NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO, CALIFORNIA 92152

NPRDC TR 76-9

AUGUST 1975

FEASIBILITY OF AND DESIGN PARAMETERS FOR A COMPUTER-BASED ATTITUDINAL RESEARCH INFORMATION SYSTEM

Diane M. Ramsey-Klee
Vivian Richman
Gio Wiederhold

NPRDC TR 76-9 August 1975

FEASIBILITY OF AND DESIGN PARAMETERS FOR A COMPUTER-BASED ATTITUDINAL RESEARCH INFORMATION SYSTEM

Diane M. Ramsey-Klee
Vivian Richman
Gio Wiederhold
R-K RESEARCH AND SYSTEM DESIGN
3947 Ridgemont Drive
Malibu, California 90265

This research was sponsored jointly by the

Organizational Effectiveness Research Programs
Psychological Sciences Division
Office of Naval Research
Arlington, Virginia 22217

and the

Navy Personnel Research and Development Center
San Diego, California 92152

Contract N00014-74-C-0396
Contract Authority Identification Number, NR 170-769

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPRDC TR 76-9	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Feasibility of and Design Parameters for a Computer-Based Attitudinal Research Information System		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Diane M. Ramsey-Klee Vivian Richman Gio Wiederhold		8. CONTRACT OR GRANT NUMBER(s) N00014-74-C-0396
9. PERFORMING ORGANIZATION NAME AND ADDRESS R-K Research and System Design 3947 Ridgemont Drive Malibu, California 90265		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 62763N NR 170-769
11. CONTROLLING OFFICE NAME AND ADDRESS Organizational Effectiveness Research Programs Office of Naval Research (Code 452) Arlington, Virginia 22217		12. REPORT DATE August 1975 13. NUMBER OF PAGES 237
14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)* Navy Personnel Research and Development Center San Diego, California 92152		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Attitudinal Research Information System
Cataloguing of Information
Cost-Benefit Analysis
Data Archives, Social Science
Data Banks, Social Science

(Continued)

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

(U) This technical report presents the findings of a study of the feasibility of developing a computer-based attitudinal research information system (RIS) for the field of Navy personnel research. The requirement for an RIS is generated from the fact that most data bases for Navy personnel and attitude research are not now retrievable and usable for secondary analysis or trend analysis. In addition to making an assessment of overall feasibility,

(Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Block 19 (Cont.) KEY WORDS

Data Base Design
Data Base Management
Indexing of Information
Information Storage and Retrieval Techniques
Interface between Users and an Information System
On-line Information System
Social Science Data Archives or Data Banks
System Design

Block 20 (Cont.) ABSTRACT

this report contains recommendations regarding how such a system should be designed, implemented, and administered. Literature reviews in several topic areas are presented, including information science, computer methodology, human factors considerations in on-line system design, and cost-benefit analysis. The audience of readers potentially interested in this report would be librarians, information scientists, and indexers; system designers, system analysts, data processing personnel, and computer programmers; economists; and managers of data processing facilities and research activities.

Specifically, the following subjects are addressed in this report: information needs and requirements of Navy personnel researchers and managers, information indexing alternatives, data base design alternatives, the operational interface between users and a computer-based information system, system design and implementation requirements for an attitudinal RIS, and cost-benefit considerations. A selected listing and description of social science data archives is included as well as a partial inventory of data bases that are candidates for inclusion in an attitudinal RIS. The conclusions and recommendations of the feasibility study team complete this report.

FOREWORD

This research and development was performed in support of the Navy Personnel Research and Development Center Attitude and Motivation Research Program and the Office of Naval Research Psychological Sciences Program.

The technical monitor for this contract was Dr. Laurie Broedling of this Center who helped conceptualize the simulation of the hypothetical attitudinal research information system (RIS) and evaluate the substantive content of the report.

Appreciation is expressed to Doctors John Nagay, Bert T. King, and Eugene E. Gloye of the Office of Naval Research for their scientific and administrative assistance, and to Ms. Ingeborg M. Kuhn, Graduate School of Business, Stanford University, for her contributions to the section on cost-benefit considerations.

J. J. CLARKIN
Commanding Officer

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SECTION 1. OBJECTIVES AND OVERVIEW

In May of 1974 R-K Research and System Design embarked on a 1-year study of the feasibility of establishing a computer-based attitudinal research information system (RIS) to facilitate more efficient use of personnel survey data within the naval personnel research community. The decision to conduct this feasibility study grew out of an awareness of the need for designing and developing a computer tool to assist in introducing a more programmatic approach to Navy personnel planning and program policy making in which information gathering and analysis could be coordinated within a single system. Since the bulk of the in-house research in this domain is conducted by the Navy Personnel Research and Development Center (NPRDC), the feasibility study focused on the information needs and requirements of this institution. However, other potential users of an attitudinal RIS also were kept in mind. It is expected that an RIS would increase the efficiency of this area of activity by standardizing data collection, data documentation, and storage and retrieval procedures; by making complex research designs and analytic techniques readily available to individual personnel researchers; and by building upon information already available from Navy personnel files and from past attitudinal and opinion studies as well as from previous research on organizational behavior. The ultimate scope of an attitudinal RIS might extend beyond the data base holdings of the Navy personnel research community to include data bases from other Navy activities and from Navy contractors that would be of use in personnel research.

Figure 1 depicts one model of the primary and secondary research process as it could be expanded by the Navy personnel research community. The left-hand side of this figure portrays the *primary research* process. An operational problem is identified; a research design to help solve the problem is devised; the sampling methodology is defined; the primary data are collected and then analyzed; the results and findings are interpreted; and recommendations based on the research findings are submitted to Navy personnel planners and decision makers, usually in the form of a research report. Presently, within the Navy personnel research community, the functional unit for the collection, analysis, storage, and retrieval of data on naval personnel is the individual study. This independence of effort results in a certain amount of redundancy of the information gathered, some duplication of effort, and a lack of coordination and standardization. The fleet has been surveyed repeatedly in individual studies without any really effective method for coordinating these investigations and placing them in the larger context of an overall personnel research program. The global problem, then, is that no programmatic, systematic, or economically efficient approach to information gathering and exploitation presently exists. A computer-based attitudinal RIS represents an attractive proposition for helping to answer this need.

The right-hand side of Figure 1 portrays the *secondary research* process, a way of going about research that has been largely ignored in the past, but which is gaining in popularity and acceptance. A great deal of time and money is invested in collecting the primary data required for conducting an individual research study. However, once such a project is completed, the data base in the form of questionnaire responses, behavioral observations, punched

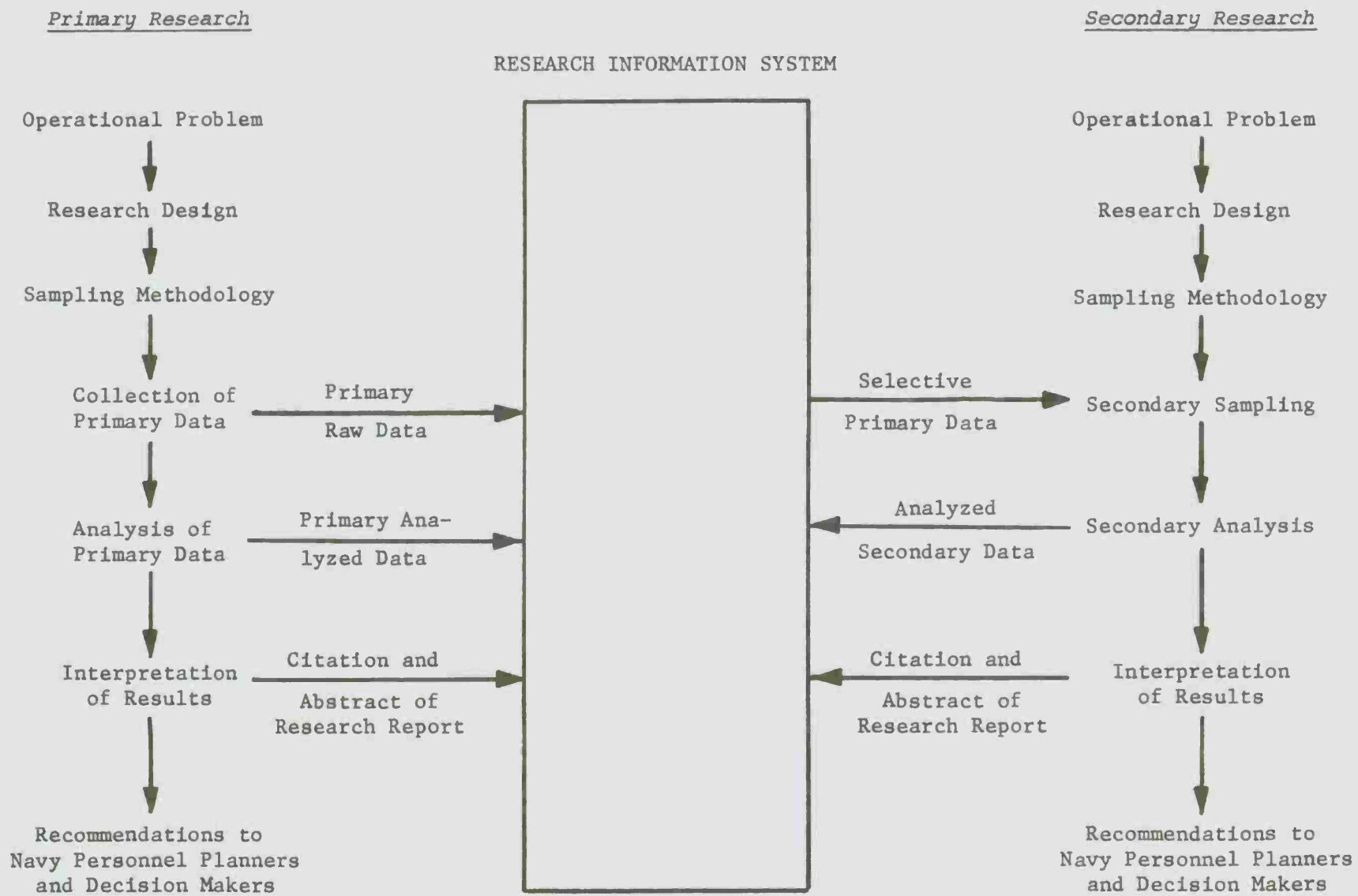


Figure 1. Model of the Primary and Secondary Research Process.

cards, or magnetic tape frequently is placed on the shelf, never to be used again. Such a data base may contain valuable information that could be utilized by another investigator in a later study. This, then, is the fabric of secondary research---use of primary data bases collected in earlier research by a second study conducted at a later point in time.

Referring again to the right-hand side of Figure 1, one can see that secondary research proceeds in much the same manner that primary research does, but with one important difference. As with primary research, secondary research begins with the identification of an operational problem followed by the devising of a research design to shed some light on the problem. In parallel with primary research, a sampling methodology is defined, but in the case of secondary research the data to be sampled reside in the primary data bases collected previously. The sampling methodology yields selective primary data to be used in the secondary analysis. The results and findings of this secondary analysis are interpreted, and as with primary research results, recommendations based on the research findings are submitted to Navy personnel planners and decision makers. However, secondary research rarely is carried out in the Navy setting because there is no defined repository for primary data bases, permitting convenient and easy access. Additionally, there is no catalog indicating which primary data bases exist, and documentation of the contents of existing data bases typically is lacking or inadequate. A computer-based attitudinal RIS could provide an important vehicle to help bridge this gap.

As a starting point for this feasibility study, two sources of information were tapped to provide a sound foundation for the work to follow. First, a survey was conducted of social science data archives now in existence to better understand how other research-oriented institutions were addressing the problems of information management. The results of this survey are contained in Section 2 and in Appendix A. Second, a series of interviews were conducted with NPRDC researchers, support staff, and managers in an attempt to identify and specify their information needs and requirements. Although the focus was on this institution, interviews also were conducted with individuals in other Navy activities and with representatives from other services, individuals responsible for generating and maintaining data bases of use in personnel research. Two products resulted from these interviews. An inventory was prepared summarizing the data bases known to be in existence at NPRDC that are candidates for inclusion in an attitudinal RIS. Not all of the data bases extant at NPRDC were identified, but the ones that were located are important, visible ones as well as representative of past, current, and future holdings. This partial inventory is presented in Appendix B.

The second product to result from the interviews with NPRDC personnel was a simulation of a hypothetical attitudinal RIS. The objective of this effort was to embody some tentative design concepts and principles in a simulated model of an interactive information system in order to determine which aspects of the model would receive a favorable reaction when viewed by potential users. The simulation was carried out by means of a cathode-ray terminal connected to a tape deck and a peripheral printer. The contents of the demonstration of the hypothetical RIS were predetermined and stored on magnetic tape cassettes from which information could be called up for display on the

video screen of the terminal. The option to print a hard copy of any display viewed on the terminal also was available. Every effort was made to make the simulation realistic, information rich, and responsive to the information needs and requirements that had been expressed during the series of interviews. A demonstration of the simulated system took approximately 50 minutes and was given 12 times to small groups of viewers from NPRDC and the Navy Health Research Center (5 to 10 individuals at each session). Another presentation of the simulated system was made to members of the Manpower Research and Development Committee at the Office of Naval Research in Arlington, Virginia using viewgraphs of the CRT displays. Feedback from these sessions helped clarify which aspects of the simulation were considered useful in facilitating information retrieval, analysis, and synthesis and which aspects were considered less responsive or superfluous. Thus, the information needs and requirements of the Navy personnel research community were more sharply defined, and served as the basis for developing the technical aspects of system implementation, specifying system design parameters, and identifying economic considerations. Section 3 contains a discussion of the information needs and requirements of Navy personnel researchers and managers as identified in the series of interviews that were conducted and as gleaned from the comments of individuals viewing the simulation of the hypothetical attitudinal RIS. Selected displays from the demonstration of the hypothetical system are included both in Section 3 and in Section 6 where variables that enhance the operational interface between users and a computer-based information system are discussed.

Sections 4, 5, and 6 of this report were prepared with two purposes in mind. First, it was considered important that any future implementation of either a manual or a computer-based attitudinal RIS should have a strong theoretical base. Second, a theoretical discussion of the variables and alternatives involved in information indexing, data base design, and user-computer interface should elucidate the key issues and trade-offs, and form the basis and justification for arriving at a definition of system design and implementation requirements. Consequently, in Section 4 the reader will find a fundamental discussion of information indexing alternatives. This section should be of particular interest and value to librarians, information scientists, and indexers. Section 5 embodies a basic discussion of data base design alternatives, and is directed primarily to system designers, system analysts, data processing personnel, and computer programmers.

In Section 6 an analysis of the factors likely to ensure an optimal interface between users and a computer-based information system is presented. Types of interfaces are discussed as well as training approaches that should be considered to minimize the effort required to learn how to use a system. Section 6 also includes a delineation of the variables important in determining user acceptance, and incorporates pertinent displays from the simulated attitudinal RIS that illustrate certain desirable features of an interactive system. This section should be of particular interest to those individuals concerned with maximizing the utilization of computer terminals and other peripheral devices and with taking advantage of the potential of computer-assisted instructional techniques for teaching users, via a terminal, how to interact effectively with a computer-based information system. Sections 4, 5, and 6 all include a summary for those readers interested only in a recapitu-

lation of the major conceptual issues discussed. All three of these theoretical sections also include an annotated bibliography and a glossary of key terms to aid the reader in assimilating and applying the information content.

The technical core of this feasibility study is contained in Section 7 where the system design and implementation requirements for an attitudinal RIS are delineated. Aspects considered in this section are information indexing requirements, file organization and estimated volume, file storage requirements, communication requirements, projected user load, processing required for disk and computational references, system and programming support facilities required, and operating system facilities required. System specifications are provided, and three possible computer hardware and software configurations fulfilling these requirements are enumerated. The manpower required to operate and maintain the system also is projected. This section along with Sections 4, 5, and 6 provide indispensable information for whomever might be delegated to conduct the implementation of an attitudinal RIS, whether it were to be done in-house or by outside contractor. The question of the technical feasibility of implementing such an information system is answered in Section 7; the economic feasibility is considered in Section 8.

In Section 8 cost-benefit considerations are reviewed as they might apply to the possible implementation of an attitudinal RIS. The identification and measurement of cost categories is considered as well as the much more difficult task of identifying anticipated benefits and placing a value on these benefits. An analytical model for cost-benefit analysis in this context is offered, and alternatives for establishing a charge structure for services provided by an attitudinal RIS are discussed.

The final section of this report, Section 9, presents the conclusions and recommendations of the interdisciplinary team that conducted this feasibility study. This last section together with Sections 1, 3, 7, and 8 present the major sources of information resulting from this study on which a management decision can be based regarding the desirability, efficacy, and economic suitability of implementing an attitudinal RIS.

SECTION 2. A SURVEY OF SOCIAL SCIENCE DATA ARCHIVES

This brief survey is included in this report for the purpose of providing both an historical perspective of the development of and a description of the current state of the art of data archives in the social sciences. The survey was made to better understand how other research-oriented institutions are addressing the problem of information management so that any useful approaches could be considered in the design of an attitudinal research information system.

At the beginning of this century there was a shift of emphasis away from the individual and from the single historical event to an awareness of mass attitudes and actions. There was a realization that small, slowly accumulating changes account for much historical and political movement. However, the data needed for this type of historical interpretation were too plentiful for precomputer analysis. Historians were overwhelmed by the prospect of analyzing the volume of information collected and recorded on various social phenomena, such as the voting records of citizens and their representatives, economic statistics, birth and death records, social and economic mobility, education, union membership, and religious affiliation.

The Hollerith card became the key to the data processing activities which made possible the management of large quantities of accumulated data. The 80-column punched card was a technological response to the spiraling needs of the U.S. Census, which had collected information for a decade on the country's expanding population. These data were too voluminous for manual analysis. From this early method of data input to electrical accounting machines, the evolution to the computer has been of great assistance in data processing where the magnitude of data has prevented manual operations.

The creation of the social science data bank or data archive was made possible by the capabilities of computing technology. Computers readily accept textual and numerical input, can perform a variety of statistical operations, and then produce analyzed output. The social science data bank or archive is a depository of data that social scientists, policy-makers, and others may use for teaching, research, decision-making, and other purposes. The U.S. Government probably produces the most data useful for studying social problems, and it is probably the largest user of these data. The data archive developed out of the social scientist's needs to exploit the masses of governmental, academic, and commercial data accumulated over the last three decades. These archives are necessary for researchers and policy makers to be able to compare current research findings with previous findings. The archive can provide a set of comparable but different data collections over time that are needed for trend analysis. The archive also allows data collections to benefit researchers other than the original data collector. Data collecting is a very expensive operation, and there is a great economic benefit if these data collections are made accessible to researchers who can use the collections for other analytic purposes such as in secondary analysis.

The function of the social science data archive is the efficient management of files of collected information or data so that these data are available and useful for social scientists. The data stored may be statistical

data, computer-usable collections prepared by social research organizations, historical and administrative records, published texts, reports, and even speeches. In contrast to a data bank, the data stored in a bibliographic retrieval system consist of references to documents. After examining references retrieved in response to a search request, a user then must decide which references appear to be pertinent to his information needs and the source documents must be located. Some data archives do include bibliographic search services as a supplement, but this is not their primary function. The primary purpose of the data bank is to collect, organize, and maintain raw or cleaned data for analysis by the user community that it serves.

Initially, the archive must *acquire information* which includes locating relevant data. Second, the information must be processed in preparation for distribution and analysis. The *creation of the data base* includes transforming the data to facilitate computer analysis, converting data to standard coding schemes to allow for comparisons and analysis, and locating and correcting errors in the data and the documentation. Thirdly, *information maintenance* insures that the data are up-to-date and accurate and prevents the permanent loss of information by preserving extra copies of data and documentation. Finally, the archive must *provide services to users*. These services may be in the form of supplying duplicates of and/or subsets from data sets, furnishing computer analysis of data, and/or providing consultations in mathematics, statistics, research methodology, and programming. One service to social scientists made possible by the advent of the computer is the development of packaged programs for standard statistical analysis. Three major statistical packages are the following: Biomedical Computer Programs (BMD), Statistical Programs for Social Science (SPSS), and Organized Set of Integrated Routines for Investigations with Statistics (OSIRIS).

There are three types of archival organizations. The first type provides the user with "clean" data, enabling the social scientist undertaking secondary analysis to immediately begin his research with the available data rather than suffer the long delays that data "cleaning" involves. The archive staff cleans the data acquired by determining whether data, codebooks, and ancillary information agree. Cleaning is a complex operation involving extended communication by mail and telephone with the data collection organization. The resulting data are of high quality, and the user can be confident that there will be few discrepancies between data and documentation and that the documentation will be understandable. The Inter-University Consortium for Political Research at the University of Michigan provides the user with clean data. However, the user must accept the philosophy and values of the archive preparing the data and documentation. Disadvantages are the expense of the data preparation, and the delay in the availability of data from the archive.

The second type of archival organization requires that any data and documentation transformations must be controlled by the user. The rationale is that it is difficult to predict the usage of some data collections, and therefore, the expense of pre-use-cleaning cannot be justified. The user is not constrained by an archive's philosophy about cleaning and data preparation. However, the user has the burden of determining discrepancies between data and documentation and locating inconsistencies within the data. Another disadvantage is that the data-cleaning task may be duplicated by other users.

The third alternative to data archiving is a combination of the above two approaches. There is some kind of human interface between the data source and the user who supplies the information required for cleaning data and handling discrepancies. However, the secondary user has considerable control over the processing and analysis of the available data. These users have to rely on the skills of the archive's personnel and may have to become familiar with several different computer programs and procedures. The archive specialist then organizes the resulting information for the future benefit of other potential users. This approach is used by the Data and Program Library Service of the Social Science Data and Computation Center at the University of Wisconsin and by the Social Science Information Center at the University of Pittsburgh.

Archives can be characterized by the scope of their data collections and by the geographical range of their services. Data collections may be highly specialized or they may be more general; they may serve only their local geographic area or even supply service worldwide. Archives that specialize in local services tend to respond better to the diverse needs of their users, while national organizations can only respond in a more general way since they do not appreciate the particular requirements of local environments. Archives with high usage rates reflect an attention to the varying needs of their users. The Inter-University Consortium for Political Research owes its success to involving its users in its data-acquisition policies. This institution also offers training programs for its users and potential users.

Appendix A provides a selected listing and description of a large number of social science data archives in the United States. This listing provides a survey of the types of data storage collections that are already in existence and the kinds of user communities associated with each. The more important social science archives that exist in other countries also have been included. The U.S. Census currently uses computer technology, and investigators can access the magnetic tapes of the various compilations through institutions around the country. Both private and university organizations offer census services, such as display of printouts of census tapes, analysis, and consultation. Since there are a growing number of these institutions, only a few are listed in Appendix A. Other valuable data in computer-processable form, such as public opinion polls, are also accessible to investigators. Most of the records compiled by government agencies have been turned into a data bank as an aid to further study. These records are available at a central repository, the Inter-University Consortium for Political Research, which houses and distributes these files.

ANNOTATED BIBLIOGRAPHY

1. Bisco, R. L. (Ed.). *Data Bases, Computers, and the Social Sciences.* New York: Wiley-Interscience, 1970.

   This book is a product of the Fourth Annual Conference of the Council of Social Science Data Archives. It includes post-edited versions of some of the papers that were presented concerning the interrelations between data banks, computer technology, and the needs of the social sciences.

2. *Data Access News.* Arlington, Virginia: Clearinghouse and Laboratory for Census Data (CLCD), (published six to eight times per year).

   CLCD is operated by Data Use and Access Laboratories (DUAL) with a grant from the National Science Foundation. *Data Access News* provides news of interest to users of statistics, and provides advance notice on technical matters to be covered in other DUALabs' publications.

3. *Encyclopedia of Information Systems and Services* (2nd international ed.). (Ann Arbor, Michigan?): A. T. Kruzas Associates, 1974.

   This book describes and analyzes approximately 1,750 organizations that provide information services based on storage and representation of structured information with output on a recurring or demand basis.

4. Hyman, H. H. *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities.* New York: Wiley, 1972.

   This book contains convenient lists of research designs for secondary analysis of survey data, problems amenable to study, archives, and other sources of data. It includes detailed case studies of the research process and of discovery and productivity in secondary analysis. In addition, there is an exhaustive bibliography and a detailed, analytic table of contents for easy reference.

5. Ruben, J., & Widmann, R. L. Information systems applications in the humanities. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 7). Washington, D.C.: American Society for Information Science, 1972. Pp. 439-469.

   This article is on the aid computers have lent to the humanities, with an emphasis on historical trends and analysis.

6. *S S Data.* Iowa City, Iowa: Laboratory of Political Research, The University of Iowa, (published quarterly).

   This newsletter of Social Science Archival Acquisitions, published quarterly, is a means of communicating information on the acquisitions of data archives to social science researchers.

7. *Social Science Data Archives in the United States.* New York: Council of Social Science Data Archives, 1967.

   This book lists and describes 25 social science data archives operating in the U.S. as of 1967.

