME

.TITLE ADDED

+++ TYPE AGAIN +++

.SUBJECT ADDED ENTRY

 TRY AGAIN.  THIS ONE IS SO-O-O-O EASY!

.TITLE ENTRY?

 T I T L E ADDED ENTRY.  NON FELIX SUM!

 WELL, MAYBE YOU CAN REDEEM YOURSELF WITH THIS FINAL ONE.
          III.   TEMPS  PERDUS  SERIES.

.SERIES ADDED ENTRY

 VERY GOOD INDEED.  THAT COVERS MOST OF THE TYPES OF ADDED ENTRY YOU
ARE LIKELY TO ENCOUNTER, WITH THE EXCEPTION OF  ILLUSTRATOR ADDED
ENTRIES, WHICH ARE FAIRLY RARE BECAUSE THEY ARE USED ONLY WHEN THE
IDENTITY OF THE ILLUSTRATOR IS OF ESPECIAL INTEREST.
```

BASIC SOURCE PROGRAM FOR ABOVE

```
KANDINSKY:  T :
              /;
            T : THERE IS A CODE IN EFFECT FOR TRACINGS:
              /
              / THOSE NUMBERED WITH  ARABIC NUMERALS ARE FOR SUBJECT ADDED ENTRIES.
              /
              / THOSE NUMBERED WITH  ROMAN NUMERALS ARE FOR OTHER ELEMENTS, SUCH AS
               TITLES.;
            T :
              /;
            T : FROM WHAT  I HAVE JUST SAID, WHAT MIGHT THIS BE?
              /
              /     1.   ROME--HISTORY--REPUBLIC, 265-30 B.C.;
```

```
MECHIKKU:   A :;

            R :DEBUG;
            C :GJUMP TO AUTHOR;

            R :SUBJECT ADDED ENTRY,  SUBJECT ADDED ENTRY, TRACING SUBJECT, SUBJECT
               TRACING,  TRACING SUBJECT,  SUBJECT TRACING;
            G : A SUBJECT ADDED GNTRY TRACING.  YOU COULD TELL BY THE  ARABIC NUMERAL
               EVEN IF IT WEREN'T CLEAR FROM THE HEADING ITSELF.;
            C :GJUMP TO ZZ538;

            R : TRACING, TRACING;
            G : YES, BUT FOR WHAT KIND OF ENTRY?;
            C :GJUMP TO MECHIKKU;

            R : ENTRY, ENTRY;
            G : WHAT KIND OF ENTRY AND WHAT KIND OF DEVICE?  TRY AGAIN.;
            C :GJUMP TO MECHIKKU;

            R : TITLE, TITLE,  T I T L E;
            G : NO.  IT CCULDN'T POSSIBLY BE A TRACING FOR A  T I T L E ADDED ENTRY.
               THE  ARABIC NUMERAL TELLS YOU THIS.  TRY AGAIN.;
            C :GMARK STALLY,
               GJUMP TO MECHIKKU;
            C :USE (Z535, Z536, Z537) ON (ZAP),
               TALLY ZAP,
               MARK (A..NILNUM, A..NILNUM, A..NILNUM, STALLY) ON (ZAP),
               JUMP TO (A..NILLOC, MECHIKKU, MECHIKKU, ZZ538) ON (ZAP);
Z535:       T : TRY AGAIN.  THE CCMPLETE TERM.;
Z536:       T : DID YOU FORGET "ADDED" ENTRY?  TRY AGAIN, ANYWAY.;
Z537:       T : THE ANSWER IS " TRACING OF (CR FOR) A SUBJECT ADDED ENTRY," OR SIMPLY
               " SUBJECT ADDED ENTRY TRACING.";
ZZ538:      C :MARK ZAP;
            N :ZZ538;


            C :IF STALLY
                  THEN ADD 2 TO LVCLLY,
               CLEAR STALLY;
BENET:      T :
               /;
            T : I WANT YOU TO TRY IDENTIFYING THE TYPES OF TRACING LISTED BELOW, FROM
               EXAMPLES WHICH  I WILL GIVE YOU IN A MOMENT;
TRACINGS:   T :
               /              TITLE ADDED ENTRY
               /              JOINT AUTHOR ADDED ENTRY
               /              SERIES ADDED ENTRY
               /              SUBJECT ADDED ENTRY
               /              TRANSLATOR ADDED ENTRY;
            T :
               / READY?  TAKE YOUR TIME.  WHENEVER YOU WANT TO LOOK AT THE LIST, JUST
               TYPE THE WORDS " REPEAT LIST.";
            T :
               /     SUPPOSE WE HAVE A BOOK ENTITLED " THE  REPUBLIC OF  ROME," BY
               JOINT AUTHORS  RENE  L.  PAGE AND  JEAN  LAVAL.  THE FORMER HAS BEEN
               SELECTED AS THE PRINCIPAL AUTHOR FOR PURPOSES OF MAIN ENTRY.  THE BOOK
               IS A TRANSLATION FROM THE  FRENCH BY  A. B.  SEA.  IT IS PART OF THE "
               TEMPS  PERDUS  SERIES".;
            T :
               /;
            T : WHAT IS THIS TRACING:     1.    ROME--HISTORY--REPUBLIC, 265-30 B.C.;