SECTION 3. INFORMATION NEEDS AND REQUIREMENTS OF NAVY PERSONNEL RESEARCHERS AND MANAGERS

Introduction

In order to make judgments regarding system design goals, it was necessary to identify the information needs and requirements of those individuals responsible for conducting Navy attitudinal research. There are several groups of such individuals---in-house research personnel, research contractors, and in-house research administrators. Since the absolute number of individuals in these three groups is fairly large, it was decided to focus on those organizational units responsible for the majority of the attitudinal research. This emphasis led to inclusion of researchers, support staff, and managers within NPRDC; researchers and support personnel at the Navy Health Research Center (NHRC); ONR scientific officers and managers; individuals who might be involved in specific ways in RIS development (e.g., related computer support); and representatives from other services. Two approaches were used to collect information---interviews and feedback from a demonstration of a simulated RIS. The interviews were conducted first and consisted of semi-structured sessions to determine the information needs and requirements of the interviewees. In the course of these interviews, it was found that most people could not enunciate the specifics of what they would like to have in an information system. They were more informative regarding what they presently do not have. Consequently, a simulated model of what an RIS might be able to do was developed and demonstrated. This "straw man" approach was designed to give potential users something concrete to react to, and, from these reactions, more specific information was garnered regarding user needs and requirements. This section reports in detail the findings from the interviews, the objective and content of the simulated RIS, and the conclusions to be drawn from the reactions of viewers of the demonstration.

Findings from the Interviews

The task of identifying the information needs and requirements of Navy personnel researchers and managers was an ongoing effort that continued throughout the total period of the feasibility study. Early in the contract year the principal investigator spent a day at the Navy Personnel Research and Development Center conducting initial interviews with NPRDC researchers concerning their needs and requirements for a computer-supported attitudinal research information system based on their research objectives and supporting data bases. From these discussions an agenda was constructed for another round of more intensive interviews which were conducted over a 3-day period by two members of the feasibility study team. A third round of interviews with NPRDC researchers spanning three more days was conducted by two members of the feasibility study team a month later. Each interview consisted of an hour's discussion. Without exception, every one of the approximately 50 people contacted was cooperative, helpful, and candid. Therefore, the conclusions to be derived from these interviews are considered to be valid and reasonably representative of the information needs and problems confronting researchers and research managers in the area of Navy attitudinal research.

Figure 2 is a reproduction of the information collection form used during the interviews in order to identify and describe existing data bases at NPRDC. If the interviewee was also able to provide documentation showing how the data base was formatted either on punched cards, magnetic tape, or disk pack, this information was collected along with examples of questionnaires and survey instruments. All of this information was organized into a large notebook compilation, arranged alphabetically by name of the person interviewed. From this organized compilation, an inventory was prepared summarizing the data bases known to be in existence at NPRDC that are candidates for inclusion in an attitudinal research information system. By no means were all of the data bases extant at NPRDC identified, but the ones that were located are important, visible ones as well as representative of past, current, and future holdings.

A draft of this inventory was distributed to the individuals who had participated in the three rounds of interviews, and their review and comments were solicited. More than half of these persons responded by making corrections and/or additions to the inventory. It is assumed that those individuals who did not respond were satisfied with the description of their data bases. An updated version of the inventory of data bases, reflecting the changes and additions made by all of the reviewers, is presented in Appendix B. The inventory is in alphabetic sequence by organizational code numbers.

In Appendix B at the end of the listing of data bases and other information systems at NPRDC, the reader will find a listing of data bases at the Navy Health Research Center that contain personnel data. This listing is followed by a description of related interviews that were conducted with representatives from the Manpower Research and Data Analysis Center (MARDAC) and with computer personnel from the Naval Electronics Laboratory Center (NELC) and the Naval Undersea Center (NUC).

In addition to providing information to develop the inventory of NPRDC data bases, the interviews with NPRDC personnel surfaced a whole host of problems associated with data collection, data processing, and data management. The problems expressed fall into three categories: problems associated with the characteristics of data base design and file maintenance and with computer software and hardware; problems intrinsic to doing research regardless of the setting in which the research is conducted; and problems reflective of the difficulty of managing a large-scale research complex within an even larger monolithic military organization. Each of these problem areas will be discussed in turn.

The general picture that emerged from the interviews is that NPRDC researchers do not have any systematic way of knowing what the Center's data base holdings are except for their own data bases and those of others that they may have happened to learn about. Hopefully, the partial data base inventory presented in Appendix B will constitute a first step toward remedying this situation. Somewhat more than a man-month of effort went into the preparation of the inventory, and it is recommended that someone be given the responsibility to keep it up-to-date and to assure that the updated inventory receives Center-wide dissemination at regular intervals. Maintaining the inventory by computer so it would be available on demand also should be

1. Person Interviewed_____

   Code No._____; Phone No._____; Bldg. No._____

2. Description of Mission of Work Group and Problems Encountered in Carrying Out Mission with Regard to Data Acquisition, Collation, Processing, and Analysis

   a. Objectives of Mission:

   b. Problems Encountered:

3. Existing or Planned Data Base: Brief Tag Name_____

   Full Name_____

   Sample Size and Description_____

   Entire Population?_____ Random Sample?_____ Stratified Sample?_____

   Longitudinal Study?_____ Dates Covered_____ How Often Are the Data Collected or Received?_____

   Unit of Data Collection: Individual_____ Work Unit_____ Ship_____ Other_____

   Explain Other_____

   If questionnaire returns, provide example of the questionnaire.

   Number of items in the questionnaire_____

   How stored now?_____

   If on punched cards or magnetic tape, provide a complete description of the format of the file. Number of fields in the file_____

Figure 2. Information Collection Form for Existing Data Bases at NPRDC.

considered. The large notebook compilation of information about existing data bases from which Appendix B was prepared will be given to NPRDC for their use in maintaining and updating the inventory, or for use by any of NPRDC's contractors if this kind of compilation would aid in the performance of their work statement.

Inadequate documentation of existing data bases clearly is a chronic problem in Navy research as it so frequently is in other research settings. Some data base formats have been documented beautifully, and an example of this excellent type of documentation is presented later in this section. In other cases, documentation is totally lacking or inadequate. In order to include a data base in a computer-based research information system, its documentation must be complete so that it can be incorporated into the schema of the overall system data collection. The role of and nature of a data base schema is discussed in Section 5. The visibility that a data base attains once it is incorporated into an interactive information system should serve as an impetus to document its contents and format as carefully as possible. Regulations requiring documentation according to a prescribed and standard format would go a long way toward helping to solve the documentation problem. Maintenance of the documentation is also an important requirement, particularly if the format of a data base is changed in any way. The most logical individual to assume this responsibility is the Tape Librarian.

In addition to the problem caused by poor or inadequate documentation of magnetic tape files created by Navy personnel researchers themselves, another significant problem revolves around coding and format changes made over time by the Navy Bureau of Personnel (BuPers) that introduce incompatibilities in data if one wants to analyze or select from historical data sets. Rewriting all of the data files involved into a common format would be prohibitive because of time and cost. For a particular research project, with enormous effort, problems of lack of compatibility among related files can be overcome by sheer determination and dedication of purpose, but the price is high.

The present inability to search files in a dynamic manner in order to provide quick responses to inquiries from Navy management is a direct result of the lack of file documentation and the absence of a data base design. The kind of design needed is complex, costly, and dependent on the availability of particular kinds of computer hardware and software. These conclusions and their ramifications with regard to the design criteria for an attitudinal research information system are treated in Sections 7 and 9.

The difficulties expressed by the individuals charged with providing computer support to the personnel research community reflect problems encountered in any computing facility where less than optimal resources are available, where dependency on other computer facilities for data tapes is necessitated, and where coordination of information about who is requesting what from whom is lacking.

From the point of view of the computer support personnel, these are the problems confronting them. They have two groups of individuals with whom they must interface. First, they must try to meet the needs of individual personnel researchers with regard to building and maintaining their data bases and

providing the necessary software to manipulate and analyze the data. Second, they must establish workable channels for securing needed criterion and other data from other components of the Navy. Standard statistical analysis packages have been made available, and software for searching and updating sequential or indexed sequential files has been developed. However, any request to search a file on an unindexed attribute or to relate multiple attributes for an individual retrieved from several files is impossible now without special programming since there is no overall data base design that provides for these information processing requirements. Thus, the available computer support, as compared against an ideal standard, is hardware and software limited, inflexible, and often unresponsive.

Another problem for computer personnel relates to the difficulty of reading magnetic tape files created by other Navy computer facilities. These tapes may be prepared on different computers that are incompatible to processing by NPRDC. Sometimes the tapes are received without any documentation. The contents are unknown much less the speed at which the data should be read, the packing density, and/or the format for interpreting and manipulating the data. If the correct speed to read the tape can be found by trial and error, the resulting printout of the contents of the tape is an alphanumeric hodgepodge. Many of the tapes are deciphered successfully; however, hundreds of tapes are being archived by NPRDC for which no documentation exists. No one knows how to use them; yet no one can justify recycling them for other data storage. In addition, the storage requirements for all of these tapes presents a problem since they need to be stored under conditions of humidity and temperature control.

Sometimes tape files are received in compressed form to permit a higher packing density of the data. However, to be able to ascertain what the data are, they have to be decompressed. This process requires a special decompression routine which may have to be programmed ad hoc. Computer personnel at NPRDC expressed a need for a more formal coordination policy with BuPers so that they might have better knowledge of and control over which files are being requested. At times, magnetic tapes have arrived at NPRDC without computer personnel knowing who to inform of their arrival because they did not know the tapes had even been requested.

The problems intrinsic to doing research are well known to any research-oriented organization. Therefore, it was not surprising to hear these problems expressed once again and in the same old refrain. The Navy attitude surveyors were frustrated when they obtained a disappointing percentage of returns to a questionnaire survey, causing problems in extrapolating from the biased sample to the overall population. Those researchers using a longitudinal design remarked that one had to begin with enormous sample sizes in order to counteract the loss of cases over time if the final sample size was to be considered of respectable magnitude. The always present problem of how to deal with incomplete or missing data was mentioned over and over again. Another problem concerned the very real difficulty of being constrained by the requirement to work with anonymous data, that is, data generated by individuals who voluntarily choose to remain unidentified, or data produced only under the condition that these data must not be associated with the data provider. In the case of anonymous data, there is no way to relate unidenti-

fied responses to other data that would be valuable for ascertaining relationships among variables, classifying individuals based on discriminant functions, or predicting future behavior and performance by regression techniques. The researcher is forced to fall back on a less precise research design in which data are aggregated by work unit, by ship, by ship type, or perhaps even by fleet. Ascertaining the relationship of these aggregated sets of responses to appropriate criterion data becomes a particularly difficult task, probably impossible in the full sense of the objective and certainly challenging of innovative ways for generating and analyzing data that have been subjected to microaggregation. The literature on this subject is not abundant since most researchers confronted with the problem walk away from it and do something else with their intellectual resources. The minimal literature on this extremely difficult issue seems to be coming from those individuals concerned with the protection of privacy and/or confidentiality of individual data in computer systems (e.g., see Feige & Watts, 1970).

Another problem mentioned frequently was the difficulty of obtaining criterion data or even finding out where these data might be obtained. One researcher interviewed stated the problem this way: "Predictor variables exist in abundance; it's the criterion data that are difficult to get." Criterion data usually are difficult to obtain because they are collected and controlled by a different organizational entity than the one to which the researcher belongs. So typically the researcher must go abegging for his criterion data, hoping that the organization having the needed data will be cooperative in providing it in a usable form. NPRDC is no different in this regard. Their criterion data consist of variables such as test scores, school grades, service history, and performance data. These data are garnered from many quarters of the U.S. Navy, all requests dependent upon the establishment of a need for the data and a procedure for delivering the needed data in a usable form on a timely basis. One of the biggest providers of criterion data to NPRDC is BuPers. In the midst of fulfilling their own mission and converting from one computer system to another, BuPers also has tried to provide other Navy components with requested data. Sometimes the data provided have been exasperating because they lacked adequate documentation to use them and interpret them easily; sometimes the very data needed were deleted because of policy decisions regarding the confidentiality of sensitive data. For example, the performance data field in magnetic tapes received from BuPers may purposefully have been made blank. While difficulty in acquiring criterion data is indicative of the general problem confronting all researchers, it is particularly revealing of the special problem NPRDC has in securing needed criterion variables.

The problems facing managers of personnel and training research and development are more global than the specific problems discussed thus far, and their solution is much more challenging and difficult. Research managers need to have access to information about the status and funding of all ongoing activities, and they may require this information to be displayed in various ways. They need this information to initiate remedial action if projects are behind schedule, to brief and to report to others on the objectives and products of NPRDC research, to plan intelligently for the future, and to establish priorities. They need to be able to coordinate their activities with those of other Navy and DoD elements in order to strengthen links, eliminate

overlap, and/or fill in gaps. Also, they should be aware of related research being conducted by Navy contractors. If requests for special information are made by BuPers management or by others within the Navy management hierarchy, the objective is to be as responsive as possible, as quickly as possible. In order to meet this objective, managers need a well-designed computer-support system, a knowledge of what information is contained in their data base holdings and how to access this information, and an up-to-date inventory of the talents and expertise of NPRDC staff. The inescapable conclusion is that NPRDC management needs a computer-based management information system that they can interact with in order to retrieve any needed information on a timely basis. The availability of this kind of information would form the basis for the development of subsystems concerned with intra-Center communication and with coordination of activities between NPRDC and other components of the Navy and DoD.

Not all of the problems facing research managers that have been identified can be ameliorated by the advent of an attitudinal RIS. Other information systems are under design and development to deal with certain subsets of the problems confronting research management. The STAR (Status of Program Report) system is being developed as a management information system for reporting on the progress of NPRDC research projects. CENDEX comprises an inventory of the talents and interests of the NPRDC staff, part of a larger effort to establish an intra-Center information-communication system to make more efficient use of in-house talents. Also under development is a Research and Development Coordination System (RDCS). Since NPRDC is charged with coordinating the Navy's personnel research and development activities as well as relating them to similar projects in the Army and Air Force, the objective of the RDCS is to strengthen links, eliminate overlap, and/or fill gaps. In addition, the CCOPS (Control and Coordination of Personnel Surveys) system has been established to control and coordinate all personnel survey administration in the Navy in order to eliminate duplication of effort and to prevent oversurveying. The unique contribution that an attitudinal RIS would make to helping to solve the problems of information management would be to systematize the way in which attitudinal research data bases are documented, indexed, and stored. The benefits to be gained from the use of an RIS would be more exhaustive and current coverage of data base holdings, increased accuracy of information, more flexibility in methods for analysis and presentation of information, and ability to locate and present required information on a more timely basis. However, the largest benefit expected to accrue from the advent of an attitudinal RIS would be in the area of secondary analysis. This benefit comprises the cost of collecting survey data and preparing it in a machine-processable form that can be avoided by the secondary data analyst.

Simulation of a Hypothetical Attitudinal RIS

In order to further identify and more sharply define the information needs and requirements of Navy personnel researchers and managers, a simulation of a hypothetical attitudinal RIS was developed. The objective of this effort was to embody some tentative design concepts and principles in a simulated model of an interactive information system in order to determine which aspects of the model would receive a favorable reaction when viewed by potential users.

The simulation was carried out by means of a cathode-ray terminal connected to a tape deck and a peripheral printer. The contents of the demonstration of the hypothetical RIS were predetermined and stored on magnetic tape cassettes from which information could be called up for display on the video screen of the terminal. The transfer rate of data among these three pieces of equipment was 1,200 baud or approximately 120 characters per second. This speed of response is reasonably acceptable for man-machine interaction; however, experienced system users would consider this rate too slow. Speed of system response is considered in detail in Section 6 where the variables involved in designing an optimal man-machine interface are discussed.

The basic premise underlying the model was that system users should have a convenient way of finding out what data base holdings exist and then should be able to explore the contents of any holding in order to decide upon an appropriate strategy for using and analyzing data of interest. Every effort was made to make the simulation realistic, information rich, and responsive to the information needs and requirements that had been expressed during the series of interviews. A logical decision tree structure was used to design and develop the simulated hypothetical attitudinal RIS. As a very first step, the user signs onto the system by providing his user number, and then he is given a brief introduction to the objectives of the system and instruction in how he may access various kinds of information. Figure 3 is a printout of the initial display presented to a user by the simulated attitudinal RIS. The user is informed that there are three primary ways of approaching the information contained in the RIS---via an INDEX OF TYPES OF HOLDINGS, via a SUBJECT CATEGORY INDEX, or via a RESEARCHER NAME INDEX. He also is informed that whichever avenue of approach he chooses initially, he ultimately will be led to the other two options somewhere in his search and exploration. In Figure 3 the user selected the first option by placing an "X" to the left of INDEX OF TYPES OF HOLDINGS. This selection was accomplished by positioning a controllable cursor to the left of the line chosen and typing an "X" (see the discussion of CRT terminals in Section 6).

The RIS then displays the INDEX OF TYPES OF HOLDINGS STORED IN THE ATTITUDINAL RESEARCH INFORMATION SYSTEM as shown in Figure 4. The user is instructed that he can request an expanded definition of any category shown in the index by placing an "X" to the left of the line of interest. In the example shown in Figure 4, the user selected DATA BASES ON MAGNETIC TAPE, SAMPLE LISTINGS. The categories shown in Figure 4 probably are not completely inclusive of the types of holdings which ultimately might be contained in an RIS, but they are inclusive of the holdings enumerated in Appendix B. Thus, this index represents a substantial beginning to understanding the scope and breadth of information that might be contained in an actual RIS.

Figure 5 portrays a printout of a portion of the data bases on magnetic tape holdings of the simulated attitudinal RIS. In this hypothetical interaction with the RIS, let us suppose that the user was interested in seeing a sample listing of the job preferences of U.S. Marine Corps recruits (Tape Number 900490). The system then would display the information shown in Figure 6. Thus, the individual viewing this display would be able to discern

Please sign on to the system by giving your user number next: 57

ATTITUDINAL RESEARCH INFORMATION SYSTEM

This system is designed to assist in the performance of Navy personnel attitude research by facilitating secondary analysis, trend analysis, statistical analysis, forecasting, modeling, and simulation. Another objective is to instruct users about the nature of current and past research data bases included in the system. There are three primary ways of approaching the information contained in this system.

X INDEX OF TYPES OF HOLDINGS is categorized according to the material form in which the subject matter originally was obtained (e.g., abstract, magnetic tape, inventory, questionnaire).

SUBJECT CATEGORY INDEX indicates the system contents organized according to their subject matter.

RESEARCHER INDEX organizes the system contents by the names of the researchers associated with each portion of the holdings.

The three indexes itemized above are arranged in alphabetic order, and they are interrelated by cross references. Whichever avenue of approach for your search that you choose initially, you ultimately will be led to the other two options somewhere in your search. For example, if you begin your search through the INDEX OF TYPES OF HOLDINGS, you will have the option somewhere in your search of approaching the type of holding you are interested in by subject categories or by researcher, if appropriate.

(NOTE: For an expanded version of any of these three indexes, request the index of your choice by placing an "X" to the left of the first line of description.)

Figure 3. Initial Display Presented to a User by the Simulated Attitudinal RIS.

INDEX OF TYPES OF HOLDINGS STORED
IN THE ATTITUDINAL RESEARCH INFORMATION SYSTEM

X Data Bases on Magnetic Tape, Sample Listings

Job Task Inventory (see Task Inventory)

Medical Records

- Narrative Summaries of Drug Abusers
- Narrative Summaries of Neurotic Patients
- Narrative Summaries of Psychotic Patients

Personal Data Sets

Questionnaire Techniques

Questions and Questionnaires

Reports, Abstracts of Attitudinal Research

- Author Index
- Manpower Systems
- Personnel Measurement
- Personnel Systems
- Studies, Attitudes and Motivation (Individual Level)
- Studies, Demographic
- Studies, Group Dynamics (Work Group Level)
- Studies, Organizational (Activity, Unit Level)

Statistical Software Packages

Survey Questions and Questionnaires (see Questions and Questionnaires)

Task Inventory

(NOTE: For an expanded definition of any category shown in this index, request the specific category by placing an "X" the left of the line.)

Figure 4. Display of the Index of Types of Holdings Stored in the Simulated Attitudinal RIS.

DATA BASES ON MAGNETIC TAPE

TAPE NUMBER	FILE NAME	RECORD LENGTH	BLOCK SIZE	PERSON RESPONSIBLE	PHONE NUMBER
402290-402489	Length of Service by Pay Grade	36	6048	Lonsdale, N.	225-6721
	DESCRIPTION: This file consists of 200 magnetic tapes containing 2-way matrices of length of service data by pay grade for all enlisted ratings since 1965. Each matrix shows 31 length of service (LOS) categories (1 through 30 years plus 31 years and over) vs. the 9 enlisted pay grades.				
900490	USMC Preference Options (Files 2 through 9)	80	6400	Rafacz, B. A.	225-2176
	DESCRIPTION: This project determined the feasibility of granting first-term USMC recruits their job preference option. From an initial sample of 14,000 men, the sample reduced to 7,452 and 2,480 men at the remaining two data collection periods. Samples were further reduced in size because not all men reported complete data.				
U486AH	OCR Performance Evaluation	140	4480	Royle, M. H.	225-2283
	DESCRIPTION: This file consists of OCR performance evaluations for Pay Grades E5 to E9 covering the period 1967 to the present. There are some missing data because of trouble reading the OCR forms. N = 100,000 approximately for Pay Grades E5 and E6; N = 50,000 approximately for Pay Grades E7 through E9.				
903420	Civil Engineer Corps Career Motivation Data	135	2880	Somer, E. P.	225-2191
	DESCRIPTION: This file contains the questionnaire responses of Civil Engineer Corps officers and former officers to a survey regarding their career motivation. N=2,051 for active duty officers and N=1,641 for officers who had left the CEC.				
U402-U418	Monthly Extract of Master Enlisted Tape	90	7200	Suiter, R.	225-6452
	DESCRIPTION: This file consists of 16 magnetic tapes of key variables extracted monthly from the Master Enlisted Tape.				

(NOTE: You may choose one of three possible options at this point. Indicate your choice by placing an "X" to the left of the option below that you select.)

X I want to see a sample listing of TAPE NUMBER (specify): 900490

I want to initiate a job request involving one of these magnetic tapes. Please set up the necessary Job Control Language.

I want to see the Subject Index for the DATA BASES ON MAGNETIC TAPE.

Figure 5. Display of a Portion of the Data Bases on Magnetic Tape Holdings of the Simulated Attitudinal RIS.

TAPE 900490

FILE NO.	DESCRIPTION	NO. OF RECORDS	NO. OF MEN
1	*Disregard This File*★		
2	Admin.One.Orig.Data	13,624	13,624
3	Admin.Two.Orig.Data	9,949	9,949
4	Admin.Three.Orig.Data	4,474	4,474
5	Admin.One.Two.RD3.MMS	7,509	2,503
6	USMC.Tape.RD3MMS.ADM2	9,926	9,926
7	Three.Admin.Plus.RD3.MMS	12,720	3,180
8	Admin.TwoNot.InOne	8,382	2,794
9	Descp.ofAll.Files	380	----

★ This is an earlier version of File No. 9.

If you want to see the format for and a sample listing of any of the files contained on this tape, please specify the file number below.

FILE NO. 2

Figure 6. Display of the Files Written on Tape 900490 Containing the Job Preferences of U.S. Marine Corps Recruits.

that the initial sample of USMC recruits consisted of 13,624 men; but as subsequent job preference data were collected, the sample size reduced ultimately to 4,474. In Figure 5 the system viewer was told additionally that because not all men reported complete data, the final complete sample consisted of only 2,480 men. Armed with these facts as background information, the hypothetical viewer of the RIS now requests to see a sample listing of File No. 2 on Tape 900490, the file containing the job preferences of the original USMC sample. Figure 7 shows the documentation of the contents of this file. From this figure it can be seen that the original data format probably was 80-column punched cards, subsequently written onto magnetic tape. The documentation describes all of the fields and indicates information needed to read the tape and interpret it properly. In addition, the system viewer is provided with a sample listing of the data for the first ten men of File No. 2 of Tape 900490. In Figure 7, fictitious names have been substituted. The viewer can perceive at a glance that this data base has certain limitations. First, some fields have data missing. Second, the file does not contain an identification number, such as Social Security Number, which would make it possible to find relevant data for each individual in other data bases. The only field potentially in common is the name of the individual, but a merger of separate files would be better accomplished based on a more unique identifying field. Despite the problems just enumerated in using this data base, the documentation is excellent. A potential user knows exactly what problems he has to deal with, and computer personnel know precisely how to read and manipulate the tape.

Suppose in Figure 5 that the user of the RIS had requested instead to see a sample listing of 2-way matrices of length of service (LOS) data by pay grade for all enlisted ratings since 1965 (Tapes 402290-402489). He might have been presented with the display shown in Figure 8 in which the FY '75 LOS by Pay Grade matrix for aircraft maintenance technicians is cross tabulated. This matrix shows 31 length of service (LOS) categories (1 through 30 years plus 31 years and over) versus the nine enlisted pay grades. In the example shown in Figure 8, the system user chooses to retrieve similar LOS by Pay Grade matrices for three additional rates. In Figure 9 the user learns that the retrieval process involves more than a simple fetch of a sample listing. The system generates and displays the Job Control Language needed to retrieve other LOS by Pay Grade matrices from tape files that might have to be mounted on tape drives at the computer center. In the example shown in Figure 9, the user also has the option of having his output printed on the high-speed line printer at the computer center or on the 1200-baud printer at the computer terminal where he signed on (for a definition of baud rate, see the discussion of CRT terminals and the glossary of key terms in Section 6). In this instance, the user chooses the local terminal to receive the printout. In the final part of his interaction with the RIS, the user signs off the system as shown in Figure 9. However, he had two other options to continue his interaction by either returning to the INDEX OF TYPES OF HOLDINGS or returning to the DATA BASES ON MAGNETIC TAPE display. Additional options could be included depending on user feedback of other desirable kinds of information displays that might logically be required at this point.

TAPE 900490: FILE NO. 2 - RECORD LENGTH=80, BLOCK SIZE=6400, RECFM=FB
ADMINISTRATION 1 OF MAPS(REV.3), N=13,624

RECORD POSITION	INFORMATION
1-3	Blank
4-18	Name (Last: first)
19-27	SSN
28	Blank
29-30	Occupational Choice
31-32	Blank
33-34	First Choice on Entering USMC
35	Blank
36-37	Second Choice on Entering USMC
38	Blank
39-40	Third Choice on Entering USMC
41	Blank
42-47	Questions 2-7, Part B
48-74	Blank
75-76	Year of Administration
77-78	Month of Administration
79	Location (1=Parris Island; 2=San Diego)
80	Administration (1, 2, or 3)

BELOW IS A SAMPLE LISTING OF THE DATA ON THE FIRST TEN MEN OF FILE NO. 2 OF TAPE 900490:

AYERS EDWARD R	58	07	03	06	113222	721111
BAILEY GEORGE J	23	04	05	06	112113	721111
BOLEN CLARENCE P	01	12	10	09	123333	721111
BURROWS CARL	58	03	06	14	211223	721221
FONTANA MARK	46	03	04	05	111322	721121
LANG ROBERT M	08	01	02	03	112222	721221
McKNIGHT PAUL	58	14	01	07	111123	721121
TOLSTEAD WILLIAM				1	112	721121
WIRTH SIDNEY	35	03	04	13	211222	721111
WOLCOTT MARTIN G		05			3	721221

See the Tape Librarian in the User Support office of the Data Analysis Department, Building 602, Room 118, if you wish to use this tape in your research.

Figure 7. Display of the Documentation for File No. 2 of Tape 900490.

LOS	E1	E2	E3	E4	E5	E6	E7	E8	E9	TOTAL
01	60	214	30	2	0	0	0	0	0	306
02	14	287	791	68	0	0	0	0	0	1160
03	4	17	356	427	14	0	0	0	0	818
04	1	4	49	261	74	0	0	0	0	389
05	0	0	11	115	111	0	0	0	0	237
06	1	0	3	28	80	0	0	0	0	112
07	0	0	6	16	81	2	0	0	0	105
08	0	0	0	11	57	12	0	0	0	80
09	0	0	1	4	35	18	0	0	0	58
10	0	0	0	1	29	11	0	0	0	41
11	0	0	0	0	24	19	0	0	0	43
12	0	0	0	0	11	19	1	0	0	31
13	0	0	0	0	17	28	3	0	0	48
14	0	0	0	2	16	44	2	1	0	65
15	0	0	0	0	6	49	7	0	0	62
16	0	0	0	0	9	41	19	1	0	70
17	0	0	0	0	3	41	12	1	0	57
18	0	0	0	0	0	24	19	0	0	43
19	0	0	0	1	4	30	33	4	1	73
20	0	0	0	0	9	47	32	6	3	97
21	0	0	0	0	3	16	15	11	0	45
22	0	0	0	0	0	3	11	1	1	16
23	0	0	0	0	0	7	10	2	1	20
24	0	0	0	0	0	2	5	6	3	16
25	0	0	0	0	0	0	2	3	1	6
26	0	0	0	0	0	0	3	2	2	7
27	0	0	0	0	0	0	2	2	0	4
28	0	0	0	0	0	0	2	1	2	5
29	0	0	0	0	0	0	0	1	0	1
30	0	0	0	0	0	1	2	1	1	5
31	0	0	0	0	0	0	3	0	1	4
TOTAL	80	522	1247	936	583	414	183	43	16	4024

If you wish to retrieve other LOS by Pay Grade matrices, specify the rate(s) as follows: BT, CS, RM

Figure 8. Sample Listing of the FY '75 LOS by Pay Grade Matrix for Aircraft Maintenance Technicians.

The Job Control Language (JCL) to retrieve other LOS by Pay Grade matrices is shown below. It will be transmitted automatically to the computer when you indicate the location you prefer for printing the output.

```
//B019LPGM JOB (B019,4,,12K),BROEDLING
//JOBLIB DD DSN=FS.B019.LOSPAYGD,DISP=(SHR,PASS)
//BATCH EXEC PGM=EXPRGRAM,REGION=225K
//PRGMPRNT DD SYSOUT=C
//PROGRAM0 DD DSN=FS.B019.LOSPAYGD(PROGRAM0),DISP=(SHR,PASS)
//PRT     DD SYSOUT=C,DCB=(BLKSIZE=6048,LRECL=80,RECFM=FBA)
//OPT     DD DSNAME=&OPT,DISP=(NEW,PASS),UNIT=SYSDA,SPACE=(TRK,(1))
//IDT  DD DSNAME=&IDT.DISP=(NEW,PASS),UNIT=SYSDA,SPACE=(200,(1,5))
//TMP  DD DSN=&TMP,DISP=(,PASS),UNIT=SYSDA,SPACE=(1200,(2,5),RLSE)
//DLM     DD DSNAME=&DLM,DISP=(NEW,PASS),UNIT=SYSDA,
//              SPACE=(520,(2,5),RLSE)
//FRQ  DD DSN=FS.B019.NAVYFREQ,DISP=OLD
//CRD   DD *
//PROGRAM1  DD DSN=FS.B019.LOSPAYGD(PROGRAM1),DISP=(SHR,PASS)
```

Please indicate where you want your output to be printed by placing an "X" to the left of one of the two choices listed below:

```
        High-speed line printer at the computer center
   X    1200-baud printer at the computer terminal where
           you signed on
```

(NOTE: You may choose one of three possible options at this point. Indicate your choice by placing an "X" to the left of the first line of the option below that you select.)

```
        Return to the INDEX OF TYPES OF HOLDINGS Stored in the
           Attitudinal Research Information System

        Return to DATA BASES ON MAGNETIC TAPE display

    X   Sign Off the System
```

Figure 9. Job Control Language To Retrieve Other LOS by Pay Grade Matrices, Printout Choices, and Subsequent User Options.

The RIS also could maintain a log of user activity. From this log, reports for those individuals managing the RIS could be prepared on a regular basis indicating how the system was being used. In addition to the amount of time users were spending interacting with the system (determined from sign-on and sign-off times), management also could be informed about what kinds of information were being accessed most frequently. Thus, data bases with high utilization rates would become candidates for disk storage and further refinement (such as multi-indexing), and data bases with low utilization might be maintained off-line in order to make room for information of higher priority and interest. The existence of the rarely utilized data bases, however, would be maintained in the INDEX OF TYPES OF HOLDINGS. This constant monitoring of user-system interaction would provide the management data needed to make intelligent decisions about which aspects of the system should be emphasized and upgraded, and which aspects should receive a lower priority of attention.

Another way that an attitudinal RIS could be invaluable is in responding to inquiries from the operational Navy about which attitudes have been surveyed. Suppose an inquiry was received from Navy management in Washington, D.C. requesting the results of any surveys of naval personnel concerning their attitudes toward drug abuse. In this instance, the user of the simulated attitudinal RIS would have selected QUESTIONS AND QUESTIONNAIRES in Figure 4. The system then responds by presenting the display shown in Figure 10. Thus, the viewer is provided with an index of search terms for interrogating the RIS holdings of questions and questionnaires. From this index he can see that some questions about drug abuse have been asked in the past. He also learns that the RIS contains narrative summaries of drug abusers treated at the Naval Drug Rehabilitation Center, Miramar. By selecting the search term DRUG ABUSE, the user is able to call up all questions on this subject asked in previous surveys. Figure 11 shows the initial portion of questions retrieved by this search. The user first learns that there are 10 questions or items related to drug abuse that have been asked in questionnaires or surveys administered since 1970. The question content is provided along with sampling characteristics and item statistics. Thus, from a perusal of the first two questions retrieved, it is possible to discover that the same question was asked two years in a row in two different survey instruments. The distribution of responses reveals that there was a slight shift in attitudes expressed toward a lesser feeling of freedom to tell a supervisor about a drug problem in one's work group. For simulation purposes, the response data, the psychometric data, and the method used to distribute the questionnaires are fictitious. However, the questions are real questions included in these two surveys. Upon request the RIS would display the remaining eight questions dealing with attitudes about drug abuse, and the system user would have the option to make a hard-copy printout of any display that interested him.

Returning to Figure 4, suppose the system user wanted to see a data set that he had generated for his own research support. In this case, he would have selected PERSONAL DATA SETS and might have been presented with a display similar to the one shown in Figure 12. Figure 12 depicts six possible ways that system users might find to adapt the RIS's capabilities to their own information needs. A researcher might want to formulate a set of abstracts and references on some particular topic. Maintaining this set by computer would facilitate updating it and printing copies of it in response to requests

QUESTIONS AND QUESTIONNAIRES
INDEX OF SEARCH TERMS

Attitudinal

- Alcohol Abuse
- X Drug Abuse (see also "Medical Records" for narrative summaries of drug abusers treated at the Naval Drug Rehabilitation Center, Miramar, California)
- Homeporting
- Intercultural Relations
- Job Duties
  - Activities
  - Command Climate and Services
  - Counseling
  - Peer Leadership
  - Supervisor Leadership
  - Training
- Race Relations

Demographic

- Personal Information
  - Education
  - Personal Data
    - Age
    - Ethnic Origin
    - Marital Status
    - Sex
- Service-Related Information
  - Enlistment History and Plans
  - Job History

(NOTE: To retrieve related questions, place an "X" to the left of the search term that you select.)

Figure 10. Index of Search Terms for Retrieving Related Questions and Questionnaires.

THERE ARE 10 QUESTIONS OR ITEMS RELATED TO DRUG ABUSE THAT HAVE BEEN ASKED IN QUESTIONNAIRES OR SURVEYS ADMINISTERED SINCE 1970. THE SAMPLING CHARACTERISTICS AND ITEM STATISTICS FOR THESE 10 QUESTIONS ARE GIVEN BELOW:

1. To what extent would you feel free to tell your supervisor about a drug problem in your work group?

1)	To a very little extent	8%
2)	To a little extent	38%
3)	To some extent	41%
4)	To a great extent	10%
5)	To a very great extent	3%

The above question was included in the Navy Human Resources Management Survey as Question 81, and administered by mail in 1973 to 3,739 enlisted men. Its test-retest reliability after an interval of six weeks was found to be .78, as measured from a sample of 250 enlisted men.

2. To what extent would you feel free to tell your supervisor about a drug problem in your work group?

1)	To a very little extent	13%
2)	To a little extent	41%
3)	To some extent	34%
4)	To a great extent	9%
5)	To a very great extent	3%

The above question was included in the Navy Human Goals Survey as Question 106, and administered by mail in 1972 to 4,121 enlisted men.

Figure 11. Initial Portion of Questions Retrieved by a Search on the Search Term DRUG ABUSE.

PERSONAL DATA SETS

	NAME OF RESEARCHER	PHONE NUMBER	NAME OF DATA SET
	Broedling, L.	225-2181	Abstracts and References to Social Science Data Archives
X	Lonsdale, N.	225-6721	Tape Library Log of LOS by Pay Grade Data Base
	Moonan, W. J.	225-2408	FIND (Retrieval System for Personal Collection of Research Documents)
	Robertson, D. W.	225-2283	Draft of Section 2, Research Report SRR 75-6
	Royle, M. H.	225-2283	Job Control Language Statements for Frequently Used Computer Runs
	Thomas, E.	225-2396	Names and Addresses of Responders to the QUICK Survey

(NOTE: These data sets are proprietary to the researchers who created them. A personal data set can be accessed only by supplying the correct code number in the following field: 2076, and placing an "X" to the left of your name.)

Figure 12. Display of Personal Data Sets That Might Be Incorporated into an Attitudinal RIS.

from others. The LOS by Pay Grade data base is very large, consisting of 200 or more tapes. A tape library log of this data base could be maintained by the RIS to preserve the documentation of each tape file and to keep track of who might be using each tape. The RIS also could provide for the storage and retrieval of personal collections of research documents. If the software supporting the RIS had a text editor, drafts of research reports could be input to the RIS by means of a terminal keyboard, easily revised by means of the text editor, and printed on demand. For frequently used statistical programs, a personal file of job control language statements could be maintained by the RIS for easy and rapid access rather than having to manually prepare or assemble these punched cards for each new computer job. This method of maintaining the JCL statements also would help to minimize the occasions for introducing new keypunching errors. And finally, the RIS might be used to maintain rosters of individuals responding to surveys, replies to a letter of invitation to attend a professional meeting, directories of personnel, a talent inventory, or whatever need a researcher had for organized, easily accessible information. Types of personal data sets that might be generated probably would reflect the ingenuity of individual researchers in taking advantage of the RIS's capabilities. By limiting access only to valid code numbers or by some other mechanism to protect the privacy of proprietary data sets, the researcher's personal files would be confidential, to be shared only at his own discretion.

There is a fairly plentiful literature on computer support for personal files. In fact, Lancaster and Fayen (1973, pp. 296-310) devote an entire chapter to this topic in their book entitled *Information Retrieval On-Line*. They point out that the building of personal document collections is an essential ingredient in research and analytical activities. The personal collection is recognized by them to be the first source to which a scientist or other practitioner will turn when the need for information arises. Foskett (1970) and Jahoda (1970) have published monographs on how to organize and index personal files and document collections, and other articles describing computer support of the researcher's personal data sets are available (Burton & Yerke, 1969; Glantz, 1970; Yerke, 1970).

Returning once again to Figure 4, the user of the simulated attitudinal RIS is given the option of finding out about statistical software packages accessible through the system. If he had selected this option, he might have been presented with the display shown in Figure 13. In Figure 13 the user learns that three sets of statistical programs are available through the RIS ---Biomedical Computer Programs (BMD), Statistical Programs for Social Science (SPSS), and Organized Set of Integrated Routines for Investigations with Statistics (OSIRIS). The user is offered the further opportunity of seeing an index for any of these three statistical software packages. Let us suppose that the user has just received a new set of raw data and he wants to learn about the distributional properties of this data set. He might want to plot the raw data in the form of a histogram. Let us further assume that this researcher is most familiar with the BMD package, and so he requests to see an index to the BMD programs since BMD manuals typically do not exist in every researcher's office. Figure 14 is a display of this index. Since a histogram is a descriptive statistical technique, the researcher probably would request more detailed descriptions of the Class D programs. This additional informa-

STATISTICAL SOFTWARE PACKAGES

The Attitudinal Research Information System makes directly available to researchers three sets of statistical programs---Biomedical Computer Programs (BMD), Statistical Programs for Social Science (SPSS), and Organized Set of Integrated Routines for Investigations with Statistics (OSIRIS). Manuals describing how to use these programs are available in the User Support office of the Data Analysis Department, Building 602, Room 11B. Consultants are available by appointment to answer questions and assist you in selecting and setting up appropriate statistical analyses of your data. Phone 225-2181 to make an appointment.

If you would like to see an index for any of these three statistical software packages, please place an "X" to the left of one of the three choices offered below:

____ BMD
____ SPSS
____ OSIRIS

Figure 13. Display of Statistical Software Packages Available through the Simulated Attitudinal RIS.

INDEX OF BMD PROGRAMS

Class D	Description and Tabulation
Class F	Contingency Tables
Class M	Multivariate Analyses
Class R	Regression Analysis
Class S	Special Programs
Class T	Time Series Analysis
Class V	Variance Analysis

If you want to see more detailed descriptions of any of these classes of programs, please specify which class below.

Class D

Figure 14. Display of the Index of BMD Programs.

tion is displayed in Figure 15 where the requester can see that there are two BMD programs that generate histograms. For the requester's purposes, BMD05D appears to be the appropriate program to use, so he asks to see a brief description of this program, which is displayed in Figure 16. This display describes the nature of BMD05D in enough detail so that the researcher can decide if it meets his statistical requirements. The RIS then provides four options, three of them involving jumps to other branches of the logical decision tree and one permitting the user to conclude his interaction with the system.

As a final example of how a user might interact with an attitudinal RIS to obtain needed information, suppose that a researcher wants to find out if any previous research had been conducted in the Navy on career motivation and incentives during recruit training. In Figure 4 this user would have seen that the RIS holdings include abstracts of attitudinal research reports. Under this category he could have narrowed his search by selecting abstracts of studies on attitudes and motivation at the individual level. He might have been presented then with the display shown in Figure 17 in which a subject index for abstracts of attitudinal research reports based on studies of attitudes and motivation at the individual level is provided. By scanning this subject index, the researcher finds out that some research has been done on career motivation and incentives and also on recruit training. Furthermore, he is informed that there are 37 abstracts in the RIS dealing with career motivation and incentives and seven abstracts dealing with recruit training. By phrasing his search criteria as a boolean expression, he can request the logical product of these two subject names. Thus, he will retrieve all abstracts dealing with both subjects, but not those abstracts dealing with either subject alone. If the system user does not know how to formulate a boolean search expression, the RIS offers instructional assistance.

Although there are other information systems for retrieval of research documents or their surrogates (abstracts), none of these systems provides comprehensive coverage of all of the available literature. For example, important working papers, technical memoranda, and interim reports would not find their way into the Defense Documentation Center (DDC) system or the National Technical Information Service (NTIS) system. The DDC system contains a compendium of both classified and unclassified technical literature dating back to 1958 on research, development, and technology produced by the Department of Defense and this organization's contractors. The NTIS system receives unclassified input from several government agencies including NASA, AEC, DDC, and other DoD agencies. This facility, in addition to being a clearinghouse for technical information, attempts to provide a search service of their document holdings. The Army Research Institute has recently let a contract for the development of a bibliographic retrieval system for documents in military sociology since this literature typically is not included in the holdings of DDC or NTIS. A commercial proprietary on-line document retrieval service is provided by Lockheed's DIALOG system, which includes an impressive list of users including the AEC and the Office of Education's ERIC network. However, no classified literature is contained in this system, and there is a fee for service. The NPRDC library has a Lockheed DIALOG terminal.

CLASS D BMD PROGRAMS
DESCRIPTION AND TABULATION

PROGRAM NUMBER	PROGRAM NAME
BMD01D	Simple Data Description
BMD02D	Correlation with Transgeneration
BMD03D	Correlation with Item Deletion
BMD04D	Alphanumeric Frequency Count
BMD05D	General Plot including Histogram
BMD06D	Description of Strata
BMD07D	Description of Strata with Histograms
BMD08D	Cross-tabulation with Variable Stacking
BMD09D	Cross-tabulation, Incomplete Data
BMD10D	Data Patterns for Dichotomies
BMD11D	Data Patterns for Polychotomies
BMD12D	Asymmetric Correlation with Missing Data
BMD13D	t Program

If you would like to see a brief description of any particular BMD program in the Class D series, please specify it below by its program number.

BMD05D

Figure 15. Description and Tabulation of the Class D BMD Programs.

BRIEF DESCRIPTION OF BMD05D
GENERAL PLOT INCLUDING HISTOGRAM

This program produces scatter plots or histograms. Two methods of scatter plotting are available.

The first method gives a one-page graph which has 50 units vertically and 100 units horizontally. The points are automatically scaled to conform to these dimensions.

The second method gives a multiple-page graph with as many units vertically as there are values of the base variable. The values of the base variable (cases) must be ordered and consecutive. The base variable is not scaled.

A one-page histogram can be produced, with a maximum of 34 intervals. The width of the interval must be specified.

(NOTE: You may choose one of four possible options at this point. Indicate your choice by placing an "X" to the left of the first line of the option below that you select.)

Return to the INDEX OF TYPES OF HOLDINGS Stored in the Attitudinal Research Information System

Return to the STATISTICAL SOFTWARE PACKAGES display

Return to the INDEX OF BMD PROGRAMS

Sign Off the System

Figure 16. Display of a Brief Description of BMD05D: General Plot Including Histogram.

SUBJECT INDEX FOR ABSTRACTS OF ATTITUDINAL RESEARCH REPORTS
BASED ON STUDIES OF ATTITUDES AND MOTIVATION (INDIVIDUAL LEVEL)

SUBJECT NAME	NUMBER OF ABSTRACTS
Advancement	12
Assignments	8
Career Motivation and Incentives	37
Category IV Men	2
Homeporting	5
Housing	
Military, Family	6
Bachelor	3
Job Training	4
Methodological Studies	3
Survey Reliability	1
Survey Design	2
Moonlighting	2
Overseas Living Conditions	6
Pay and Financial Benefits	8
Recruit Training	7
Recruitment and Enlistment	8
Re-enlistment and Enlistment	21
Retirees and Veterans	3
Reserves	5
Sea Duty	3
Ship Facilities	2
Uniforms	8
Wives' Attitudes	7

NOTE: This collection of abstracts can be searched by individual subject name or by any boolean combination of subject names. If you want to see examples of how to formulate a boolean search expression, indicate YES or NO next:
If NO, enter your boolean search expression next:

Figure 17. Display of a Subject Index for Abstracts of Research Reports on the Attitudes and Motivation of Individuals.

Since none of these document collections are complete, in as much as the turnaround time to access any of these systems may be considerable, and because there is a charge for the services provided, it makes sense for the RIS to maintain a specialized collection of abstracts of research reports pertinent to the day-to-day working interests of Navy personnel researchers. The individual researcher would be able to find out at a convenient local terminal what literature was available to elucidate a particular research problem. The availability of the complete document from one of the national information services would be noted so that the document could be ordered if desired. But even more important, the existence of informal research memoranda and reports not archived in these national information centers would be made known so that personnel researchers could obtain access to these other information sources through mechanisms provided by the RIS.

All of the simulated RIS displays reproduced in this section were created with the purpose of helping hypothetical system users find out what data are available and how to take advantage of their availability. Users of the simulated system were given the opportunity to explore various pathways in the logical decision tree structure, they were offered information in increasing amounts of detail as they pursued particular logical pathways to their conclusion, and they were given options to branch to other displays in the decision tree that were likely candidates for what to do next. In addition, instructional aids were made available to the user as needed. Additional displays from the simulated attitudinal RIS are reproduced in Section 6 where factors facilitating the operational interface between the user and the system are discussed in detail.

By and large, the reaction of those individuals who viewed the demonstration of the simulated system was favorable. Over 60 people attended the 12 demonstration sessions, and they offered a gamut of comments ranging from unsophisticated questions to very penetrating observations. In some instances, viewers forgot that the demonstration was only a simulation of a hypothetical system, and they were disappointed to learn that they could not explore certain interesting pathways because the progression through the displays had to be made in a predetermined sequence. From these reactions it can be concluded that the simulated RIS was realistic enough to create the illusion that it was being driven by a sophisticated computer system.

Some viewers remarked that what they saw was interesting, but they themselves probably would not use the system very much. Other viewers asked how soon the system would be in operation since they really could use it immediately. This range of reactions suggests that in the advent of a real RIS, some researchers would be avid users while other individuals would continue to do research in the usual way. Other questions asked were, "Could I use the system even if I can't type?" and "How would I learn to use the RIS?" Factors crucial to user acceptance of interactive information systems are discussed in Section 6. Section 6 also contains a discussion of various ways of interfacing with an information system, including the role of a user surrogate, and a detailed account of various training approaches is provided.

One important requirement that was mentioned by several individuals during the demonstrations was the desirability of having a graphics plotter for producing figures for reports. Copier-like devices are now available which permit the reproduction of images displayed on a graphics terminal. In Section 7 a graphics terminal with hard-copy output capability has been included in the system design specifications.

The accumulation of knowledge and feedback about information needs and requirements, as described in this section, served as the basis for developing the technical aspects of system implementation, specifying system design parameters, and identifying economic considerations. Sections 7 and 8 deal with these topics in detail.

ANNOTATED BIBLIOGRAPHY

1. Burton, H. D., & Yerke, T. B. FAMULUS: A computer-based system for augmenting personal documentation efforts. *Proc. Amer. Soc. Information Sci.*, 1969, *6*, 53-56.

FAMULUS is a computer-based system designed to support the documentation activities of the individual scientist with minimum interference in his information-organizing habits and instincts. It operates economically on large, third-generation computers.

2. Fiege, E. L., & Watts, H. W. Protection of privacy through microaggregation. In Bisco, R. L. (Ed.), *Data Bases, Computers, and the Social Sciences*. New York: Wiley-Interscience, 1970.

This article discusses the advantages and disadvantages of partial aggregation of data to replace microunit observations, which are typically regarded as confidential, by nonconfidential mean values for the grouped observations.

3. Foskett, A. C. *A Guide to Personal Indexes Using Edge-Notched, Uniterm and Peek-a-Boo Cards* (2nd ed.). Hamden, Conn.: Archon Books, 1970.

A monograph on how to organize and index personal files and document collections.

4. Glantz, R. S. SHOEBOX---A personal file handling system for textual data. *AFIPS Conf. Proc. Fall Joint Computer Conf.*, 1970, *37*, 535-545.

A discussion of the design considerations for SHOEBOX, its basic operation, and human factors features. The SHOEBOX system, a part of the MITRE Corporation's long-term effort in the development of text-processing systems, is designed to be the electronic analog of a personal desk file drawer.