BENEZET:    A :;

            R :DEBUG;
            C :GJUMP TO AUTHOR;

            R : SUBJECT ADDED ENTRY, SUBJECT ADDED ENTRY;
            G : GOOD.  HOW ABOUT:        2.   ROME--HISTORY--CIVIL  WAR 49-48 B.C.?;
            C :GJUMP TO ZZ540;
```

```
           R :LIST;
           C :GUSE TRACINGS,
              GJUMP TO BENEZET;
           C :USE (Z539, Z54C) ON (ZAP),
              TALLY ZAP,
              TALLY (A..NILNUM, A..NILNUM, LVOLLY) ON (ZAP),
              JUMP TO (A..NILLOC, BENEZET, ZZ540) ON (ZAP);
Z539:      T : TRY AGAIN.  THREE WORDS.;
Z54(:      T : THAT WAS A SUBJECT ADDED ENTRY.
              /
              / HOW ABOUT:                  2. ROME--HISTORY--CIVIL  WAR, 49-48 B.C.?;
ZZ540:     C :MARK ZAP;
           N :ZZ540;

PENGURION: A :;

           R :DEBUG;

           C :GJUMP TO AUTFOR;

           R :SUBJECT ADDED ENTRY,SAMF, SAMF, LIKEWISF,LIKEWISE,DITTO, DITTO,$";
           G : GOOD.
              /
              / HOW ABOUT:        3. ROME--OFFICIALS AND EMPLOYEES;
           C :GJUMP TO ZZ542;

           R :LIST;
           C :GUSF TRACINGS,
              GJUMP TO BENGURION;
           C :USE (Z541, Z542) ON (ZAP),
              TALLY ZAP,
              TALLY (A..NILNUM, A..NILNUM, LVOLLY) ON (ZAP),
              JUMP TO (A..NILLOC, BENGURION, ZZ542) ON (ZAP);
Z541:      T : TRY AGAIN.  THREE WORDS.;
Z542:      T : THAT WAS ALSO A SUBJECT ADDED ENTRY.
              / TRY THIS ONE:          3.    ROME--OFFICIALS AND EMPLOYEES;
ZZ542:     C :MARK ZAP;
           N :ZZ542;

BENCIT:    A :;

           R :DEBUG;
           C :GJUMP TO AUTHOR;

           R :SUBJECT ADDED ENTRY,SAMF, SAMF, LIKEWISF,LIKFWISE,DITTO, DITTO,$";
           G : FINE.  NOW WHAT IS THIS?       II. LAVAL,  JEAN, 1930-     , JT.
              AUTH.;
           C :GJUMP TO ZZ544;

           R :LIST;
           C :GUSE TRACINGS,
              GJUMP TO BENCIT;
           C :USF (7543, 7544) ON (ZAP),
              TALLY ZAP,
              TALLY (A..NILNUM, A..NILNUM, LVOLLY) ON (ZAP),
              JUMP TO (A..NILLCC, BENOIT, ZZ544) ON (ZAP);
7543:      T : TRY AGAIN.  ALL THREE WORDS, PLEASE.;
7544:      T : THAT, AGAIN, WAS A SUBJECT ADDED ENTRY.
              /
              / WHAT ABOUT                   II.  LAVAL,  JEAN, 1930-      , JT. AUTH.;
ZZ544:     C :MARK ZAP;
           N :ZZ544;

BENSON:    A :;

           R :DEBUG;
           C :GJUMP TO AUTHOP;

           R : JOINT AUTHOR ADDED ENTRY, JOINT AUTHOR ADDED ENTRY;
           G : YES.  NOW WHAT IS THIS?       I.   SEA,  A. B., 1940-      , TR.;
```

```
            C :GJUMP TO ZZ546;