5. Jahoda, G. *Information Storage and Retrieval Systems for Individual Researchers*. New York: Wiley, 1970.

This book is intended for the researcher in any field who wishes to start or improve an index to his document collection. It was a major work in 1970 dealing with manual document retrieval systems. A variety of indexes are discussed such as coordinate indexes, edge-notched cards, optical coincidence systems, and citation indexes.

6. Lancaster, F. W., & Fayen, E. G. *Information Retrieval On-Line* (a Wiley-Becker & Hayes Series book). Los Angeles: Melville Publishing Co., 1973. Pp. 296-310.

Chapter 13 of this book is devoted to a discussion of *On-Line Support for Personal Files*. Personal file handling in the following computer systems is described: FAMULUS; RIMS, RIQS, and RFMS; SHOEBOX; AUTONOTE; TIP; SOLER; and SPIRES.

7. Yerke, T. B. Computer support of the researcher's own documentation. *Datamation*, 1970, *16*(2), 75-77.

A discussion of the considerations and constraints that formed the basis for the design of FAMULUS. (FAMULUS is not an acronym. A Famulus was the private secretary of a mediaeval scholar or alchemist. Perhaps the most well-known was Wagner, Famulus of Dr. Faustus.)

SECTION 4. INFORMATION INDEXING ALTERNATIVES

Introduction

This section provides a theoretical discussion of the variables and alternatives involved in the indexing of information for storage and subsequent retrieval. The purpose of the discussion is to help form the basis and justification for deciding which indexing alternative would be most beneficial in an attitudinal research information system and would best suit the needs of Navy personnel researchers and managers. First, the need and mechanisms for vocabulary control and the devices used with the controlled vocabulary will be discussed. Second, the various degrees of vocabulary control will be presented along with their concomitant advantages and disadvantages. The alternatives for the preparation and the organization of a controlled vocabulary are presented next, followed by an important discussion of the various cost and performance aspects of indexing and of the information system's vocabulary. A summary, annotated bibliography, and glossary of key terms are included as a convenience and a further aid to the reader.

Information may be documented and stored using various media such as books, articles, or abstracts; microfiche or microfilm; or magnetic tapes, disks, or data cells. These data collections when organized become a data base that can be interrogated by users to retrieve needed information. In any system that retrieves information, manually or by computer, a method is needed for the collection, organization, and selective retrieval of data from a collection of information.

The raison d'etre for any information storage and retrieval system is to serve the needs of its users. Therefore, the units that form the data base must be collected and organized specifically to coincide with the particular interests and needs of the users. Similarly, the method for retrieving this information must coincide with user interests and needs. The information units contained in an information storage and retrieval system can be retrieved in toto or their surrogates can be retrieved in substitute. A bibliographic data base consists of document surrogates or document records containing full bibliographic citation plus assigned indexing terms. The document surrogate is a representation of the document and consists of descriptive elements that form a unique record. Some elements within the record can serve as retrieval keys for accessing and retrieving the needed information. The decision of which elements to use in document description and which elements to use as retrieval keys should be optimized to best fit user needs. Searches also can be implemented on nonindexed data bases in a computer system. These free text or natural language searches operate on the full text of documents, on document abstracts, or on a set of freely assigned indexing terms. These systems will be discussed in more detail in the subsection *Degrees of Vocabulary Control*. The retrieval objects, for example, can be documents, research projects, names of institutions, qualifications and names of experts, job offerings, or data elements such as census data (Soergel, 1974, p. 3). In this section the retrieval objects usually are referred to as documents or information, but it should be understood that the theory presented here could refer to any type of retrieval object in an information storage and retrieval system.

There are a number of standard bibliographic elements included in a document representation: type of document (e.g., book, article, report), personal author/corporate entry, author affiliation, title, edition, place of publication/publisher, date of publication, collation, and document identifying number. Once the decision is made regarding the bibliographic elements that will be used to describe the document, the actual selection of these elements from a document is relatively straightforward. However, the process of assigning the indexing or subject terms is not simple since the entire document offers itself for the selection process. Once the decision is made as to which subject aspects of a document are of importance to users, the terminology for expressing these subject aspects---the *indexing terms*---must be decided upon. The indexing terms chosen as descriptive elements are access points in the search process and constitute an indexing language. In the design of an indexing language two steps are taken. First, the decision is made as to which aspects of the recorded information are of probable importance to potential users now and in the future. Second, it is decided how these aspects should be expressed so that there will be a good possibility of their matching the words potential users will choose for expressing their information requirements. *The function of bringing the language of the searcher into coincidence with the language of the indexer is served by the controlled vocabulary.* A system without any vocabulary control, such as a natural language system, places an excessive burden on the user and severely limits retrieval recall since there is less chance of a match between words as they occur in text and the words of the searcher in making his request. Before undertaking a discussion of retrieval systems with varying degrees of vocabulary control and the natural language retrieval systems devoid of any vocabulary control, it will be useful to specify the types of devices used in an indexing language for vocabulary control and improved recall (the retrieval of relevant document) or precision (the system holding back nonrelevant documents).

Vocabulary Control and Indexing Language Devices

"The controlled vocabulary exists primarily to assist the searcher in the information retrieval system" (Lancaster & Fayen, 1973, p. 244) by bringing the searcher's language into coincidence with the indexer's language. There is a lack of precision in the English language that creates ambiguity since many words have more than one meaning, and a single idea or concept can be expressed in a variety of ways by using different words. The purpose of vocabulary control is to precisely relate words to an idea or concept in order to maximize the likelihood that both the indexer and the searcher will use identical terminology when referring to the content of a given document. The function of vocabulary control in an information retrieval system is to:

1. insure that a concept is always indexed by the same term (word or group of words),

2. establish which of several synonyms or near synonyms is to be used as a retrieval key and provide references to this key from the possible variants,

3. indicate a relationship between broader concepts (generic terms) and related narrower concepts (more specific terms),

4. indicate a relationship between similar concepts other than the formal generic-specific relationship, and

5. differentiate homographs by the addition of a qualifying word or phrase, such as ATTITUDE (AIRCRAFT) and ATTITUDE (BEHAVIOR).

There are a number of commonly used devices (see Table 1) for the control of a vocabulary in an indexing language. One device that is used to insure that a concept is always indexed by the same term is the *scope note*. Scope notes describe the limit of subject matter embraced by each term. These notes may define a term, limit its use to a concept or certain concepts, or distinguish a term from other terms in the vocabulary. An example of a scope note for the term ATTITUDE is the following:

> ATTITUDE is an enduring, learned predisposition to behave, think, or feel in a consistent way towards persons or things.

TABLE 1

THE MOST COMMONLY USED VOCABULARY CONTROL DEVICES AND THEIR FUNCTION

Thesaurus Control Devices	*Subject Heading Control Devices*	*Function of Device*
scope note	scope note	Consistent indexing of a concept
use (used for, UF)	see (seen from, SF, x)	Synonym and near-synonym control Control of word form
broader term, BT	see also, sa (xx)	Generic-specific relationship
narrower term, NT	see also, sa (xx)	Generic-specific relationship
related term, RT	see also, sa (xx)	Relationship other than generic-specific

Synonyms, two or more words having the same or nearly the same meaning, and near synonyms are controlled by *use* or *see* references which point from an equivalent term to the preferred term that should be used as a retrieval key. The reciprocal indicator for a *use* reference is *used for* or UF. The reciprocal indicator for a *see* reference is *see from* or SF or *x*. For example, where CANNABIS is the acceptable retrieval key, the synonym marihuana would point to the preferred term as follows:

marihuana *use* CANNABIS

The reciprocal is indicated as follow:

CANNABIS
UF marihuana

Use and *see* references also are used to refer from a specific term that should not be used while indexing to the appropriate, more generic term that must be used in its place. This type of reference occurs when it is perceived that users are not really interested in a concept in a specific sense but that the broader concept is of more interest. For example, the users of a system may not be interested in the drugs dilaudid or talwin specifically but in the more generic chemical substance which is morphine. The indexer and user would be led to the preferred indexing term as follows:

dilaudid *use* MORPHINE
talwin *use* MORPHINE

The reciprocal indicator *used for* or UF is used as follows:

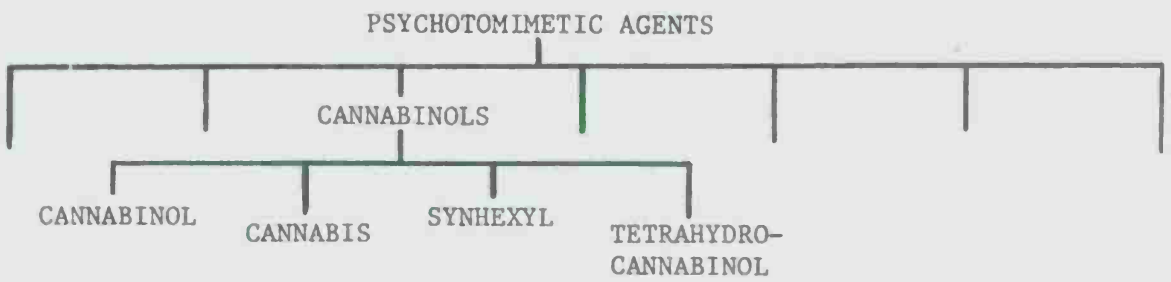
MORPHINE
UF dilaudid
talwin

The *use* or *see* reference also directs from abbreviations, alternative word forms and spellings, and alternative word sequences for indexing terms composed of two or more words to the preferred form of the term. Whether the *use* form or *see* form is used for this reference is a matter of convention, but the purpose remains the same, that is, to direct from a term that cannot be used in indexing and searching to a term that can be used.

The generic-specific relationships between terms can be displayed by means of a formal classificatory organization, which imposes a hierarchical structure on a vocabulary as shown in the following example:

PSYCHOTOMIMETIC AGENTS
●
●
●
CANNABINOLS
CANNABINOL
CANNABIS
SYNHEXYL
TETRAHYDROCANNABINOL

or



It also is possible to impose a hierarchical structure on a vocabulary by means of an appropriate network of cross-references, such as *broader term* (BT) and *narrower term* (NT) or the alternative *see also* reference. These would be illustrated as follows:

CANNABINOLS
BT PSYCHOTOMIMETIC AGENTS
NT CANNABINOL
CANNABIS
SYNHEXYL
TETRAHYDROCANNABINOL

or

PSYCHOTOMIMETIC AGENTS
UF HALLUCINOGENS
BT SEDATIVES AND STIMULANTS
NT CANNABINOLS

The *see also* reference ties together related terms and does not distinguish between the generic-specific relationship and other types of term relationships. The following example illustrates the generic-specific type of relationship:

CANNABINOLS
SA CANNABINOL
CANNABIS
SYNHEXYL
TETRAHYDROCANNABINOL
xx PSYCHOTOMIMETIC AGENTS

In the above example there is no *see also* reference to PSYCHOTOMIMETIC AGENTS since this would not be considered useful. A user searching the index at a level as specific as CANNABINOLS is not likely to find the broad area of PSYCHOTOMIMETIC AGENTS to be of much value. However, as indicated by the *xx* reciprocal reference, there is a *see also* reference from PSYCHOTOMIMETIC AGENTS to the more specific level CANNABINOLS to indicate that a user can narrow his search to the more specific level of drugs.

The RT or *related term* reference is used to indicate a relationship between terms other than the generic-specific type of relationship. The terms

may be related semantically or include related activities, materials, and processes. The indication of related terms plays a suggestive role in a controlled vocabulary since these references serve to suggest to the user other terms that may be of use to him in his search. The *see also* reference also can be used to indicate a relationship between terms, whether it is a generic-specific type of relationship or another type of term relationship. Usually the *related term* reference is automatically reciprocated as in the following example:

MARITAL RELATIONS
RT MARRIAGE COUNSELING

MARRIAGE COUNSELING
RT MARITAL RELATIONS

or

MARITAL RELATIONS
SA MARRIAGE COUNSELING
xx MARRIAGE COUNSELING

and

MARRIAGE COUNSELING
SA MARITAL RELATIONS
xx MARITAL RELATIONS

The *see also* reference is reciprocated by the symbol *xx*. In the above example the *xx* symbol indicates that there is an analogous *see also* reference from MARRIAGE COUNSELING to MARITAL RELATIONS. However, MARRIAGE COUNSELING does not have to refer the user to MARITAL RELATIONS if this is not considered to be a useful or meaningful relationship as in the following example:

MARITAL RELATIONS
SA MARRIAGE COUNSELING

and

MARRIAGE COUNSELING
SA COUNSELING
xx COUNSELING
MARITAL RELATIONS

A *homograph* is a word with the same spelling as another but with a different meaning and origin. In a classification scheme homographs would be distinguished by the context in which they appear in the classificatory structure and they would be distinguished during the search by the use of their alphabetic or numeric code in the search request. For example:

A Aeronautics
Aa Attitude

B Behavior
Ba Attitude

Homographs also may be distinguished by the use of modifying terms or parenthetical qualifiers. For example:

AIRCRAFT ATTITUDE
BEHAVIORAL ATTITUDE

or

ATTITUDE, AIRCRAFT
ATTITUDE, BEHAVIOR

or

ATTITUDE (AIRCRAFT)
ATTITUDE (BEHAVIOR)

The vocabulary control devices just described are used to improve *recall*. However, several other devices used for increasing retrieval *precision* have been given recent attention in the literature on information retrieval. These precision devices do not control the indexing vocabulary itself but instead are used with the indexing terms themselves in order to improve retrieval precision. Recall and precision will be discussed as performance measures in more detail in the subsection *Cost Performance Aspects of Indexing and Vocabulary Variables*. However, it is necessary to briefly define these two measures in this discussion of indexing language devices. Recall refers to the degree of success achieved in retrieving relevant information from a retrieval system, and precision refers to the ability of the system to hold back irrelevant information. Recall devices for vocabulary control and precision devices form the syntax of the indexing language (see Table 2).

TABLE 2

INDEXING LANGUAGE DEVICES USED FOR
INCREASING RECALL AND PRECISION

Recall Devices	*Precision Devices*
Control of synonyms	Coordination (subheadings)
Control of near synonyms	Linking of terms
Control of word forms	Role indicators
Hierarchical grouping	Term weighting
Grouping by statistical association (clumping, clustering)	Word distance indicators

The *linking device* is a precision tool that ties together related terms and separates terms that are not related. For example, the indexing terms for an article dealing with the drug abuse of enlisted men and their wives' attitude towards overseas living conditions are as follows: DRUG ABUSE, ENLISTED MEN, WIVES, ATTITUDE, OVERSEAS, and LIVING CONDITIONS. This article would be falsely retrieved (a false drop) for the following search requests:

1. DRUG ABUSE of WIVES of ENLISTED MEN,

2. The ATTITUDE of ENLISTED MEN towards OVERSEAS LIVING CONDITIONS,

3. The ATTITUDE of OVERSEAS ENLISTED MEN towards WIVES, and

4. The ATTITUDE of WIVES towards DRUG ABUSE.

False drops can be avoided by linking the related terms together in the following manner:

ENLISTED MEN - DRUG ABUSE

WIVES - ATTITUDE - OVERSEAS - LIVING CONDITIONS

For the above article to be retrieved, the search request must contain the related terms in the same link, consequently eliminating the possibility of this article being retrieved falsely for the above four searches.

However, while links indicate a relationship between terms, they do not indicate the *type* of relationship between terms. The problem of incorrect term relationships is solved by another precision tool, the *role indicator*. Suppose we have an article about wives' attitudes toward the Navy which are affected by living conditions. This article would be indexed as follows: WIVES, ATTITUDE, NAVY, and LIVING CONDITIONS. This article would not be of interest to the searcher who wants articles on the ATTITUDE of WIVES towards LIVING CONDITIONS or the ATTITUDE of the NAVY towards WIVES. In order to avoid false drops because of incorrect term relationships, the type of roles that occur in this particular instance could be indicated. For example, let:

1 = Principal subjects
2 = Independent variables, causes
3 = Dependent variables, effects
4 = Passive recipients

Then the above article would be indexed as follows:

WIVES (1)
ATTITUDE (3)
NAVY (4)
LIVING CONDITIONS (2)

Thus false drops will be avoided if the search strategy matches the irrelevant document at the term level, but does not match it at the term-role level.

IFP's (Institut Francais du Pétrole) PRETEXT software can handle links and role indicators, in addition to boolean, weighted, and ignore logic (Williams, 1974, p. 243).

Precision devices are expensive to apply in an information storage and retrieval system, and to be cost-effective the added input and manipulation costs involved in the use of these devices should be offset by an appreciable saving in screening time at output, that is, precision should be greatly increased (King & Bryant, 1971, p. 152). So far the justification for the use of links and roles has not been proven with a possible exception in the field of chemistry (Lancaster & Gillespie, 1970, p. 51) as indicated by the value of certain roles in the studies of Montague and Van Oot (Lancaster, 1972, p. 128). Mullison and his colleagues in the chemical field found that roles allowed higher precision, but caused lower recall, reduced indexing consistency, and increased indexing time since there is a need to apply more complex judgments to the indexing process (Lancaster & Gillespie, 1970, p. 51; Artandi, 1970, p. 146). Lancaster states that roles also complicate the term/query relationship since the "search terms will form only part of some complete interrelationship of terms used in the indexing" of a document (Lancaster, 1972, p. 128). The searcher does not know all of the roles that the search terms may play, since these roles may be governed by other terms used in indexing that the searcher knows nothing about.

Cleverdon in his 1966 Cranfield study found that the precision device of *coordination* was more effective than the use of links (Lancaster, 1972, p. 128). Coordination relates or coordinates terms at the time of indexing (pre-coordination) or at the time of searching (post-coordination). In a pre-coordinate vocabulary the terms WIVES and ATTITUDE would be intersected or combined to become the logical product WIVES' ATTITUDE, which would be the accepted indexing term in the system. The concept or term relationships are built into the indexing language itself. In a post-coordinate vocabulary such as Taube's Uniterm system (Lancaster, 1972, p. 5), the single-word indexing terms WIVES and ATTITUDE would be coordinated at the time of searching. The formal relationship among terms may be defined by means of *boolean algebra* in the search request. Most modern retrieval systems use both pre-coordination and post-coordination. Concepts that frequently occur together are combined in a pre-coordinate term, and this term then may be combined with others in the search request.

The pre-coordination device of subheadings can be used instead of role indicators to solve the problem of incorrect term relationships. They also can be used as link indicators and have proven to be extremely valuable in MEDLARS (Lancaster, 1968a, pp. 91-95). A single subheading can be used with several indexing terms in the vocabulary. For example, ATTITUDE can be pre-coordinated with the subject involved as in WIVES/ATTITUDE or ENLISTED MEN/ ATTITUDE.

Boolean logic for search and retrieval is an important tool for improving precision since it expresses the logical relationships among the concepts of a document and manipulates these concepts at the time of searching. Boolean algebra uses the following algebraic notations to express logical relationships:

+ (OR) = the logical sum
* (AND) = the logical product
- (NOT) = the negation

If a searcher wants to retrieve information on attitudes towards drugs other than marihuana, the search statement would read ATTITUDES (AND) DRUGS, but (NOT) MARIHUANA. If a searcher wants to retrieve information on attitudes towards drugs or towards alcohol, then the search statement would read ATTITUDES (AND) DRUGS (OR) ALCOHOL.

An additional recall device that should be mentioned here is *stemming* or *truncation*, a method of reducing words to root form as a means of controlling all variant forms of a word such as singular and plural forms. This device improves recall since all possible aspects of a basic concept are retrieved, but precision is reduced since the fine distinctions of meaning are lost. For example, the distinction between LIVE, LIVEABLE, LIVELY, and LIVEN would be lost if they were reduced to their root form LIVE. For the same reason loss of precision also occurs when recall is improved by controlling synonyms.

Degrees of Vocabulary Control

Most information storage and retrieval systems operate with complete vocabulary control, that is, a controlled vocabulary is used by both indexers at the input stage and searchers at the output stage. However, with the advent of computers complete texts can be stored and searched without any form of vocabulary control. The natural language of the searcher is searched against the natural language of the text and retrieval occurs where a match is made. Such a system is known as a *free text* or *natural language retrieval system*. There is a continuum from no vocabulary control through various degrees of vocabulary control. However, to effectively operate a retrieval system, some degree of vocabulary control is necessary.

Controlled Vocabulary Indexing and Controlled Vocabulary Searching. An information storage and retrieval system with a pre-controlled vocabulary is one in which a controlled vocabulary is used by both indexers and searchers. There are a number of advantages to a system with such complete vocabulary control. The first advantage is that a particular concept is always represented in the same way in the system, thereby reducing the intellectual burden on the searcher who is seeking all relevant information on a particular topic. He is spared the task of trying to think of all possible terms that might be used to express the subject in which he is interested. For example, if a searcher is interested in the subject, marihuana, with a controlled vocabulary he need not think of all the other terms that might have been used in relevant documents, such as cannabinols, cannabis, hemp, hashish, hash, bhang, grass, weed, pot, Mary Jane, and Acapulco Gold. Recall of relevant documents would decrease if he did not think of all possible terms. Since it is not likely that a searcher will think of all possible relevant terms, a controlled vocabulary increases recall by retrieving a larger proportion of relevant documents.

Another advantage of the controlled vocabulary is that it facilitates generic searches since the relationship is shown between the broader, more generic concepts and the related narrower, more specific terms. For example, if a searcher is interested in all the documents in a system that relate to cannabinols, through a hierarchical structure or a system of cross-references he will be led from CANNABINOLS to the more specific types of cannabinols such as CANNABIS, CANNABINOL, SYNHEXYL, and TETRAHYDROCANNABINOL.

There are a number of disadvantages in the use of pre-controlled vocabularies. Controlled vocabulary indexing costs are the largest expense in an information storage and retrieval system since controlled vocabularies tend to be expensive to construct, to use, and to update. While cost of construction can be partially decreased by the use of a controlled vocabulary already available in the subject area of interest (Jahoda, 1970, p. 27), the cost of indexing itself is high. To quote Jonker (1964), "...the cost of indexing, the intellectual work of assigning descriptive 'terms' to an item of information, is usually very much higher than the cost of the clerical work of entering the data into the system or the cost of an adequate machine retrieval system" (p. 12).

Another disadvantage of the controlled vocabulary is that it tends to be nonspecific in relation to the natural language of the document and to the natural language of the searcher. Therefore, while the controlled vocabulary increases recall, it can cause precision failures related to specificity by retrieving some documents that are irrelevant to the user's needs. Consider the example of a searcher who is only interested in documents about SYNHEXYL; however, the vocabulary of the system is not specific enough to allow a depth of indexing to this level. Therefore, instead of the vocabulary uniquely identifying this term, all information on this subject becomes indexed under the more general term CANNABINOLS. The search must be conducted on the term CANNABINOLS which retrieves information on SYNHEXYL, but also retrieves unwanted documents on cannabis, cannabinol, and tetrahydrocannabinol.

A third disadvantage of a controlled vocabulary is the difficulty and time involved in learning it in order to use it effectively. The more the system is dependent on a carefully controlled vocabulary and the more complex are the indexing conventions, the more difficult it will be for the indexer and searcher to learn how to use it. Information specialists often are required to translate search requests into the system's vocabulary in order to make the use of large systems more effective. A natural language is not dependent upon indexing conventions and does not need to be learned, and therefore is easier for the user to apply in search strategies if he is not concerned with finding all of the relevant documents in the system. However, the difficulty for the user in using a controlled vocabulary can be overcome by the use of a large *entry vocabulary* which translates his natural language terms into their equivalents in the controlled language of the system.

An entry vocabulary provides entries into the system from various natural language terms that do not appear in the controlled vocabulary (Lancaster & Fayen, 1973, p. 247). Synonyms and more specific terms will lead the user to the appropriate controlled term in the vocabulary, thereby increasing the chance of coincidence between the language of the searcher and the controlled

language of the system. A full entry vocabulary eliminates or partially eliminates the need for an information specialist to conduct the searches. A large entry vocabulary also alleviates the potentially frustrating situation of multiple term look-ups that take place in constructing a search strategy.

In an on-line, pre-controlled vocabulary system, a large entry vocabulary serves as an important aid to the searcher. This capacity exists with systems such as MEDLARS, ORBIT, and RECON (Lancaster & Fayen, 1973, p. 218). Another aid to searchers in an on-line, pre-controlled vocabulary system would be the *explode* capacity. This capacity allows a generic search to be conducted on a complete category of terms, such as CANNABINOLS, without the need to list all of the more specific terms. The ability of a searcher to be able to interact with and browse through the controlled vocabulary and adjust his search strategy while the search is in progress is an important advantage of both manual (Jahoda, 1970, p. 14) and on-line information retrieval systems (Lancaster, 1972, p. 211-212). The controlled vocabulary can be of great assistance to a user if he has patience since concepts are always represented in the same way and a user is led to related terms, including synonyms, more specific and/or more generic terms, and terms semantically related in other ways.