            R :LIST;
            C :GUSE TRACINGS,
               GJUMP TO BENSON;
            C :USE (Z545, Z546) ON (ZAP),
               TALLY ZAP,
               TALLY (A..NILNUM,. A..NILNUM, LVOLLY) ON (ZAP),
               JUMP TO (A..NILLCC, BENSON, ZZ546) ON (ZAP);
Z545:       T : TRY AGAIN.  ALL FOUR WORDS, PLEASE.;
Z546:       T : THAT ONE WAS A  JOINT  AUTHOR  ADDED  ENTRY.
              /
              / TRY                          I.   SEA, A. B., 1940-      , TR.;
ZZ546:      C :MARK ZAP;
            N :ZZ546;

BENTLINK:   A :;

            R :DEBUG;
            C :GJUMP TO AUTHOR;

            R : TRANSLATOR ADDEC ENTRY, TRANSLATOR ADDED ENTRY;
            G : YOU'RE DOING BEAUTIFULLY.  HERE'S ANOTHER:    IV.   THE  REPUBLIC OF
                ROME;
            C :GJUMP TO ZZ548;

            R :LIST;
            C :GUSE TRACINGS,
               GJUMP TO BENTLINK;

            R :ADDED ENTRY, ADDED ENT*;
            G : WHAT KIND OF ADDEC ENTRY?  WHAT DOES THE "TR." MEAN?  PLEASE REPEAT.
                (THREE WORDS.);
            C :GJUMP TO BENTLINK;
            C :USE (Z547, Z548) ON (ZAP),
               TALLY ZAP,
               TALLY (A..NILNUM, A..NILNUM, LVOLLY) ON (ZAP),
               JUMP TO (A..NILLOC, BENTLINK, ZZ548) ON (ZAP);
Z547:       T : TRY AGAIN.;
Z548:       T : YOU SHOULD HAVE ANSWERED " TRANSLATOR ADDED ENTRY."
              /
              / TRY ANOTHER:                    IV.    THE   REPUBLIC OF   ROME.;
ZZ548:      C :MARK ZAP;
            N :ZZ548;

BENTLEY:    A :;

            R :DEBUG;
            C :GJUMP TO AUTHOR;

            R : TITLE ADDED ENTRY, TITLE ADDED ENTRY ;
            G : KEEP IT UP  ONE MORE:        III.   TEMPS   PERDUS   SERIES.;
```

```
             C :GJUMP TO ZZ550;

             R :LIST;
             C :GUSE TRACINGS,
                GJUMP TO BENTLEY;
             C :USE (Z549, Z550) ON (ZAP),
                TALLY ZAP,
                TALLY (A..NILNUM, A..NILNUM, LVOLLY) ON (ZAP),
                JUMP TO (A..NILLCC, BENTLEY, Z7550) ON (ZAP);
Z549:        T : TRY AGAIN.  THIS CNE IS SO-O-O-O EASY ;
Z550:        T : T I T L F ADDFD ENTRY.  NCN FFLIX SUM
                /
                / WELL, MAYRE YCU CAN REDEFM YCURSELF WITH THIS FINAL ONE:
                         III.    TEMPS  PFRDUS  SFRIES.;
Z7550:       C :MARK ZAP;
             N :ZZ550;

RENTON:      A :;

             R :DEBUG;
             C :GJUMP TO AUTHOR;

             R : SERIES ADDED ENTRY, SFRIES ADDED ENTRY;
             G : VERY GOOD INDEEC.  THAT COVERS MOST OF THE TYPES OF ADDED ENTRY YOU
                ARE LIKELY TO ENCOUNTER, WITH THE EXCEPTION OF  ILLUSTRATOR ADDED
                ENTRIES, WHICH ARE FAIRLY RARE BECAUSE THEY ARE USED ONLY WHEN THE
                IDENTITY OF THE ILLUSTRATOR IS OF ESPECIAL INTEREST.;
             C :GJUMP TO ZZ553;

             R :LIST;
             C :GUSF TRACINGS,
                GJUMP TO RENTON;
             C :USF (Z551, Z552) ON (ZAP),
                TALLY ZAP,
                TALLY (A..NILNUM, A..NILNUM, LVOLLY) ON (ZAP),
                JUMP TO (A..NILLCC, BENTON, ZZ553) ON (ZAP);
Z551:        T : TRY AGAIN.;
Z552:        T : THAT, "NAME", WAS A  SERIES ADDED ENTRY.;
ZZ553:       C :MARK ZAP;
             N :ZZ553;
BERENGARIA:  T :
                /;
             T : WHAT IF YOU FIND NO TRACINGS AT THE BOTTOM OF A CARD?
                /
                / IT MEANS THAT THERE ARE (VERY FEW) (3) (NO) (INDESCRIBABLE)
                (NON-DISTINCTIVE) ADDED ENTRIES FOR THAT PARTICULAR BOOK.;

             A :;

             R :"NEGAT";
             G : CORRECT.  IT HAPPENS.;
             C :GJUMP TO ZZ555;
```