Controlled Vocabulary Indexing and Natural Language Searching. In an on-line system it is possible both to index with a controlled vocabulary and to conduct a search in a user's own natural language. With a large entry vocabulary, the natural language terms can be automatically converted into system terms. This feature differs from the case of a manual system and some on-line systems where the searcher locates the accepted system term by looking up the term in the indexing language through the use of an entry vocabulary, and then uses the system term in the search. In a system that performs this conversion automatically, the searcher is spared the task of a look-up operation, and the illusion is created of a completely natural language search. However, the disadvantages of this type of system are the cost of constructing and loading a voluminous look-up operation and the fact that the entry vocabulary can never be complete. Also, because of the illusion created, the searcher may be surprised to see the search output is more generic or broader than his more specific request. An important problem with natural language searches is that of the language itself, since many words have multiple meanings which can be interpreted only in context, that is, by the surrounding text. The simplest way to solve this problem would be to display alternative meanings of words to the searcher so that he may choose the correct term.

Another possibility for conducting natural language searching is the translation of natural language questions into the controlled vocabulary of a system. Although there is still much work to be done in this area, Kellogg (1966, 1967, 1968) has described some procedures for such a translation.

Natural Language Indexing and Natural Language Searching. In a natural language system the data base input consists of the complete text of the document, a complete abstract, or a set of freely assigned indexing terms without any vocabulary control. The searcher uses his own language as the search input. The arguments in favor of such a system are that the cost of indexing is minimized; the documents are processed and ready for retrieval in much less time; there is complete specificity; there is no need for a vocabulary look-up

operation; and the language of the documents is more likely to match the language of the searcher who is probably a subject specialist from the same scientific community from which the documents emerged. Probably the best known commercially available natural language system is Data Central (Lancaster & Fayen, 1973, p. 37). Most of the advantages inherent in a natural language system have concomitant disadvantages.

The first requirement of a natural language system that stores and processes complete or abstracted document text is that the data base be in machine-readable form, which is a costly process. There is no indexing cost and the user interested in some, but not all, information on a particular topic can obtain it quickly since the time required to input the information is less when the need for indexing is eliminated. If the document file consists of freely assigned indexing terms, the time and cost of indexing is still minimal when compared to the intellectual effort and considerations that occur with the use of a controlled vocabulary. The free assignment of indexing terms---sometimes called *keyword indexing*---can be done manually by relatively unskilled indexers and can be converted into an automatic operation, referred to as *automatic indexing*. There is no vocabulary control exerted at the indexing stage and essentially the indexing language is the natural language of the text. All words in the document, which are in the natural language of the document, would become indexing terms except for the words on a *stop list*. The stop list eliminates from further processing the common words that do not in themselves indicate subject matter and consists of pronouns, articles, prepositions, conjunctions, copula and auxiliary verbs, and quantitative adjectives. Typically, the stop list will reduce the vocabulary of a document approximately 50 percent. In addition to the stop list, H. P. Luhn developed other early techniques for automatic indexing. Words occurring infrequently (for example, occurring only twice in the text) also can be removed as indexing terms (Lancaster, 1972, p. 154).

While free term indexing eliminates human indexing errors and more importantly is a less costly indexing method, it creates errors caused by language ambiguity that only human judgment can interpret accurately and, therefore, increases the searching costs since there is now more weeding out of irrelevant documents at the output stage of retrieval. Psycholinguistic studies indicate that humans can select a particular sentence interpretation only on the basis of probabilistic cues that result from linguistic and environmental contexts. There is no computer system capable yet of operating at this level of subtlety.

It is argued that there is less information loss in a natural language system since the entire document may be stored and available for searching (Lancaster, 1972, p. 139). However, the input costs are high if the data base is not already in machine-readable form; and if there is no vocabulary control, some relevant information may never be retrieved. A further advantage claimed for a natural language system is that there is no *vocabulary lag* since a new word will get into the system as soon as it appears in the literature. This advantage may serve the needs of searching current materials; however, for retrospective searching there is the problem of having one concept represented by more than one term.

A major disadvantage of having the document file consist of the full text of documents or an inverted subject file* of freely assigned indexing terms is that the file size can become large and unwieldy in large natural language data bases (Brandhorst & Eckert, 1972, p. 382). A large file is one which consists of 200,000,000 characters or more. As found in a study by Janning (Borko, 1967, pp. 49-50), a free vocabulary grows to a point where some control becomes necessary.

The natural language system allows complete specificity since the subject matter of the document is expressed in the document by the most specific language possible. This specificity enables the searcher to retrieve information on a precise topic rather than on the broader concept which might be the level of specificity of a controlled vocabulary. However, this feature may create a problem of overspecificity for the searcher who is interested in the broader aspects of a subject. The problem of the natural language system is that it is impossible to conduct generic searches since there is not a controlled vocabulary that has related or grouped the broader concept with its more specific counterparts. Generic searching must be conducted by the *ad hoc* summation of search terms which is very time-consuming and creates an intellectual burden for the searcher.

One prevalent problem in natural language systems is precision failure due to false coordinations and incorrect term relationships (Lancaster, 1972, p. 113; Lancaster & Fayen, 1973, p. 147). False coordination occurs when two terms that are retrieval keys turn out to be essentially unrelated in the document though they both occur in the document. The more words there are to search against, the greater the chances are for false coordination. Incorrect term relationship is also a cause of false drops in a natural language system, that is, the two terms used as retrieval keys are directly related in the documents retrieved but not in the way that the searcher wants them related. For example, if a searcher is interested in wives' attitudes and he searches on the term WIVES and ATTITUDE, irrelevant documents will be retrieved that do not deal with wives' attitudes though they are about attitudes and are also about wives. This problem also occurs in Uniterm systems which are free of the pre-coordination of indexing terms since the vocabulary consists only of single-word indexing terms. One method that is used to partially alleviate this problem is the use of *word distance indicators* (Lancaster & Fayen, 1973, p. 251) which substitute for links, roles, pre-coordination of terms, and other similar devices in more conventional systems. Word distance indicators permit text retrieval only under specified conditions, such as when the two terms appear (or co-occur) in the same paragraph, the same sentence, or within a certain number of words of each other.

Another source of precision failure in a natural language system is caused by the exhaustivity of topics represented in the document file. If a topic's absence is noted in a document or if a topic is mentioned only briefly, the irrelevant document still will be retrieved when a search is made for this topic. This failure can be remedied somewhat by *weighting*, a procedure

*An *inverted subject file* groups together all those documents or records that have a particular subject in common.

which provides that a term must occur a minimal number of times in a document before the document will be retrieved. Therefore, documents that make slight reference to a topic would not be retrieved even though in some cases the document may still be relevant. Since there is no application of human judgment as to the importance of a certain topic, these types of recall failures will occur.

Although the searcher is spared the translation or thesaurus look-up process, the greatest disadvantage of a natural language system is the intellectual burden placed on the searcher who is more interested in recall than in fast turnaround. Recall failure is one of the weakest features of a natural language system (Brandhorst & Eckert, 1972, p. 382). The choice of search terms is not limited by a controlled vocabulary and a searcher must think of all possible ways of expressing the topic of interest by considering synonyms, variant spellings, broader and narrower terms, and related terms. The time and energy expended to think of all possible approaches to retrieval can be considerable, and failure to do so is the most significant cause of recall failure, a finding from the EARS evaluation (Lancaster & Fayen, 1973, p. 252).

Natural Language Indexing and Controlled Vocabulary Searching. In the 1960's there was a peak of interest in thesaurus construction, and in the 1970's there was much activity in the area of natural language text processing by computer (Brandhorst & Eckert, 1972, p. 384). However, some vocabulary control or a thesaurus may be as essential for natural language-based systems as it is for systems that have controlled vocabulary indexing at the input stage (Brandhorst & Eckert, 1972, p. 382). This controlled vocabulary would be an aid for searching the stored natural language data base and can be stored internally in the computer or be external to the system in printed form. The search thesaurus would serve the purpose of improving recall and reducing the inverted file size by controlling synonyms, grouping related terms together in order to facilitate generic searches, and handling variant names and spellings. Such a thesaurus is referred to by Lancaster as a *post-controlled vocabulary* (Lancaster & Fayen, 1973, p. 252).

The post-controlled vocabulary mitigates two of the weakest features of natural language systems. It lessens the intellectual burden on the searcher to consider all possible search approaches, and it improves recall by enhancing the search strategy. At the same time the advantages of a natural language system can be retained, that is, specificity of the data base allows highly precise or specific searches, and the language of the data base is very similar to the language of the subject specialist searcher. However, the problem of language ambiguity still would not be controlled by human judgment at the input stage.

If the thesaurus is stored internally in an on-line system, a searcher can request that all synonyms for his input term be automatically used in the search. He also could incorporate all related terms into his search strategy without the need to key in each separate term. For example, a search could be made on all the specific cannabinols without keying in each separate type of cannabinol. It should be remembered that a search thesaurus is not used for indexing but only for constructing search strategies. It is not actually involved in the control of the system's vocabulary, but is used merely to bring related words together to help in the construction of the search strategy.

A search device which groups related words in natural language systems is stemming or word truncation where the search is made on a portion of a word only (the word root). Left truncation is permitted as well as right truncation. For example, in the Pittsburgh system (Lancaster, 1972, p. 139), a search conducted on ABANDON# (where # signifies that the search does not care which letters follow the word root) would pull in the following words: ABANDON, ABANDONS, ABANDONING, ABANDONMENT, and ABANDONED. Truncation has been discussed previously in the subsection *Vocabulary Control and Indexing Language Devices*.

Previously, weighting and word distance indicators were described as precision devices that are used in the search strategies of natural language systems. As an additional precision tool *automatic syntactic analysis* (sentence parsing) programs exist that determine structural dependencies between words in a sentence. However, natural language systems have operated successfully with devices no more sophisticated than word distance indicators and weighting. "In the last five years there has been an evident trend toward simplification in information retrieval.... Generally speaking, syntactic analysis has proven to be unnecessary and unjustifiable in terms of cost-effectiveness. The same can be said of roles and other relational indicators" (Lancaster, 1972, p. 145).

The construction of a search thesaurus requires a considerable amount of intellectual effort and thus is quite costly. However, such a search aid raises the performance level of natural language systems to such a degree that it compensates for the cost of its construction, and also lessens the intellectual and time burden placed on the searcher. The aid of the computer could be enlisted in constructing a search thesaurus, sometimes referred to as a *growing thesaurus* (Lancaster & Fayen, 1973, p. 255). After allowing the system to operate for a while without any searching aid, the search strategies that have been conducted can be analyzed. These collected data can be of use in constructing the search thesaurus. Since the users are collaborating in the building of the thesaurus, it will best suit their needs. In Salton's SMART system (Salton, 1971), much work has been done on the development of post-controlled vocabularies for use with natural language, on-line retrieval systems where the thesauri are humanly prepared. The experimental SMART system developed by Salton, first at Harvard and later at Cornell University, operates on free text and will accept questions in English sentence form. The system has experimented with the use of stem-suffix cutoffs, weighting, humanly constructed synonym dictionaries, machine-generated thesauri, word discrimination lists, word distance indicators, and syntactic phrases.

Some limitations placed on the efficient use of a natural language system by Senko (Lancaster, 1972, p. 150) are the following:

- Each document, or document portion, should be less than 200 or 300 words,
- The entire data base should be less than 1,000,000 words because of the large storage requirements, and

- The document vocabulary and/or the search request vocabulary should be relatively constrained or else it will become too large and unwieldy.

An advantage of the natural language data base is that it can be transferred more easily than the indexed data base from one information center to another, avoiding many of the problems of compatibility or convertibility.

From the above discussion it can be seen that there are both advantages and disadvantages to either controlled vocabulary data bases or natural language data bases, but in all cases some degree of vocabulary control is necessary for adequate performance. Generally, controlled vocabulary indexing helps recall, minimizes time and costs at the searching stage by reducing the intellectual burden on the searcher, and has greater costs at the input stage because of the intellectual burden of indexing. Natural language input decreases input effort and costs, making data available faster, and increases specificity at the expense of an increased intellectual effort for the searcher, lowering recall, and increasing output costs by the addition of sophisticated searching aids. It should be noted that hybrid systems using controlled vocabulary indexing and natural language nonindexed data bases do exist together. Lockheed's on-line DIALOG system in the NPRDC library in San Diego is an example of a hybrid system since it has the capacity to operate on full text as well as with controlled vocabulary indexes.

The Organization of an Indexing Vocabulary

In a pre-controlled vocabulary system the ultimate effectiveness of the entire information storage and retrieval system depends upon the adequacy of the indexing language, its organization, and the consistency of its use. The indexing terms are the descriptive labels that characterize the subject content of the stored information. The indexing terms act as access points or retrieval keys to the stored information that they describe. The ultimate goal of indexing is to improve the effectiveness of retrieval. The indexing term can take the form of a combination of words---also called a *subject heading*, as in a pre-coordinate vocabulary, or the form of a single word---also called a *Uniterm* or *keyword*, as in a post-coordinate vocabulary. An indexing language is composed of a complete set of indexing terms that lead to the appropriate indexing term. The size and composition of the indexing language are much more critical to the performance of a retrieval system than the organization structure. However, the terms in an indexing language must be organized and displayed in a way that is useful to both the indexer and the searcher. The basic forms of displaying an indexing language are the classification scheme, the alphabetic subject index, and the thesaurus.

The Classification Scheme. The historical rationale behind classification theory is to organize all knowledge into groups of related items where the related items are brought together by carefully applied principles of division. When the theory of classification is applied to indexing, the indexing terms chosen to describe the items of information usually are selected before any indexing is begun. The relationship among the selected indexing terms is set in a rigid fashion and arranged in a logical order

starting from the very general and going to the specific, with the specific always being included in the general. Some examples of classification schemes are Ranganathan's Colon Classification, Dewey's Decimal Classification, and the Universal Decimal Classification. See Figure 18 for a sample display of a classification scheme. However, with the expansion of knowledge and the concomitant diffuseness of meaning of an item of information, the clear and logical distinctions between classes and groups became more difficult. Usually the classes or terms were selected a priori or before any indexing actually began. The classes were forced upon a subject field. With the recognition that an item of information has a diffusion of meaning and can be looked at from a different point of view depending on user needs and areas of interest, the rigidity of the traditional classification theory was not practical. The current use of word classification is any method creating relations, generic or other, between individual semantic units, regardless of the degree of hierarchy, or more simply it is the putting together of like things. The classes (or groups) determined should be recognizable and nameable. Therefore, word classification now is applied to any indexing language in which the relationships between terms are indicated. The advantage of the hierarchical form that results from the classification process draws attention to omissions and incomplete hierarchies. It also imposes a discipline that helps to avoid loose terminology.

In recent years work has been done on the *automatically derived class* which also can be referred to as *mathematical classification* (Richmond, 1972, pp. 87-88; Prywes & Smith, 1972, pp. 141-152). These classes will not be as clearly defined as humanly derived classes but still can be useful for search purposes. Automatic classification is done a posteriori and relies exclusively on the content of the data base itself. Decisions and processing occur only after the data base is made available (Lang & Zagorsky, 1973). Automatic classification groups related terms together on the basis of statistical characteristics of these terms. Class membership is determined by certain strengths of positional relationships among the class members and no question of meaning arises during the procedure. *Clumping* and *clustering* are two examples of mathematical classification techniques. While automatic classification is not a difficult methodology to utilize, it is found generally to produce results inferior to a humanly constructed thesaurus (Lancaster, 1972, p. 159). Salton has concluded in his work on the SMART project that a manually constructed thesaurus performs better on the same collection than any thesaurus using automatic or semiautomatic term grouping (Summit & Firschein, 1974, p. 316).

The Alphabetic Subject Index. The alphabetic subject index is one in which the indexing terms are arranged in a single alphabetic sequence. Since the arrangement of the indexing terms is alphabetic, related information tends to be dispersed throughout the index. In order to facilitate generic searching, related terms are brought together by means of the *see also* reference. However, the *see also* reference lumps together all types of term relationships---the broader, narrower, and other term relationships. The *see also* reference provides for a sort of hidden classification. The *see* reference is used to control synonyms. Traditionally, the indexing terms in the subject index were referred to as subject headings and included both single words and phrases. Terms consisting of two or three words were pre-coordinated into

Lal	Systems of instruction (Monitorial etc.)
Lam	Lessons
Lan	Rote
Lap	Dictation
Lar	Recitation
Lav	Lecture, oral instruction
Leb	Discussion
Led	Seminar
Lef	Tutorial
Leg	Coaching
Lej	Demonstration
Lel	Experiment
Lem	Direct method
Len	Heuristic method
Lep	Group teaching
Les	Short course
Let	Part-time course
Lev	Correspondence course
Lew	Training methods
Lex	Apprenticeship, Student assistantship
Ley	In-service and on-the-job training
Lib	Teaching aids
Lid	Textbooks
Lif	Models
Lig	Toys
Lij	Laboratory equipment, apparatus
Lil	Animals
Lip	Plants, gardens
Lir	Walks
Lit	Journeys, travel
Liv	Visits
Lix	Apparatus (for Physical Education)
Lob	Audio-visual aids
Lod	Blackboard
Lof	Flannelgraph
Log	Prints, drawings
Loj	Pictures, posters
Lol	Photographs
Lom	Slides
Lop	Film, cinema
Lor	Radio and television
Los	Radio
Lot	Television
Lov	Closed-circuit
Low	Gramophone
Lox	Tape recorder
Loy	Teaching machines, automation, self-instruction
Lum	Museums) In schools and colleges; for
Lus	Libraries) Educational Documentation see Bux

Figure 18. A Sample Page from *The London Education Classification.*

single concepts and could be meaningful units on their own. When post-coordinate indexing developed in the 1950's, the tendency was to limit multi-word phrases or terms. Now that many information storage and retrieval systems utilize both pre-coordinate and post-coordinate concepts, the differences between the types of indexing language organization have become largely those of convention. Figure 19 illustrates two types of alphabetic subject indexes, one which does not indicate any term relationships and one which shows some term relationships.

The Post-Coordinate Index. Prior to 1940, information retrieval systems were essentially pre-coordinated systems, with either classified or alphabetic subject indexes. In the 1950's post-coordinate systems began to emerge. Post-coordinate systems essentially were simple to apply, with little or no indexing skill and subject knowledge needed. Their main attraction was that a machine could perform these indexing operations referred to as *word indexing*. In this technique the indexing units are not selected prior to the preparation of the index entries for specific documents, and the relationship among the indexing units is not indicated. Taube advocated in his Uniterm system that the indexing terms should consist of single words extracted from the text of a document. Very little or no vocabulary control was used. There are other examples of these types of post-coordinate indexes such as concordances and KWIC indexes. *Concordances* are alphabetic indexes of all the words in a document shown in their exact context (Kent, 1971, pp. 98-109 & 217-218).

In the original form of the *KWIC (Key-Word-In-Context)* or *permutation indexes*, the indexing terms were keywords limited to the words contained in the *title* of the document. KWIC indexes currently list each substantive word or keyword in any corpus of text, and each keyword becomes an entry point in a printed index. The keyword is positioned successively, that is, in order of every occurrence in the corpus of text, in a fixed position which is the center of one index line. The access key or keyword is surrounded by the other words in the text to form one index line (see Figure 20). The advent of the KWIC index stimulated some progress in making the titles of scientific papers more informative. However, titles and abstracts still do not receive the same critical scrutiny by editors and reviewers as the rest of the paper (Gannett, 1973, p. 245). The *KWOC index (Key-Word-Out-of-Context)* only differs from the KWIC index in format wherein the keyword is positioned at the left-hand margin of the index line. The *KWAC index (Key-Word-Augmented-in-Context)* is another variation of indexing from the corpus of text alone. Additional indexing words are added by subject specialists. In KWIC-type indexes and in natural language search systems, more information is provided at the access point since it is surrounded by the text of the document or the text of the title in which it appears.

The Thesaurus. Although the above types of word indexes have their uses and require little intellectual effort, the retrieval performance suffers from the same problems as any other system with a lack of vocabulary control. The thesaurus was born when it was realized that the post-coordinate retrieval system might benefit from careful vocabulary control. Because it grew out of post-coordinate systems, some of the indexing terms are meaningless on their own. However, many of the concepts have been pre-coordinated and resemble the traditional subject heading. The major display of a thesaurus lists the

Sample of the Alphabetic Section from the American Psychological Association's *Thesaurus of Psychological Index Terms*. (Note that there is no indication of the relationship between indexing terms.)

Attack Behavior
Attempted Suicide
Attendants (Institutions)
Attention
Attention Span
Attitude Change
Attitude Formation
Attitude Measurement
Attitude Measures
Attitude Similarity
Attitudes/ (see Figure 21)
Attorneys
Attribution
Audiences
Audiology
Audiometers
Audiometry
Audiotapes
Audiovisual Communications Media
Audiovisual Instruction
Auditory Cortex
Auditory Discrimination
Auditory Displays
Auditory Evoked Potentials
Auditory Feedback
Auditory Hallucinations
Auditory Localization
Auditory Masking
Auditory Neurons
Auditory Perception
Auditory Stimulation
Auditory Thresholds
Aunts
Aura
Aurally Handicapped
Australia

Sample of an Alphabetic Subject Index from the Armed Services Data Section of the *NAVY MEDISTARS Manual of Indexing Terms*. (Note that there is some indication of the relationship between indexing terms.)

AIR FORCE HISTORY
ARMED SERVICES AWARDS
ARMY HISTORY
COMBAT INJURIES
COMBAT KILLINGS
COMBAT REACTION — SA TRANSIENT SITUATIONAL DISTURBANCES, NAVY BUMED
DISCHARGE, ARMED SERVICES
DISCIPLINARY ACTION, ARMED SERVICES
ENLISTMENT
FAMILY SEPARATION REACTION, ARMED SERVICES
JOB DUTIES, ARMED SERVICES
JOB PERFORMANCE AND SATISFACTION, ARMED SERVICES
LEAVE — SA VIOLATIONS, ARMED SERVICES
LENGTH OF SERVICE
NAVY HISTORY

Figure 19. Samples of Alphabetic Indexes.

CONTEXT	KEYWORD	CONTEXT	CITATION	DOCUMENT NUMBER
	INFORMATION	ANALYSIS AND RETRIEVAL.	KENT71	345
STRUCTURED	INFORMATION	FILES.	KAIM73	560
METHODS OF	INFORMATION	HANDLING.	BOUR63	019
	INFORMATION	MANAGEMENT SYSTEM.	BARN73	585
SECURITY OF	INFORMATION	PROCESSING.	BORU72	404
VOCABULARY CONTROL FOR	INFORMATION	RETRIEVAL.	LANC72	449
MONOLINGUAL THESAURI FOR	INFORMATION	RETRIEVAL.	UNES7L	316
T SYSTEMS FOR STRUCTURED	INFORMATION	RETRIEVAL.	CRON74	676
	INFORMATION	RETRIEVAL ON-LINE.	LANC73	527
	INFORMATION	RETRIEVAL SYSTEMS; CHAR	LANC68	131
S FOR THE DEVELOPMENT OF	INFORMATION	RETRIEVAL THESAURI.	COSA67	094
ANNUAL REVIEW OF	INFORMATION	SCIENCE AND TECHNOLOGY.	CUAD66	047
THE EVALUATION OF	INFORMATION	SERVICES AND PRODUCTS.	KING71	285
LISTAR, THE LINCOLN	INFORMATION	STORAGE AND ASSOCIATIVE	ARME70	248
	INFORMATION	STORAGE AND RETRIEVAL.	BECK63	011
	INFORMATION	STORAGE AND RETRIEVAL S	JAHO70	207
RED ATTITUDINAL RESEARCH	INFORMATION	SYSTEM.	RAMS75	783
NTEGRATED, USER-ORIENTED	INFORMATION	SYSTEM.	ELLI72	487
A DATA FORMAT FOR	INFORMATION	SYSTEM FILES.	ANZE71	292
AN ON-LINE	INFORMATION	SYSTEM FOR MANAGEMENT.	DUFF69	164

Figure 20. Portion of a Permuted KWIC Index to Titles in the Information Storage and Retrieval Field.

indexing terms, also referred to as descriptors, in alphabetic order. Non-accepted terms appear in the alphabetic sequence with *use* references to the accepted terms. Generic relationships are indicated by BT and NT references. RT references bring together terms that are related in ways other than the generic-specific type of relationship. See Figure 21 for a sample display of a thesaurus. There usually is a display alternative to the alphabetic one in which a fully systematic hierarchy is presented. By the use of notations or codes, a user can look up a term in its hierarchical arrangement from the alphabetic display. The alphabetic display may not carry the BT and NT terms under any one term any deeper than a single level in the hierarchy, but it may show deeper hierarchical relationships also.

Graphic Displays. There also are graphic displays of thesauri such as circular or arrowgraph displays (Lancaster, 1972, pp. 55-65). These visual displays bring related terms into physical proximity and allow the indexer or searcher to view a complete group of these associations at a glance. However, large hierarchies involving multiple relationships and levels are difficult to display in a graphic form.

Citation Indexes. The types of access points to the documents stored in an information system usually are author, document date, serial number, and subject content. The access to the subject content of stored information has been reviewed because this is the area of most confusion and presents the most options and problems in an information storage and retrieval system. Accessing stored information by author, document date, or serial number is relatively straightforward. One type of index that has the author of cited documents as the access point is called a *citation index*. The citation index is mentioned briefly here because of frequent reference to it in the literature on indexing.

The citation index lists articles that have referred to a particular paper, therefore, going both forward and backward in time (Borko, 1967). A researcher can review all pertinent previous work relating to his subject of interest. The rationale behind the citation index is that references cited in a paper of interest to a user are likely to be on the same or a related subject. It should be understood that this type of index differs from the subject index since coding the elements of the citation is a clerical operation and there is no necessity to read or interpret the subject matter (Becker & Hayes, 1967, pp. 136-137). Little intellectual effort is required to construct this index.

The Preparation of a Controlled Vocabulary Index

The four basic steps in the construction of a controlled subject vocabulary index (Lancaster, 1972, p. 27) are as follows:

1. Identify the precise subject matter to be covered,

2. Select the appropriate terms to describe this subject area,

Attitude Tests	
Use	Attitude Measures
Attitudes/	
Used For	Opinions
Related	Alcohol Drinking Attitudes
	Attitude Change
	Attitude Formation
	Attitude Measurement
	Attitude Similarity
	Attribution
	Childrearing Attitudes
	Community Attitudes
	Consumer Attitudes
	Counselor Attitudes
	Death Attitudes
	Drug Usage Attitudes
	Employee Attitudes
	Employer Attitudes
	Family Planning Attitudes
	Handicapped (Attitudes Toward)
	Hedonism
	Job Applicant Attitudes
	Marriage Attitudes
	Occupational Attitudes
	Parental Attitudes
	Political Attitudes
	Prejudice
	Psychotherapist Attitudes
	Public Opinion
	Race Attitudes
	Religious Beliefs
	Sexual Attitudes
	Socioeconomic Class Attitudes
	Stereotyped Attitudes
	Student Attitudes
	Teacher Attitudes
	Work (Attitudes Toward)
Attorneys	
Used For	Lawyers
Related	Law Enforcement Personnel
	Personnel/
Attraction (Interpersonal)	
Use	Interpersonal Attraction
Attribution	
Broader	Social Behavior
	Social Perception
Related	Attitudes/

Slash (/) identifies *array* terms. These terms represent an array of conceptually broad subject matter and may be used for indexing and searching when a more specific term cannot be used to represent the subject matter. *Narrower* term relationships have not been illustrated in the above sample.

Figure 21. A Sample of the Relationship Section of the *Thesaurus of Psychological Index Terms* of the American Psychological Association.

3. Decide upon the exact form of these indexing terms, and

4. Organize and display the terms in a useful format.

The form of the indexing terms, their organization, and display have already been discussed in previous subsections that dealt with vocabulary control and organization. A consideration of user information needs has been recognized as being of prime importance in the design of a particular information storage and retrieval system. The subject matter to be covered by the stored information usually is generated by the individual needs of a particular group of users. The selection of the appropriate terms to describe this corpus of subject matter also should be based on user needs.

Since retrieval is achieved when there is a match made between the indexing term representing an item of information and the search term representing the interest or need of the user, the question is whether the indexing vocabulary should reflect the language of the stored documents or the language of the user. The simplest systems exist where the generators of the information (the authors of the documents), the indexers, and the users all belong to the same profession. Then the professional terms found in the text tend to be in coincidence with the terminology of the user.

Therefore, there are two different approaches that could be taken in choosing the indexing vocabulary. One approach would be to generate the vocabulary on the basis of the language of the documents in the collection. The other approach is to base the indexing vocabulary on the language of the user. If the information system is to serve users effectively, it is more important that the vocabulary be based on the language of the users (Lancaster, 1972, p. 32). However, since the vocabulary also should reflect the subject matter of the stored information, it is important that a wedding should occur between the document characteristics and the request characteristics representing the interests of potential users. This marriage can be achieved by analyzing user requests, and by consulting the document collection, dictionaries, glossaries, thesauri, and other publications in the area of the subject interest of the users. A committee of subject specialists or of users can be consulted in the choosing of the indexing terminology. Involving potential users in the construction of a vocabulary also serves to generate user interest in an information system and, therefore, helps to promote acceptance.

Some aids in the construction of indexing vocabularies are the ANSI standard on thesaurus construction (American National Standards Institute, 1971), Vickery's (1960) procedures of constructing a classification scheme, the COSATI (1967) guidelines, the Department of Defense (1967) manual, and the Office of Education (1969) rules. Other useful references may be found in Lancaster's book on vocabulary control (1972, pp. 37 & 76).

Batten (1973, p. 45) has suggested several guides to thesaurus construction such as UNESCO's (1971) guidelines. One manual that Batten suggests as being helpful is written by Aitchison and Gilchrist (1972). Soergel's (1974) book is a very thorough work that devotes Part III to a detailed description

of constructing indexing languages and thesauri. However, since thesaurus construction is a very costly process, the adaptation of an authoritative thesaurus already in existence would be advisable if one is available in the subject area of interest. For example, the *Thesaurus of Psychological Index Terms* (1974) published by the American Psychological Association might be useful for adaptation in constructing a thesaurus on attitude and motivation research.

The advantages and disadvantages inherent in generating an index or in indexing by computer have been discussed in the subsection on natural language indexing and the keyword approach to indexing. There is no system which is truly or fully based on automatic indexing or automatic index preparation that can be regarded as an operating system (Lancaster, 1972, p. 153). This assessment excludes the KWIC and KWOC indexes which automatically produce a permutation on words occurring in the document titles but go no further than this in representing the actual subject content of documents. The goal and appearance of an automatically produced index are not similar or equivalent to a humanly constructed index since the information processing capabilities of data processing equipment do not equal those of the human (Borko, 1967, pp. 99-100). A machine cannot yet match the human capability to operate effectively with the imprecision and ambiguity of the English language (Wyllys, 1967, pp. 168-172). Automatic indexing is an extraction process of substantive words from text, and the vocabulary is constructed by the criterion of frequency of occurrence, either absolute frequency or relative frequency (Lancaster, 1972, p. 155). Some progress has been made in automatic language processing such as in the areas of indexing, classifying, and abstracting, and there are some fruitful areas of application; however, much more needs to be done (Borko, 1967, p. 123).

Cost Performance Aspects of Indexing and Vocabulary Variables

Indexing and the indexing vocabulary are two subsystems of an information storage and retrieval system and both have an effect on the performance and cost of the system. Both can influence and be a source of system failure. The coverage of the data base and the completion of the collection also affect the quality of the system but these variables are independent of indexing and the indexing vocabulary. In evaluating a retrieval system the relationship between the level of performance or effectiveness of a system and the cost involved in achieving this level is important. The indexing and vocabulary variables influence some of the performance criteria and the costs of a system. These interrelationships are illustrated in Table 3. The performance criteria can be measured in terms of how effectively a system is satisfying its objectives or user needs. The cost is affected by the material/equipment/overhead, professional vs. clerical manpower, and input/output processes.

Input vs. Output Costs. The input costs are the largest costs of an information system and most of this cost factor occurs at the beginning or start of building the system. Input costs are related to the time and effort of designing or preparing the vocabulary, indexing the data base, and developing and applying the maintenance or updating procedures. The output costs

relate to the time and effort involved in designing search strategies, searching the system, and screening or filtering the output for the relevant documents. A generalization which usually is true is that there is an inverse ratio between input cost and output cost (Jahoda, 1970, p. 25). Almost invariably economies in one area increase the burden on the other. Depending upon user needs and therefore upon the system's requirements, a balance between input and output costs will have to be determined. Several factors influence the decision of whether to put the emphasis on the input processes or on the output processes. Two factors to consider are the volume of documents to be indexed and the anticipated volume of requests. If there are many documents to be indexed and few requests expected, economies of input would be a more rational trade-off. If there are few documents in a system and many requests expected, then output economies should be employed with more emphasis placed on the accuracy of the input processes. As King and Bryant (1971, p. 92) have pointed out, the input accuracy (which is affected by the emphasis on input processes) directly affects the accuracy of the system's performance.

Another factor influencing the input/output ratio is the requirement of speed. If input speed is important (as it is in the case of the current awareness needs of intelligence work where documents must get into the system as rapidly as possible), then indexing economies should be adopted. However, if output speed and rapid response are critical or more important factors (as in the case of a Poison Information Center), no economies at input are justified since there would be delayed response and reduced accuracy of output (Lancaster & Fayen, 1973, pp. 397-400).

The output or search cost is related to the response time, which is also a measurement of the performance level. Response time involves more than the actual search time, such as (1) the time that a potential user waits for a particular data base to be made available and (2) computer speed. This part of the response time is not affected by the indexing vocabulary. The portion of response time affected by the indexing methods and the indexing vocabulary is the time spent in designing the search strategies, searching the system, and screening the output for relevant documents. However, within reason, response time is rarely a prime user requirement and is always secondary to the recall and precision requirements (Lancaster & Fayen, 1973, p. 128).

Recall and Precision Ratios. For an efficient system the *filtering* process should be done by the system itself and not by the searcher. There are two measures which are used jointly to indicate the filtering capacity of the system and, therefore, are good measures of system effectiveness---the recall ratio and the precision ratio (Lancaster, 1972, pp. 107-108; Jahoda, 1970, pp. 22-24). The recall ratio is the quantitative expression of recall, which is the ability of the system to retrieve relevant documents in response to a subject request. The precision ratio is the quantitative expression of precision, which is the ability of the system to hold back nonrelevant documents. Therefore,

$$\text{Recall ratio} = \frac{\text{number of relevant documents retrieved by the system}}{\text{total number of relevant documents in the system}} \times 100$$

TABLE 3

THE EFFECTS OF INDEXING AND VOCABULARY
ON SYSTEM PERFORMANCE AND COSTS

Indexing and Vocabulary Variables	*System Effectiveness*		*Input Cost: Time and Effort*			*Output Cost: Time and Effort*		*Other Effects*
	Recall	*Precision*	*Index Construction*	*Index Use*	*Index Update*	*Search Construction*	*Search Screening*	
Highly structured, controlled vocabulary *vs.*	Higher	Higher	Higher	Higher	Higher	Lower	Lower	Higher level indexing personnel
Unstructured, uncontrolled, freely assigned keywords	Lower	Lower	Lower	Lower	Lower	Higher	Higher	Less qualified indexing personnel
Recall devices (in a controlled vocabulary) *vs.*	Higher	Lower	Higher	Higher	Higher	Lower	Higher	
Precision devices (in a controlled vocabulary)	Lower	Higher	Higher	Higher	Higher	Higher	Lower	Indexing consistency is reduced
Highly specific indexing terms *vs.*	Lower	Higher	Higher	Higher	Higher	Higher	Lower	Higher level indexing personnel; harder to achieve indexing consistency
Relatively broad indexing terms	Higher	Lower	Lower	Lower	Lower	Lower	Higher	Less qualified indexing personnel

(Continued)

TABLE 3 (CONT.)

THE EFFECTS OF INDEXING AND VOCABULARY
ON SYSTEM PERFORMANCE AND COSTS

Indexing and Vocabulary Variables	*System Effectiveness*		*Input Cost: Time and Effort*			*Output Cost: Time and Effort*		*Other Effects*
	Recall	*Precision*	*Index Construction*	*Index Use*	*Index Update*	*Search Construction*	*Search Screening*	
Indexing exhaustivity *vs.*	Higher	Lower	Higher	Higher	Higher		Higher	More terms to be matched in searching; increases search time; more chance of false coordination
Lack of indexing exhaustivity	Lower	Higher	Lower	Lower	Lower		Lower	
Large entry vocabulary	Higher							
Indexing errors: omission of important concepts	Lower				Higher			
Indexing errors: use of inappropriate terms	Lower	Lower			Higher		Higher	
Indexing policy: review indexing for quality control	Higher	Higher				Lower	Lower	Increases input costs

$$\text{Precision ratio} = \frac{\text{number of relevant documents retrieved by the system}}{\text{total number of documents retrieved by the system}} \times 100$$

or

If
- a = relevant documents retrieved
- b = relevant documents not retrieved
- c = nonrelevant documents retrieved
- d = nonrelevant documents not retrieved

Then

$$\text{Recall} = \frac{a}{a + b} \times 100$$

$$\text{Precision} = \frac{a}{a + c} \times 100$$

The precision ratio measures the efficiency with which the system is able to achieve a particular recall ratio. Often recall and precision tend to vary inversely in searching, that is, if recall is increased, precision tends to be reduced and vice versa (Lancaster, 1972, pp. 107-110). This inverse relationship is true in the case of indexing exhaustivity and vocabulary specificity, as well as in the use of recall and precision devices.

Theoretically, 100 percent recall can be achieved if all or a large portion of the documents are retrieved, placing the screening burden on the receiver of the output, but that defeats the filtering purpose of the information retrieval system. Systems must have a high degree of precision or users will become disenchanted with them and stop using them. To quote Mooers' Law, "An information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it" (Lancaster & Fayen, 1973, p. 347). Users differ in their recall and precision requirements, and the optimum balance between the two measures ultimately will be decided by user needs. Some users may be interested in a high degree of precision. In very large collections a low precision ratio might not be tolerable since this would entail the screening of a large number of retrieved documents. On the other hand, users who require most of the documents on a topic are more interested in high recall. In small collections where only a few documents are retrieved, high recall is more important and screening these few documents would not be time-consuming.

In the subsection *Vocabulary Control and Indexing Language Devices* a detailed analysis was given of how the specific devices used to increase recall and precision affect cost and performance. While recall and precision devices increase the overall accuracy of output, they also increase the cost of vocabulary preparation, indexing, and updating.

Much discussion centers around the issue of relevance since relevance is a value judgment placed on documents by individuals with information needs. The requester makes this value judgment. It is difficult to determine the true number of relevant documents contained in a system unless the requester examines all of the nonretrieved items. However, in most operating systems this examination is not practical and the best estimate of recall is made. The number of relevant documents contained in the system and not retrieved can be estimated by various methods. One method involves composing a list of documents known to be in the system through outside sources such as other information centers or published indexes. The requester then can judge this list of documents known to be in the system that are relevant to his needs and determine which have not been retrieved by the system being evaluated.

The form of the system output is important because of its effect on the precision tolerance of the user. If the system's product is in a form that facilitates rapid scanning allowing irrelevant items to be disregarded fairly easily, a user is likely to tolerate a lower precision ratio (Lancaster & Fayen, 1973, p. 129). The form of output usually is independent of indexing and vocabulary except for KWIC-type indexes and systems that search the complete text.

Controlled vs. Uncontrolled Vocabulary. The issue of the use of a controlled vocabulary in contrast to a system operating with a nonindexed data base or freely selected keywords devoid of language control has been discussed in the subsection *Degrees of Vocabulary Control*. The option is between increased input cost with associated improved output accuracy or less input cost leading to increased output cost and effort. In keyword searching, a searcher may virtually have to construct a segment of a controlled vocabulary each time he prepares his search strategy.

Specific vs. Broad Indexing Terms. An important variable that affects the cost and performance of a system is the *specificity* of the indexing terms in the indexing language (Lancaster, 1968b, pp. 68-70). Specific indexing terms allow for a precise subject representation of documents, permitting higher precision in output than a broader representation for a more specific concept. For example, PORPOISE is a more specific indexing term for the concept *porpoise* than is the broader term MAMMAL. The greater the specificity of an indexing language, the larger is the size of the indexing vocabulary and the more frequently it changes, increasing the cost of creating, applying, and updating the vocabulary. A higher level of personnel is required to apply and maintain it. A higher search precision due to the precise representation of topics saves on the searching time spent screening out irrelevant documents. The recall ratio tends to be lowered in the more specific vocabulary. One reason is that it is more difficult to achieve indexing consistency. There is more likelihood that a relevant document may have been indexed by an incorrect synonymous term. For example, the concept *porpoise* may have been indexed by the term DOLPHIN. Also, documents dealing with the concept in general are missed though they may contain useful information about the narrower concept.

In the vocabulary where subject concepts are handled by more general indexing terms, lower level personnel can be used for indexing since the vocabulary is smaller and more general. The construction of search strategies

also is simplified with a smaller, more general vocabulary. While precision may be decreased, recall may be increased through the reduction of indexing errors and inconsistencies, and by the retrieval of potentially relevant information dealt with under the broader concept. Much of the information about mammals also applies to porpoises. However, the irrelevant material on mammals, such as whales, would have to be weeded out for the searcher interested in porpoises specifically.

However, the specificity of the vocabulary must be directly related to the specificity of the requests made. It would be uneconomical and inefficient to establish a vocabulary considerably more specific than the level of specificity required by user needs. A careful analysis should be conducted to establish the level of specificity of representative requests (Lancaster, 1972, pp. 218-219).

Exhaustivity vs. Lack of Exhaustivity in Indexing. The *exhaustivity* of the indexing is another important variable affecting the cost and performance of a retrieval system. The exhaustivity of indexing is the number of indexing terms selected for each unit of information or document. As with the specificity of indexing terms, the level of indexing exhaustivity should be as exhaustive as that required by user needs. The break-even point for exhaustivity would be the "point beyond which the addition of further index terms, although adding appreciably to input costs, is not making any highly appreciable difference to the recall potential of the system" (King & Bryant, 1971, p. 155). Greater exhaustivity increases indexing time and costs, increases the average number of documents retrieved per search, and increases output screening time (Lancaster, 1972, pp. 222-223). The recall capability is increased because more concepts have been recognized and labeled. Precision is decreased since concepts that are treated in only a minor way have been indexed and are of no use to the user because they are too shallow and irrelevant. Precision also decreases because the increase in the number of indexing terms used increases the potential for false coordination (Lancaster, 1968b, pp. 66-67). Greater exhaustivity affects search time because of the increased number of terms to be matched in the searching operation (King & Bryant, 1971, p. 12).

However, a lack of indexing exhaustivity would result in concepts not being tagged that would be of interest to users, resulting in an inadequate recall ratio. Precision is increased since every document that is retrieved must deal in a substantial manner with the subject of the request, and since there are fewer indexing terms per document, the possibility of false coordination of terms in searching is decreased.

The Entry Vocabulary. The entry vocabulary, which consists of the "natural-language expressions, occurring in documents or requests, that map onto the controlled vocabulary of the system" (Lancaster, 1972, p. 219), is another important variable that is often overlooked. A large entry vocabulary may be relatively expensive to construct and update but it will have long-term benefits by reducing the intellectual burden on indexers and searchers who will not have to make the same intellectual decisions over and over again. Since the intellectual decision that was made initially by the indexer was recorded, it will not have to be made again, thus increasing indexing

consistency since the same concept will be treated in the same manner when it reappears. The larger the entry vocabulary, the fewer the intellectual decisions that will have to be made by indexers and searchers, thus reducing indexing and searching time. A lower professional level of indexing staff can be used if fewer intellectual decisions need to be made. The lower the professional level of personnel needed for indexing, the lower the salary levels will be. Since indexing consistency increases with a large entry vocabulary and search formulation is easier, recall also will be improved.

Indexing Errors and Policy. Indexing errors also affect system performance. If important concepts are missed and not tagged, the system will fail to recall them in retrieval. Also, if concepts are labeled with incorrect indexing terms, they will fail to be recalled for a search on the correct indexing term, and they will be retrieved incorrectly under the wrong indexing term. More time will have to be spent in screening out the incorrectly labeled concepts. Lancaster (1968a, p. 199) postulated in evaluating MEDLARS that a significant number of indexing omissions may be due to a lack of specific terms in the indexing vocabulary and a failure to use more general terms. Indexing errors may be decreased by an indexing policy of reviewing all indexing decisions. This practice would reduce the time necessary to weed out irrelevant material, but would increase the input costs (Bryant & King, 1971, p. 156).

Lancaster (1968a, pp. 185-203; 1972, pp. 191-203) has offered several recommendations for enhancing the performance of information storage and retrieval systems. There should be a close integration among the functions of indexing, searching, and vocabulary control. Continuous quality control by system monitoring ultimately is essential to the success of any large information storage and retrieval system. An indexing term display for the searcher is important. Abstracts or summaries are more helpful in the screening process than title tracings. *Automatic term replacement* is useful in compensating for vocabulary changes. The filing of repeated search requests is a useful searching aid. A search request form completed by a researcher without reference to the indexing vocabulary is useful to the system specialist in translating the search request into the language of the controlled vocabulary. *Postings* or *term-frequency lists*, recording the number of times a term has been used in indexing, are used to assist in developing search statements based on boolean logic. A low-frequency term may stand alone or may be related by the use of an OR operator to other terms. However, a high-frequency term should be ANDed to another term to avoid producing an excessive amount of output.

The following computer aids help in the continual assessment of the indexing vocabulary. Postings or term-frequency lists indicate those terms that have been used very infrequently and are therefore good candidates for deletion from the vocabulary. *Use-data* can be logged for a good user-oriented indexing language such as the number of requests sought under a particular term, the number of times a user enters the system with a term not in the thesaurus, and the number of times a cross-reference path is used. Another useful aid is the number of terms to which a term is related in the thesaurus, identified by TS or *thesaurus structure*.

An information storage and retrieval system can have a stimulating effect on users and researchers. If the information system adequately reflects through the indexing language the future needs of users, new areas of interest can be stimulated even though the present need for certain information was not reflected in user surveys.

Summary

The purpose of vocabulary control is to precisely relate words to an idea or concept in order to maximize the likelihood that both the indexer and the searcher will use identical terminology when referring to the content of a given document. The function of bringing the language of the searcher into coincidence with the language of the indexer is served by the controlled vocabulary. The structured, controlled vocabulary requires time and effort to construct, to use, and to maintain, while there is no construction and maintenance cost for unstructured keywords which are used freely in indexing. Indexing from a controlled vocabulary requires the selection of terms from the vocabulary which may involve a look-up operation, while free term or keyword indexing may be done by less qualified indexers who freely select keywords from the texts of documents being indexed. However, keyword searching increases the burden at the output stage and the searcher may virtually have to construct a segment of a controlled vocabulary each time that he prepares his search strategy. Recall is reduced since it is unlikely that the searcher will think of all relevant terms on which to search, and precision is reduced since language ambiguities have not been controlled at input, thus increasing the screening effort. With the use of a controlled vocabulary more effort is exerted for the input operation, increasing output accuracy, and decreasing the time and effort involved in constructing search strategies and screening output.

Vocabulary control devices and the other indexing language devices reviewed in the subsection *Vocabulary Control and Indexing Language Devices* form the syntax of the controlled indexing language and are used to improve recall or precision. In that subsection a detailed analysis was given of how the specific devices affect cost and performance. Generally, the higher the recall or precision, the higher the input costs, since indexing language syntax increases the cost of the vocabulary preparation, indexing, and updating. Roles and links have been found to be less cost effective than other precision devices such as coordination. Once again the factors influencing input economies versus output economies have to be considered. Document volume versus request volume as well as greater input speed with less output accuracy versus slower and more accurate input with faster and more accurate output would influence the decision of placing the burden at the input stage or the output stage.

In a pre-controlled vocabulary system the ultimate effectiveness of the entire information storage and retrieval system depends upon the adequacy of the indexing language, its organization, and the consistency of its use. The size and composition of the indexing language are much more critical to the performance of a retrieval system than the organization structure. However, the terms in an indexing language must be organized and displayed in a way

that is useful to both the indexer and the searcher. The basic forms of displaying an indexing language are the classification scheme, the alphabetic subject index, and the thesaurus. The thesaurus was born when it was realized that single keyword searching strategies might benefit from careful vocabulary control at the input stage. However, since thesaurus construction is a very costly process, the adaptation of an authoritative thesaurus already in existence would be advisable if one is available in the subject area of interest. The goal and appearance of an automatically produced index are not similar or equivalent to a humanly constructed index since the information processing capabilities of data processing equipment do not equal those of the human.

Most automatic information retrieval systems have found it necessary to use a great deal of intellectual effort either at the input or output stage, depending on user needs and cost limitations. The intellectual effort may be in constructing, using, and maintaining an indexing language or in the formulation of search strategies and screening output, both of which will involve compiling synonym lists and/or thesauri or classification hierarchies. There is a tendency towards simpler systems of vocabulary control at this time, since the use of very sophisticated devices has not increased benefits substantially but has increased costs considerably. Several factors, the most important being input/output volume and input/output speed and accuracy, must be considered in determining whether the input processes or the output processes are emphasized. Greater input care increases the input costs but improves output efficiency and reduces output costs. However, user needs and cost limitations determine whether greater care and expense should be put into the input operation or whether an acceptable level of performance can be attained by economizing on input costs.

An important variable that affects the cost and performance of a system is the specificity of the indexing terms in the indexing language. However, the specificity of the vocabulary must be directly related to the specificity of the requests made. It would be uneconomical and inefficient to establish a vocabulary considerably more specific than the level of specificity required by user needs.

The exhaustivity of the indexing is another important variable affecting the cost and performance of a retrieval system. As with the specificity of indexing terms, the level of indexing exhaustivity should be as exhaustive as that required by user needs. The entry vocabulary is an important variable that is often overlooked. A large entry vocabulary is relatively expensive to construct and update but it will have long-term benefits by reducing the intellectual burden on indexers and searchers who will not have to make the same intellectual decisions over and over again.

There should be a close integration between the functions of indexing, searching, and vocabulary control. Continuous quality control by system monitoring ultimately is essential to the success of any large information storage and retrieval system.

ANNOTATED BIBLIOGRAPHY

1. Aitchison, J., & Gilchrist, A. *Thesaurus Construction: A Practical Manual*. London: ASLIB, 1972.

A helpful manual on thesaurus construction that gives attention to the problems of term choice and vocabulary size. It is a result of more than a decade of vast experience, yet is less than 100 pages, has a selective bibliography, and has a useful cross-reference system.

2. American National Standards Institute. *Guidelines for Thesaurus Structure, Construction and Use*. ANSI Standards Committee Z.39 (Draft), February 1971.

This is a draft guideline for thesaurus construction. It offers a somewhat restricted choice of design based closely on the types of structures used in the Thesaurus of Engineering Terms and the Thesaurus of Engineering and Scientific Terms. This draft closely follows the LEX/COSATI guidelines and may eventually become a U.S. national standard on the subject.

3. Artandi, S. Document description and representation. In C. A. Cuadra (Ed.), *Annual Reveiw of Information Science and Technology* (Vol. 5). Chicago: Encyclopaedia Britannica, 1970. Pp. 143-167.

This review indicates that automatic document description and representation is a long way from success. A description is given on the progress made in: understanding the indexing process; determining the importance of the query in search effectiveness; exploring the role of the thesaurus as an aid to users; and achieving greater integration of indexing, query formulation, and vocabulary control.

4. Batten, W. E. Document description and representation. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 8). Washington, D.C.: American Society for Information Science, 1973. Pp. 43-68.

A review of modern trends in thesauri, classification, subject content and document representation, and catalog design. Batten also evaluates manually and automatically generated indexes.

5. Becker, J., & Hayes, R. M. *Information Storage and Retrieval: Tools, Elements, Theories*. New York: Wiley, 1963.

This book provides a foundation and structure within which developments in information retrieval and allied fields can be viewed for their relationship and interaction with each other.

6. Borko, H. Design of information systems and services. In C. A. Cuadra (Ed.), *Annual Review of Information Science and Technology* (Vol. 2). Chicago: Encyclopaedia Britannica, 1967. Pp. 35-86. (a)

A review of systems analysis concepts and of the procedures for applying these concepts to the design of information systems.

7. Borko, H. Indexing and classification. In H. Borko (Ed.), *Automated Language Processing*. New York: Wiley, 1967. Pp. 99-125. (b)

A state-of-the-art review of the methods of automatic indexing and classification and the advantages and disadvantages of automatic methods as compared to manual methods.

8. Brandhorst, W. T., & Eckert, P. F. Document retrieval and dissemination systems. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 7). Washington, D.C.: American Society for Information Science, 1972. Pp. 379-437.

A thorough review of trends, advances, and problems of computerized document retrieval and dissemination systems. Noncomputerized retrieval techniques are not covered. Included is a table listing 76 institutions with computerized document retrieval systems, such as the Navy MEDISTARS system. There is information on the type of system, the associated hardware and software, the developmental status, and other system characteristics.

9. Cleverdon, C. W., Mills, J., & Keen, E. M. Factors determining the performance of indexing terms. In *Studies in Indexing and Cataloging*. Detroit: Management Information Services, 1970. Pp. 1-424.

This classic work describes quantitative experiments on indexing variables and their effect on retrieval performance.

10. Committee on Scientific and Technical Information (COSATI). *Guidelines for the Development of Information Retrieval Thesauri*. Washington, D.C.: COSATI, 1967.

This guideline records the rules and conventions established by COSATI for thesaurus construction after the accumulation of a considerable amount of experience.

11. Department of Defense. *Manual for Building a Technical Thesaurus*. Washington, D.C., 1966. AD 633 279. (Revised and published as Appendix 1 of the *Thesaurus of Engineering and Scientific Terms*, 1967.)

This is a manual to aid in the construction of a technical thesaurus. The rules and conventions resulted from the experience accumulated by Project LEX (a project of the Department of Defense).

12. Gannett, E. K. Primary publication systems and services. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 8). Washington, D.C.: American Society for Information Science, 1973. Pp. 243-275.

This paper discusses changes in the generation, productions, and delivery of primary publications, particularly in the scientific and technical fields.

13. Harris, J. L. Document description and representation. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 9). Washington, D.C.: American Society for Information Science, 1974. Pp. 81-117.

This article deals with efforts made toward standardization in document description and to a lesser extent in representation. Some attention also is given to the new awareness that a need exists for standard bibliographic description of nonbook materials.

14. Jahoda, G. *Information Storage and Retrieval Systems for Individual Researchers*. New York: Wiley-Interscience, 1970.

This book is intended for the researcher in any field who wishes to start or improve an index to his document collection. It was a major work in 1970 dealing with manual document retrieval systems. A variety of indexes are discussed such as coordinate indexes, edge-notched cards, optical coincidence systems, and citation indexes.

15. Jonker, F. *Indexing Theory, Indexing Methods and Search Devices*. New York: Scarecrow Press, 1964.

This short and simple text is on the theory of and the different methods of indexing and of searching.

16. Kellogg, C. H. *An Approach to the On-Line Interrogation of Structured Files of Facts Using Natural Language* (SP-2431/000/00). Santa Monica, Cal.: System Development Corporation, 1966.

Kellogg provides an overview of fact retrieval systems and describes his own experimental efforts restricted to a structured data base. The approach taken is to translate English questions into data-management system file-searching procedures.

17. Kellogg, C. H. *On-Line Translation of Natural Language Questions into Artificial Language Queries* (SP-2827/000/00). Santa Monica, Cal.: System Development Corporation, 1967.

A description of procedures for the translation of natural language questions into the controlled vocabulary of a fact retrieval system.

18. Kellogg, C. H. *Data Management in Ordinary English: Examples* (TM-3919/000/00). Santa Monica, Cal.: System Development Corporation, 1968.

A description of procedures for the translation of English language questions into a highly formalized internal language of a data-management system.

19. Kent, A. *Information Analysis and Retrieval*. New York: Becker and Hayes, 1971.

A general text which fits data organizational concepts into the framework of information retrieval systems.

20. King, D. W., & Bryant, E. C. *The Evaluation of Information Services and Products*. Washington, D.C.: Information Resources Press, 1971.

This book is an excellent distillation of the experience of Westat Research Inc. in the area of evaluation of information services. A construction of a detailed system model depicts the relations among system features, costs, performance, and benefits. Guidance is given as to what to measure, how to measure, and how to interpret results. Also contains brief and lucid primers on user surveys, statistics, sampling methods, and experimental design.

21. Kinkade, R. G. (Ed.). *Thesaurus of Psychological Index Terms*. Washington, D.C.: American Psychological Association, 1974.

This thesaurus is divided into three sections, the alphabetic section with no indication of relationships, the relationship section arranged alphabetically, and the heirarchical section. Approximately 4,000 terms are included from psychology and related disciplines, but it is difficult to find the same term in each of the three sections since there is no code or notation system that guides the user from any one section to the other two.

22. Lancaster, F. W. *Evaluation of the MEDLARS Demand Search Service*. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Public Health Service, 1968. (a)

A description of an evaluation study of MEDLARS with conclusions and recommendations. These findings and recommendations for enhancing the performance of MEDLARS may be applicable to other information storage and retrieval systems.

23. Lancaster, F. W. *Information Retrieval Systems; Characteristics, Testing, and Evaluation*. New York: Wiley, 1968. (b)

A textbook with brief but clear treatment of the historical background, the conceptual and mathematical bases, and examples of applications of viable and widely applicable procedural features of information retrieval systems. Half of the book has a descriptive emphasis on human-performed procedures, while the last half is devoted to methods and concepts of system evaluation from the Cranfield and MEDLARS evaluation projects. The discussion of operational efficiency centers around search accuracy and is not taken up with cost. The economic efficiency discussion considers search accuracy, other benefit and effectiveness criteria, together with cost data, and a trade-off analysis.

24. Lancaster, F. W. *Vocabulary Control for Information Retrieval*. Washington, D.C.: Information Resources Press, 1972.

This work is a very good overall treatment on the influence of vocabulary and vocabulary control mechanisms upon retrieval performance, vocabulary compatibility, and cost-effectiveness.

25. Lancaster, F. W., & Fayen, E. G. *Information Retrieval On-Line* (a Wiley-Becker & Hayes Series book). Los Angeles: Melville Publishing Co., 1973.

This book provides a broad survey of the characteristics, capabilities, and limitations of then current information retrieval systems operated in an on-line interactive mode.

26. Lancaster, F. W., & Gillespie, C. J. Design and evaluation of information systems. In C. A. Cuadra (Ed.), *Annual Review of Information Science and Technology* (Vol. 5). Chicago: Encyclopaedia Britannica, 1970. Pp. 33-70.

This is a review of the literature on design and evaluation of information storage and retrieval systems. The emphasis is on important trends of interest to system designers and includes discussions on trends at national and international levels; complete system design discussions; and discussions on the design of individual system components, including indexing, the vocabulary, searching, and data processing.

27. Lang, A. L., & Zagorsky, S. W. *Implementation of an Automatic, a Posteriori, Hierarchical Classification System* (Technical Report No. 73-05). University of Pennsylvania, The Moore School of Electrical Engineering, Contract Nonr N00014-67-A-0216-0014 - NR 049-153, Office of Naval Research, January 1973.

This paper is a description of a semi-automatic indexing and classification system that uses a human interface with the indexing routines. There is a discussion of the theory, methods, and long-run advantages of automatic indexing and classification.

28. Office of Education. *Rules for Thesaurus Preparation*. Washington, D.C.: Educational Resources Information Center, 1969.

These rules were compiled after considerable experience by the Educational Resources Information Center (ERIC) of the Office of Education. In developing the ERIC thesaurus, the emphasis was on the use of natural language expressions that would be meaningful and recognizable to users.

29. Prywes, N. S., & Smith, D. P. Organization of information. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 7). Washington, D.C.: American Society for Information Science, 1972. Pp. 103-158.

A description is given of formal and automatic tools that help designers and users to organize and exchange information in large and complex data bases and to improve data processing. There is a review of computer languages that have been developed specifically to specify organization of information.

30. Richmond, P. A. Document description and representation. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 7). Washington, D.C.: American Society for Information Science, 1972. Pp.73-102.

This review covers the areas of descriptive analysis and representative or subject analysis of documents, with more attention given to document description.

31. Salton, G. *Automatic Information Organization and Retrieval*. New York: McGraw-Hill, 1968.

SMART is an experimental retrieval system developed by Salton, first at Harvard University and later at Cornell University. This book is a most detailed description of this experimental natural language system and of the theories underlying it.

32. Salton, G. (Comp.). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

A very detailed description of the SMART system, of the theories underlying it, and of the conclusions drawn from this very important experimental system.

33. Salton, G. (Ed.). *Reports on Text Analysis, Dynamic Indexing, Feedback Searches, Dictionary Construction and File Organization*. Ithaca, N.Y.: Cornell University, Department of Computer Science, December 1972.

The SMART Project is a major source of work in language processing techniques. The 1972 report is on the basic SMART system, interactive feedback, dictionary construction, and cluster file organization.

34. Sharp, J. R. *Some Fundamentals of Information Retrieval*. London: A. Deutsch, 1968.

This text is on conventional indexes with a bias against automatic procedures.

35. Simmons, R. F. Automated language processing. In C. A. Cuadra (Ed.), *Annual Review of Information Science and Technology* (Vol. 1). New York: Interscience Publishers (Wiley), 1966. Pp. 137-169.

One major area of this review is on computational linguistics in addition to computerized syntactic analysis and question-answering programs. Some attention is given to mechanical translation, automatic abstracting, stylistic analysis, text editing, and type composition.

36. Scergel, D. *Indexing Languages and Thesauri: Construction and Maintenance* (a Wiley-Becker & Hayes Series book). Los Angeles: Melville Publishing Co., 1974.

This text provides a comprehensive treatment of the function, structure, and construction of indexing languages and thesauri in both manual and computerized information storage and retrieval systems. It evaluates alternatives in design, and construction and maintenance procedures for indexing languages and thesauri in different situations. It is supplemented by an extensive critical bibliography and guide devices to specific subject matter.

37. Stevens, M. E. *Automatic Indexing: A State-of-the-Art Report*. Washington, D.C.: U.S. Department of Commerce, National Bureau of Standards, 30 March 1965.

An excellent and complete review of automatic indexing with a pro-automatic indexing bias.

38. Summit, R. K., & Firschein, O. Document retrieval systems and techniques. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 9). Washington, D.C.: American Society for Information Science, 1974. Pp. 285-331.

A review of the literature on systems design, modeling, and evaluation; file structure design; and the user interface. This paper also reports on experiments in search procedures and indexing, query formulation, file organization, and language processing.

39. United Nations Educational, Scientific and Cultural Organization. *Guidelines for the Establishment and Development of Monolingual Thesauri for Information Retrieval*. Paris: United Nations Educational, Scientific and Cultural Organization, 22 December 1971. ED 059 749.

These guidelines are based on the idea that a thesaurus should be constructed "according to a body of knowledge." While this book provides a valuable check list, no direct guidelines are given on how to choose descriptors. No one particular thesaurual structure is suggested.

40. Vickery, B. C. *Faceted Classification*. London: ASLIB, 1960.

A useful book that describes procedures for constructing a faceted classification. Facet analysis involves sorting candidate terms into homogeneous, mutually exclusive facets on the basis of a single characteristic of division.

41. Vickery, B. C. *Techniques of Information Retrieval*. London: Butterworths, 1970.

A textbook that deals at length with problems of document description and representation.

42. Williams, M. E. Use of machine-readable data bases. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 9). Washington, D.C.: American Society for Information Science, 1974. Pp. 221-284.

This chapter covers the 1973 as well as earlier papers on data bases, considered here to be organized sets of machine-readable records containing bibliographic or document-related data. There is no coverage of data bases whose content is primarily numeric.

43. Wyllys, R. E. Extracting and abstracting by computer. In H. Borko (Ed.), *Automated Language Processing*. New York: Wiley, 1967. Pp. 127-179.

This review describes the progress, the advantages, and the problems of automatically produced abstracts. Future possibilities for the techniques of automatic abstracting and extracting is envisioned.

GLOSSARY OF KEY TERMS

Alphabetic Subject Index: The alphabetic subject index is one in which the indexing terms are arranged in a single alphabetic sequence.

Automatic Classification: A computerized technique for grouping related terms or documents together on the basis of statistical characteristics of these entities. Class membership is determined by certain strengths of positional relationships among the class members and no question of meaning arises during the procedure. SEE ALSO: *Clumping or Clustering*.

Automatic Indexing: Automatic indexing is the application of computers to automatically reduce a document to an indexed representation.

Automatic Syntactic Analysis: Sentence parsing or the determination of structural dependencies between words in a sentence that is done by means of computer programs.

Boolean Search Logic: The use of algebraic notations to express logical relationships among indexing terms and the manipulation of these terms at the time of searching.

Citation Index: An index in which the index headings are the citations that an author makes to other documents.

Classification Scheme: Traditionally, the classification scheme organized all knowledge into groups of related items where the related items were brought together by carefully applied principles of division. The current use of word classification is any method creating relations, generic or other, between individual semantic units, regardless of the degree of hierarchy.

Clumping or *Clustering*: A mathemathical technique for the automatic grouping of "like" terms or documents.

Concordance: An alphabetical index of all the words in a text or corpus of texts, showing every contextual occurrence of a word.

Controlled Vocabulary: The vocabulary of an information system that relates words to an idea or concept; brings related words together; and controls synonyms as well as variant names and spellings.

Coordination: Relating or coordinating terms at the time of indexing (pre-coordination) or at the time of searching (post-coordination).

Entry Vocabulary: An entry vocabulary is a vocabulary of natural language expressions, occurring in documents and requests, with appropriate mappings to the controlled terms of the system.

Exhaustivity: The exhaustivity of indexing is the number of indexing terms selected for each unit of information or for each document.

Filtering: The filtering process is the ability of the system to retrieve relevant documents in response to a request while holding back nonrelevant documents.

Free Term Indexing (or *Keyword Indexing*): The free assignment of indexing terms to documents without the use of vocabulary control at the indexing stage.

Free Text Retrieval System SEE *Natural Language Retrieval System*

Growing Thesaurus: A search thesaurus constructed, with the aid of a computer, from the analysis of the search strategies used in searching a natural language system.

Homograph: A word with the same spelling as another but with a different meaning and origin.

Indexing Term: The label or descriptor chosen to describe or represent a subject aspect of a document.

Inverted Subject File: This file groups together all those documents or records that have a particular subject in common.

Keyword (or *Uniterm*): A keyword is the single-word form of an indexing term, as in a post-coordinate vocabulary.

Keyword Indexing (or *Free Term Indexing*): The free assignment of indexing terms to documents without the use of vocabulary control at the indexing stage.

KWAC Index (*Key-Word-Augmented-in-Context*): A variation of the KWIC index in which additional indexing words are added to the keywords contained in a corpus of text.

KWIC Index (*Key-Word-In-Context*): The KWIC index or permutation index lists each substantive word or keyword in any corpus of text, and each keyword becomes an entry point in a printed index. The keyword is positioned successively in the center of one index line surrounded by the other words in the text.

KWOC Index (*Key-Word-Out-of-Context*): The KWOC index only differs from the KWIC index in format wherein the keyword is positioned at the left-hand margin of the index line.

Linking Device: A precision tool that ties together related terms in a document and separates terms that are not related.

Mathematical Classification SEE *Automatic Classification*

Natural Language (or *Free Text*) *Retrieval System*: A retrieval system in which the natural language of the searcher is searched against the natural language of the text and retrieval occurs where a match is made.

Permutation Index SEE *KWIC Index*

Post-Controlled Vocabulary: A controlled vocabulary that is not used for indexing, but is used an as aid in constructing strategies for searching a stored natural language data base.

Post-Coordinate Index: The indexing units are not selected prior to the preparation of the index entries for specific documents, and the relationship among the indexing units is not indicated. The index entries are manipulated at the time of searching (post-coordination) in order to derive their logical sums, products, and negations. Examples of post-coordinate indexes are concordances and KWIC indexes.

Postings SEE *Term-Frequency List*

Precision: Precision is the ability of an information retrieval system to hold back nonrelevant documents in response to a subject request.

Pre-Controlled Vocabulary: A controlled vocabulary that is used both for indexing and for searching a stored data base.

Recall: Recall is the ability of an information retrieval system to retrieve relevant documents in response to a subject request.

Related Term (RT) Reference: The RT reference is used to indicate a relationship between terms other than the generic-specific type of relationship.

Role Indicator: A precision tool that indicates the type of relationship that exists between terms used in indexing a particular unit of information.

Scope Note: The scope note describes the limit of the subject matter embraced by an indexing term.

See Reference: The *see* reference directs from a term that cannot be used in indexing and searching to a term that can be used.

See Also Reference: The *see also* reference ties together related terms and does not distinguish between the generic-specific relationship and other types of term relationships.

Specificity: The specificity of the vocabulary is the degree to which the indexing terms are able to express the precise concepts or subject representation of a document and the precise concepts of a request.

Stemming (or *Truncation*): A method of reducing words to root form as a means of controlling all variant forms of a word.

Stop List: A list of words that excludes from processing common words that do not in themselves indicate subject matter such as pronouns, articles, prepositions, conjunctions, copula and auxiliary verbs, and quantitative adjectives.

Subject Heading: A subject heading is the combination of words that form an indexing term, as in a pre-coordinate vocabulary.

Synonyms: Two or more words that have the same or nearly the same meaning.

Term-Frequency List (or *Postings*): A record of the number of times that a term has been used in indexing.

Thesaurus: The thesaurus is an alphabetic list of indexing terms or descriptors that provides for the control of synonyms and homographs, and also displays generic-specific and other relationships between terms.

Truncation SEE *Stemming*

Uniterm SEE *Keyword*

Use Reference: The *use* reference directs from a term that cannot be used in indexing and searching to a term that can be used.

Weighting: A precision device which provides that a term must occur a minimal number of times in a document before the document will be retrieved.

Word Distance Indicator: A precision device which permits text retrieval only under specified conditions based on the proximity of the appearance (or co-occurrence) of terms in the text.

Word Indexing SEE *Keyword Indexing*

SECTION 5. DATA BASE DESIGN ALTERNATIVES*

Introduction

The purpose in writing this section is two-fold: (1) to provide a theoretical base for any future implementation of either a manual or a computer-based attitudinal research information system, and (2) to present a theoretical discussion of the variables and alternatives involved in data base design in order to elucidate the key issues and trade-offs and to form the basis and justification for arriving at a definition of system design and implementation requirements for a computer-based attitudinal research information system.

A general description of a data base encompasses a collection of mutually related information, the computer hardware that is used to store the collection, and the programs that are used to manipulate it. "The complete specification of the data base for an [information] system involves precise definitions of the kinds of information to be stored in the system, the relationships of the various types of data to each other, and the physical organization of the information within the computer system. The technique of providing this specification for a particular application is known as *data base design*" (Lancaster & Fayen, 1973, p. 46). The job of the data base designer is to gather together all of the information available about the requirements to be imposed on the system and then to derive an optimal file and data base structure to support them.

A mechanism to support a data base is referred to as a *data base system*. The objectives of a data base system are to be able to (1) systematize the access to data elements, (2) refer to data elements without knowledge of record or file structure, (3) change record or file structure without affecting existing data base functions, (4) handle related files within one structure, and (5) describe the data base from different points of view so that it can become a communication medium between data generators and information seekers. Wiederhold (in preparation) has identified a number of tasks that a data base system should perform.

- The acquisition and storage of the data collection.
- The organization of the information in the data collection to show the logical relationships among the various data elements.

* The bulk of the material presented in this section has been abstracted from a textbook on data base structures and schemas written by Gio Wiederhold, to be published by McGraw-Hill. The book grew out of several courses given by the author at Stanford University, the University of California at Berkeley, and the University of California at San Francisco on data base organization, structures, and schemas. Where other authors' work is cited, appropriate references are provided.

- The updating of data elements in the data collection.
- The search for and retrieval of specific data elements in the data collection.
- The reduction of large quantities of data to usable form.

It is not unusual in data base design to find that the minimum computer adequate to store the data quantity is also adequate to perform the tasks. On the other hand, there are large-scale applications where no single computer is capable of handling all of the tasks that might be required in an active data base environment. It should be pointed out that it is not essential that storage of data bases in a computer be controlled by means of a data base system. This approach becomes desirable only when large amounts of interrelated data are stored which are of interest to multiple groups of users. In this case, the data base system serves the function of handling relations between data elements in multiple files.

In the remainder of this section, the following topics of prime importance in data base design will be discussed: file organization methods, data base systems and schemas, methods to gain reliability, protection of privacy, and data base management. A summary of the major concepts presented concludes this section, accompanied by an annotated bibliography and a glossary of key terms.

File Organization Methods

The basic information unit in an information storage and retrieval system is the *data element*, a particular value of a data entity (e.g., 19 for Years of Age). With each data element the attribute type has to be known to permit meaningful data processing; in this case, the attribute type is "Age." The individual pieces of information that comprise records are variously referred to as data elements, data fields, or data subfields. A *record* has been defined by Lancaster and Fayen (1973, p. 47) as a collection of related data elements, items, or codes that are logically treated as a unit. Spencer (1974, p. 221) elaborates on this definition by stating that a record refers to a group of logically related items that can be manipulated as a single unit during computer processing. The retrieval of data elements is the central objective of a file system. A file system provides the means to fetch entire records according to defined search keys. The term *search key* denotes a specified attribute type and value used to retrieve records that match the search key. The key portion of a record is the field against which the search key is matched. A search key may comprise multiple attribute fields of a record in order to achieve a unique relationship. An example of such a search key is a book title (which alone may not identify a book uniquely), the publisher, and the date of publication.

A *file* is defined to be a collection of similar records kept on secondary computer storage devices. Not only does a file consist of similar records, it also has a consistent organization. The basic features to be desired in information systems that store large amounts of data are rapid access for re-

trieval, convenient update, and economy of storage. In order to gain economy of storage, the data need to be stored with a minimum of redundancy. Redundancy exists when data fields are duplicated, or when the description of the contents of the data fields is repeated with every entry. Reduction of the latter form of redundancy can be aided by imposing structure, so that the position of a data element serves as a partial descriptor. If rapid retrieval to multidimensional queries is an important requirement, then data may need to be stored redundantly.

These desiderata and secondary criteria tend to conflict with each other, so that the choice of method of file organization will determine the relative performance of an information system in these areas. A good matching of the priorities determined by the application of the data base system to the capabilities provided by the file system is vital to the success of the resulting information system. "Where usage is not well defined, or where multiple usage prevents organization in terms of particular predictable needs, file organization must become a complex of structures. The aim of each substructure in the complex is to provide access under some type of entry, for which the substructure is organized to respond optimally---to bring items together which should be considered together under that particular type of access" (Becker & Hayes, 1963, pp. 267-268). However, a complex of substructures promotes redundancy in the file organization.

With the above considerations in mind, there are six basic methods of file organization that need to be weighed by the system designer: the pile file, the sequential file, the indexed sequential file, the multi-indexed file, the direct file, and the multi-ring file.

The Pile File. The *pile file* represents an unstructured minimal method of file organization. Data are collected in the order in which they arrive. They are not analyzed, categorized, or normalized. At best their order may be chronological. The records may be of variable length, and need not have similar sets of data elements. Pile files are found where data are collected prior to processing, where data are not easily organizable, and in some research on file structures. Data banks that have been established for military intelligence may have this form since the potential value of a record is difficult to assess. This type of file organization may also be used in artificial intelligence applications.

The pile file provides a basic unstructured organization which is flexible, wasteful of space when used for storage of well-structured data, easy to update, very awkward for fact finding, but amenable to exhaustive searches.

The Sequential File. The *sequential file* provides two distinct structural changes relative to the pile organization. The first improvement is that the data records are ordered into a specific sequence, and the second improvement is that the data attributes are categorized so that the individual records contain all of the data attribute values in the same order and possibly in the same position. The data attribute names then need to appear only once in the description of the file.

With a sequential file, the file organization is restricted to a limited and predetermined set of attributes. A single description applies to all records, and all records are structurally identical. If a new attribute has to be added to a record, then the entire file has to be reorganized and/or rewritten. The simplest method of file update is to add the new attribute at the end of the record as it is being rewritten. Sometimes, however, it is preferable to insert the new attribute at some position within the record. In this case, half of the records in the file on the average will have to be reorganized and rewritten to provide space for the new data item. To avoid this problem sequential files sometimes are organized initially with space allocated to spare, a practice which is wasteful of storage. The restriction that only one attribute determines the order of the file and consequently that all other attribute values are less suitable as search keys is a disadvantage common to all sequential files. Most of the data files at NPRDC have this type of organization.

The sequential file, containing a collection of ordered fixed records, is inflexible, efficient for the storage of well-structured data, difficult to update, awkward for fact finding, but very suitable for efficient, exhaustive searches.

The Indexed Sequential File. The *indexed sequential file* represents an attempt to overcome the access problem inherent in the sequential file organization without losing all of the benefits and tradition associated with sequential files. One new feature consists of an index to the file to provide better random access; another new feature provides a means to handle additions to the file. Thus, indexed sequential files have the property that the records in the file may be accessed either by the use of an index or in a sequential manner. Indexed sequential files are used in business applications where data files have to be updated regularly, but where a single key is always adequate to find the record. This condition is typical in personnel files, where knowledge of the individual's identification number is a prerequisite, or in billing files, where an account number is always available.

In the case of an indexed sequential file, only a single index is added to the sequential organization. The result is a file that can be searched and updated according to one attribute, but is otherwise still fairly inflexible. The indexed sequential file organization is reasonably efficient for the storage of well-structured data and suitable for exhaustive processing. Sequential files that are indexed only for an attribute of major importance and that experience high utilization provide an efficient method of organizing data that will be searched on the indexed attribute.

The Multi-Indexed File. A file which is accessed only through indexes is referred to as a *multi-indexed file*. There is no restriction on the allocation of space for a data item as long as a pointer exists in some index that allows its position to be determined. No attribute has to be preferred, and indexes may exist for all attributes for which a search may be contemplated. The multi-indexed file removes the constraint on sequentiality and allows many search attributes. Examples of such files are to be found in airline and rental car reservation systems, job banks, and other inventory-type applications. In these kinds of applications data rarely are processed sequentially, other than for occasional, perhaps annual, stock-taking.

A file for which complete indexes have been created is sometimes referred to as an *inverted file*. This term has its origin in bibliographic indexing. A complete index to a file that contains text in English may have as entries all of the words in the file that are not on a stop list. This type of index is essentially a vocabulary with pointers to the text where the words appear. If the pointers are augmented by a sample of the text, then the result assumes the form of a concordance.

The critical decision for the system designer is how many attributes should be indexed. An index of all attributes can easily exceed the size of the original file! In practice, there usually are some attributes for which indexing is not justified. When all of the data in a file are indexed, then all of the available information may be contained in the index, and conceivably, the data themselves could be eliminated. This possibility is referred to as a *phantom file*. An example of a phantom file can be found in the area of library cataloging. Searches through the indexes are made using logical combinations of request parameters which ultimately produce a set of call numbers for books. These call numbers are in essence the addresses to the desired retrieval objects, the books on the shelves of the library. A library which would always be accessed in this manner could store its books according to considerations other than the sequentiality of subject codes. If, instead of call numbers, the system provided the actual shelf number and position, then books could be stored according to size, to minimize space requirements, or according to frequency of access, to minimize retrieval time. Browsing in such a library might be less productive, but probably more surprising. Currently, such optimizations are, in practice, limited to computer files.

A *transposed file* offers another variation of the indexed file. A transposed file is organized toward selection or evaluation of data according to a dimension which is orthogonal to the data entry or update dimension. This type of file presents opportunities for the discovery of trends or relations which are otherwise not obvious. Many research-oriented data processing tasks take transposed data and then group, select, and correlate the values found.

The major advantage of the indexed or inverted form of file organization is that it allows access to all of the data in the file with equal ease. Since the pointers to all of the data are stored in the index, boolean search logic can be applied before any records are actually retrieved from the storage media (Lancaster & Fayen, 1973, pp. 56-57). Thus, it is possible to obtain frequency counts of records responsive to a particular search request before initiating the actual retrieval. This feature permits a searcher to expand his search parameters if too few records match his search request. Conversely, if too large a number of records match his request, then he has the option of narrowing his search by respecifying the search parameters.

However, there is a penalty to be paid for this ease of access and retrieval. Storing and maintaining an indexed or inverted file is more cumbersome than with other file organization methods because large directories must be created and maintained as well, directories which contain all of the data values in the file and the addresses of all locations where those values occur (Lancaster & Fayen, 1973, p. 57). Negus and Hall (1971, pp. 260-261) have pointed out that indexed or inverted files involve much more processing time

to update and require a greater amount of computer storage. Although searching an indexed or inverted file generally requires less computer processing time, this advantage may be more than offset by the cost penalty imposed by the increased amount of computer storage required.

In summary, the multi-indexed file is fairly flexible, reasonably efficient for data storage if indexes are selected with care, permits updates at a fair cost in complexity, allows convenient fact finding, but is awkward for exhaustive processing. A great deal of storage may be occupied by the indexes.

Direct File. The *direct file* relates most closely the attribute value used in a search to the physical capabilities of direct access mechanisms. In a direct file structure, the position of the record is determined by a randomizing computation based on the key value. No (or few) file accesses are required to locate the record other than the data record retrieval operation or the data record writing operation itself. The direct access method is extremely fast since it avoids intermediate file operations, but the method forces the data to be positioned according to a single search attribute. The searches through index tables that aid in the location of a record in the indexed file organizations are replaced by a computation whose object is to provide the record address.

Direct files are frequently used for directories, pricing tables, schedules, and similar applications. In such applications, where the record sizes are small and fixed, where fast access is essential, and where the data are always accessed according to a single key attribute, the direct file organization is uniquely suitable. However, the direct file is quite inflexible because of its mapping requirements, it has some storage overhead, and it permits updates at some cost in complexity. This method of file organization provides efficient fact retrieval according to a single dimension, but it may prove to be impossible for exhaustive searches.

Multi-Ring File. The *multi-ring file* is oriented towards efficient finding of *sets* of records that contain some common attribute value. These subsets are explicitly chained together through the use of pointers which define some order for the members of the subsets. A record may be a member of many such subsets. A header record will contain information which pertains to all of its subordinate member records.

Multi-ring structures are the basis for some of the largest data bases currently in use. Management information systems, where much of the system operation involves tabulating, summarizing, and exception reporting, have been implemented using multi-linked lists. The multi-ring file provides for a number of record types with many interconnections. Redundant and empty fields are reduced, but some space is required for the linkage pointers. Updates may be performed at a moderate cost. The ring organization provides good, but not always fast access to facts, and allows great flexibility for exhaustive or subset searches.

Only the multi-indexed file and the multi-ring file provide the capability for access to data according to more than one dimension. Whenever there are many records and the user demands are such that a variety of searches and

groupings of data seem necessary, then these two methods provide the basic building blocks for the appropriate file designs. These two methods of file organization are not routinely available now to researchers and scientists, since more typically they are part and parcel of large data management systems. These systems may provide report-generating and tabulating facilities, but rarely do they include the more sophisticated statistical techniques needed for scientific analysis. For special applications, tailor-made support systems frequently are written which use these file organization methods.

Comparison of File Methods. In Table 4 a summary of the performance characteristics of the six basic methods of file organization is presented for three criteria: optimization of storage space; ease of update; and efficiency of retrieval for facts, for a subset summary, and for exhaustive searches. If the data in the file are unstructured, the pile file provides the best use of storage space. If the data in the file are structured, either the sequential file or the direct file optimizes the use of storage space. The easiest file to update, regardless if the record size is fixed or variable, is the pile file. For fixed-length records, the indexed sequential file and the direct file are quite easy to update. None of the six types of file organization provides ease of update for variable-length records with the exception of the pile file. The multi-indexed file is exceptionally well suited for retrieving facts, followed closely in efficiency by the indexed sequential file, the direct file, and the multi-ring file. Only the multi-ring file and the multi-indexed file are specifically suited for retrieving a summary subset of data, with the multi-ring file being superior in this regard to the multi-indexed file. For exhaustive searches the sequential file has the edge, followed in efficiency by the pile file, the indexed sequential file, and the multi-ring file.

Different types of file organization may place different requirements on the contents of the *file directory*. Associated with each file may be a header or a file directory record. This record contains information describing the position and the organization of the detailed records of that file. The collection of directory records for a number of files may in turn form a file.

Data Base Systems and Schemas

A picture of the data in terms of files, records, fields, and the relation between items of data contained in these entities is appropriate for describing the static aspects of an information system. Such a data structure definition may be available in the form of a document containing guidelines to the individuals who are responsible for programming the file operations for an information system. An alternative approach is to materialize the data structure definitions in the form of a collection of computer-readable codes, a schema, which can guide file processes automatically.

The collection of information that describes the data base, when organized in a formal manner, is called the *data base schema*. The schema should be readable by the data base system and used by generalized programs to control the flow of data to the files that contain the data base. The schema is stored within the system to be accessible when needed, both to properly place incoming data into the files and to locate required data at a later time.

TABLE 4

GRADES OF PERFORMANCE ON THREE CRITERIA
FOR THE SIX BASIC METHODS OF FILE ORGANIZATION

File Method	*Optimization of Storage Space*		*Ease of Update*		*Efficiency of Retrieval*		
	Structured Data		Record Size		For Facts	For a Subset Summary	For Exhaustive Searches
	No	Yes	Fixed	Variable			
Pile	A	E	A	A	E	D	B
Sequential	F	A	F	F	F	D	A
Indexed Sequential	F	B	B	D	B	D	B
Multi-Indexed	C	D	C	C	A	B	D
Direct	F	A	B	F	B	F	F
Multi-Ring	C	B	D	D	B	A	B

LEGEND:

A - Excellent; specifically suited
B - Quite good
C - Adequate
D - Marginally useful
E - Useful only at great effort
F - Unusable; a failure

The formal schema description also provides a means for the data base users, data base designers, and programming staff to communicate and to define their concepts. The dictionaries associated with the schema aid the users in describing their requests, and perform a filtering function to improve data quality within the data base.

Figure 22 portrays the component parts of a simple file and shows the schema which describes this file. In this file a row describes a specific entity belonging to the file holdings. A column represents a data element type. In this simple file there are only four types of data elements: NAME, YEAR OF BIRTH, LENGTH OF SERVICE, and T-SCORE. A data element is the single value at the intersection of an entity row and a data type column. The schema provides the formal description for this file including for each data element its type, its length, its allowable values, and any other information needed to define it completely.

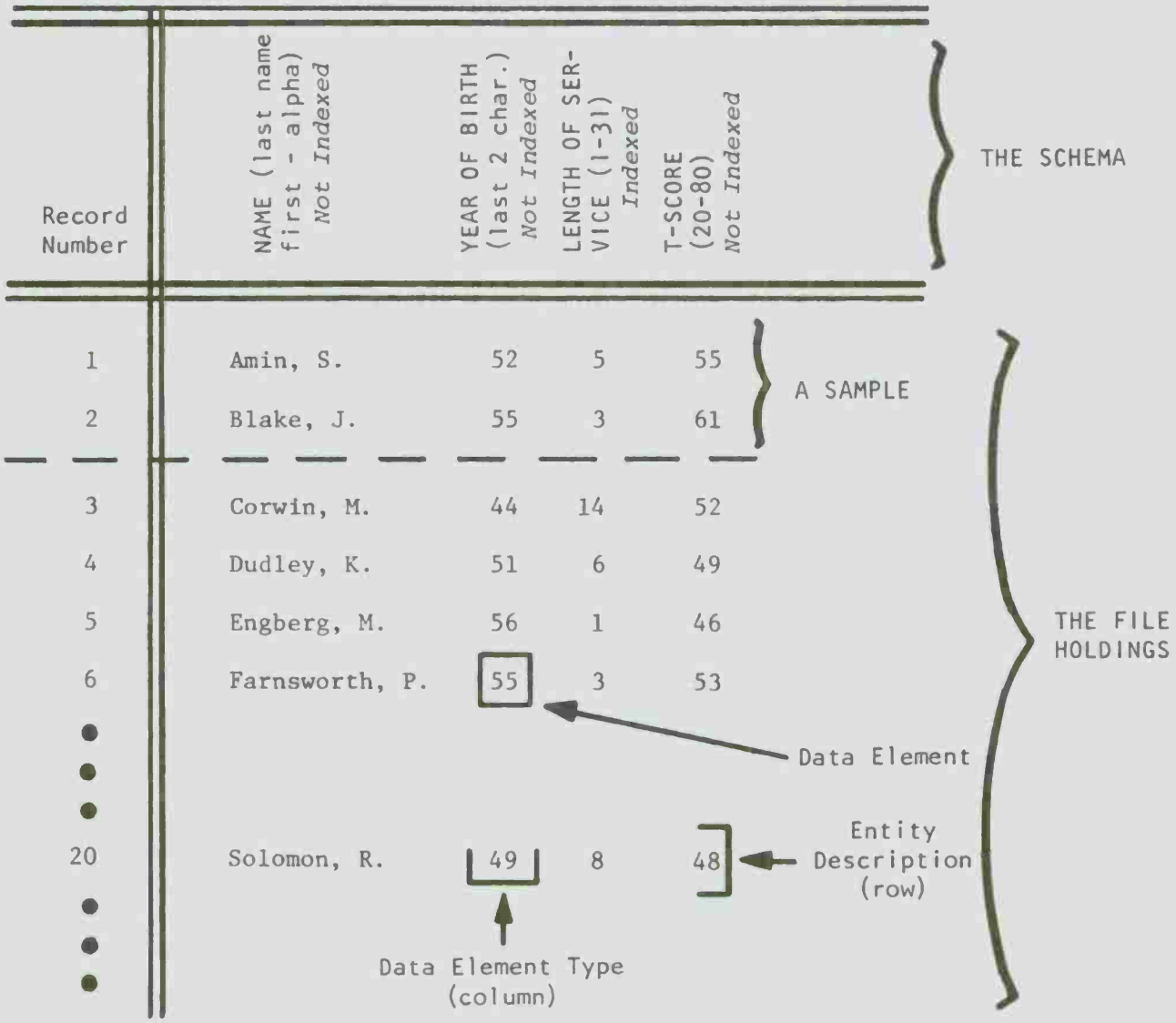
The descriptions collected in the schema can be used either in generating the data base system and its processing functions, or its use can be deferred to the execution time of the processes that manipulate the data base. In the environment of the data base, *compiling* is equivalent to using all of the information in the schema when the programs are created; *interpreting*, on the other hand, is equivalent to the use of a general program that, when called upon to carry out data base manipulations, looks at the schema to find items and determines their relationships. Compromises between the two methods are possible.

In a large data base system, the schema itself may be of massive proportions, and further structuring may be required. Each process on the data base will use only some of the available data, and thus, only a fraction of the schema entries. It makes sense, therefore, to identify subsets of the schema for specific processes. When multiple processes using different but overlapping subschemas are simultaneously active, then it is desirable for them to be able to share schema entries. Only privilege of access information, if tied to processes rather than to the data base elements, must remain separate. The schema as the repository of access privilege information is discussed in more detail later under *Protection of Privacy*.

Methods To Gain Reliability

The term *security* is used to describe the protection of computer systems and their contents from destruction. If one wishes to secure a data base, a reliable and predictable mode of operation will need to be achieved. Therefore, a protection mechanism must be provided to attain the desired control of access to the data. Finally, it is necessary to assure that there will be no destructive interference as multiple users share access to the data base.

System *reliability* is achieved by maximizing the probability that the system will do what it is instructed to do. Perfect reliability is attained when a computer system, both hardware and software, always produces correct results. One obvious problem with this criterion of reliability is the determination of correctness. Current methods of program debugging are inadequate since at best they are limited, under typical conditions, to the verification of a few sample calculations.



ACCESS PROCESS FOR AN INDIVIDUAL: (1) Search through the schema to determine where in the file the attribute NAME is stored. (2) Determine from the schema if the attribute is indexed. Since it is not indexed, (3) search through the file by last name until the desired name is located (a match) or until the next name in the file exceeds the desired name in the alphabetic order (a nonmatch).

Figure 22. A Simplified File Structure Showing Its Components and Its Schema.

In order to produce correct results, what is needed are correct data and correct algorithms. Furthermore, the computer system must carry out the algorithms correctly. In each of these areas the problem reduces to two subproblems:

1. The existence of an error has to be detected, or the absence of errors has to be proven.

2. When an error is detected, a means of correcting and recovering from the error has to be available.

A single technique may provide both a degree of detection and some restoration capability.

Computer systems are composed of many parts that are prone to failure. The system components themselves luckily have much higher reliability probabilities. The failure rate of modern electronics seems not to be affected primarily by number of operations carried out, but generally is specified in terms of operational time alone. For example, a typical mean time between failures for a highly loaded digital switching transistor is 40,000 hours, equal to a failure probability of only .000025 per hour. In practice, not all component failures will have a detectable effect on system performance because not all components are contributing to the operation of the computer at any one time. Experience has shown that many computer failures are of the transient type. Such failures may be caused by infrequent aberrations such as power fluctuations, electrical noise, and/or accumulation of static electricity.

Some parts of computer systems have much higher error rates than others. Magnetic storage devices, communication lines, and data entry and output devices are more prone to error. Human beings are essential elements of computer systems, and they evidence even higher error rates.

The reduction of data entry errors is strongly related to the quality of the transcription device and its operation. Errors attributable to mistyping or miskeying can be reduced dramatically through the use of input devices that provide a selection capability. Display screens or forms that present a list of choices, requiring data input personnel only to indicate their selection for the entry of a data element, expedite the data entry process and contribute substantially to the accuracy of the data base.

Quality control of data entry processes may be accomplished by a number of techniques. Data base systems may include format verification procedures as well as comparisons with previously recorded or recently entered data. A display, in the form of plain English language, of the meaning of codes or identification numbers used to enter data text can aid significantly in error control. If data entry is by means of a terminal, it is important that response messages clearly distinguish the difference between occasional error messages and repetitive confirmation messages. The accompaniment of an audio error signal can be very helpful used in conjunction with visual-display data entry terminals when large quantities of data are being entered.

Redundancy of information is obtained when data units contain more information than is strictly necessary (Wilcox & Mann, Eds., 1962). The data units can be checked for internal consistency using the redundant information, and if they are found to be in error, they can be corrected. Three basic methods for effecting redundancy are parity checking, duplication of data, and error-correcting codes.

A simple form of redundancy is obtained with individual data elements through the use of a *parity bit*. Characters on magnetic tape or in computer memory frequently have such a redundant bit added when they are generated initially. The count of the number of bits of value 1 in the character representation is termed the *Hamming weight* (Hamming, 1950). If the character is represented by eight bits, then the parity bit in the ninth position will be set so that the Hamming weight of the combination is odd (odd parity encoding) or even (even parity encoding). Odd parity is the preferred code since it avoids any occurrences of all zero or all blank sections on the recording medium, a condition which is necessary for assuring the detection of the existence of a character, particularly on magnetic tape. Odd parity produces a code with a minimum Hamming weight of 1. The number of bits which are different between the coded representation of the character transmitted and the actual code received is termed the *Hamming distance*, and is equal to the difference in the Hamming weights of the two codes.

Simple parity coding is applicable to the communication-oriented processes in computing. Only error detection is provided by this method, and even this detection is not complete. In addition, this method of error detection cannot detect errors related to the content of the data; it applies only to the detection of data transmission errors.

Duplication of data is another simple form of achieving redundancy in computing systems. Keypunching with subsequent verification of the contents of the punched cards by rekeying and matching is an example of duplication of data during one small interval of the information-processing cycle. Completely duplicated storage of data rarely is employed in view of the normally high reliability of computer storage devices. However, the method frequently is used with magnetic tape storage and when the data to be stored have a unique or high value. The cost of duplicate entering and maintaining of data is quite high, so the desirability of providing data security by means of back-up files must be weighed carefully against the cost. When information is copied or transmitted, usually a duplicate will exist for some length of time. It is wise to design the operations of an information system so that these original duplicates are available until accuracy of the copies has been verified.

Error-correcting codes provide many of the benefits obtained by duplication of data without the burden of all of the cost. In this method a certain number of redundant bits are added to each data element. These check bits are produced similarly to parity bits, but each check bit represents different groupings of the information bits. The groups are determined in a manner that results in the detection of which bit is in error so that it can be corrected. A specific form of an error-correcting code can only correct up to a finite number of bits; therefore, the usefulness of this method is greatest where an error extending only over a few bits occurs in relatively many data elements. This technique for improving the reliability of data is used extensively with disk files and similar devices.

Protection of Privacy

There are two aspects to the issue of protection of privacy. The first, and most commonly understood, aspect is that access to data must be *denied* to those people who do not have a right to access these data. The second, but equally important, aspect of protection is that access to all relevant data must be *guaranteed* to those people who exercise their access privilege properly.

Three elements must be considered in designing an adequate system for protecting the privacy of data: (1) the accessors, (2) the type of access desired, and (3) the data objects to be accessed. Each of these elements has to be properly identified in order to achieve control over access to data.

The Accessor's Identification Key. To the information system the external identification of an accessor is primarily the accessor's name as entered by him into the system. The accessor also may be identified by a password, which has to be typed in on request, or perhaps by a machine-readable key or badge. Typically, legitimate system users are lax in guarding the secrecy of their access protocol. They fear that they may forget it so they write it down, and perhaps may even give it to someone else who acts in their behalf. This laxity can be partially overcome by changing the access passwords frequently and/or by providing several levels of access privilege. Certain methods to guarantee the positive identification of a privileged user have been proposed and tested, all of which depend on the biological uniqueness of each human being. Three manifestations of this unique coding of individuals are fingerprints, voice prints, and retinal patterns. Since the cost to implement such a method for protecting privacy of data is substantial, few systems employ biological identification mechanisms. However, for information systems that demand ultra high control of access, these mechanisms should be considered as possibilities by the system designer.

Types of Data Access. Methods for accessing data can be categorized by type. Traditionally, distinctions have been made between permission to read data and permission to write data. Magnetic tape reels and some tape cartridges have inserts which, when removed, make writing physically impossible. Some disk drives have write protect switches. More sophisticated systems have added an execute-only privilege. Current computer systems limited to these three types of access control provide fairly unsatisfactory protection. The procedures which are part of the hardware operating system frequently have the privilege to read or write anything, anywhere, and could not function if this privilege were removed. The read privilege frequently is available to any user, and only a knowledge of the structure of another user's data is required to gain access to the files of others.

A distinction among seven types of access privileges should provide a far greater degree of protection of privacy.

READ access would grant the privilege of copying data into the accessor's working environment.

EXECUTE access would grant the privilege of use of a program or a procedure.

CHANGE access would provide the accessor with the conventional write access.

DELETE access would allow destruction of the information that a data object existed as well as destroying the data object itself.

EXTEND access would allow the addition of information to a file without the privilege to destroy previous data and without the privilege to read previous information unless those privileges also were conferred.

MOVE access would provide the capability to move data fields without the privilege to read their contents.

EXISTENCE VERIFICATION access would make it possible to determine whether a specific data element exists in order to make decisions on the invocation of further processes.

These seven distinct types of access privileges provide $2^7 = 128$ combinations as shown in Table 5. The 0-bit position is used to flag whether or not the access key itself is valid. The remaining seven bits define the types of access to be granted. The two combinations 0000000 and 1111111 will occur most frequently. The first combination is an unconditional denial of access by any type of access privilege. The second combination represents the full set of access privileges required by many users of the system. An explicit understanding of the full dimension of possible access types appropriate for specific cases is necessary in order to be able to specify the security level of a system design. When designing a new system, it is important to define the protection level which will be required for the user community.

TABLE 5

BIT ASSIGNMENT FOR A PROTECTION KEY BYTE

Bit 0	KEY Itself Is Valid
Bit 1	READ Access Is Granted
Bit 2	EXECUTE Access Is Granted
Bit 3	CHANGE Access Is Granted
Bit 4	DELETE Access Is Granted
Bit 5	EXTEND Access Is Granted
Bit 6	MOVE Access Is Granted
Bit 7	EXISTENCE VERIFICATION Access Is Granted

Data Objects To Be Locked. The number of data objects that are candidates for protection is the sum of the following:

1. Basic data elements,
2. Indexes or relations,
3. Programs or procedures, and
4. Schema entries and relations.

Adequate protection of privacy may be achieved when data that are obtainable cannot be linked to specific individuals or cannot be used to implicate facts about them. The identity of individuals and their corresponding confidential information can be effectively obscured by combining the data for several individuals into a single aggregated record. Partial aggregation replaces microunit data, which typically are regarded as confidential, by nonconfidential observations on "statistical stereotypes" (Feige & Watts, 1970). The observations on statistical stereotypes are constructed by combining microunit data into groups and then calculating the mean values for the grouped observations.

The schema is the obvious repository of privilege information, that is, the specification of permission to access data elements. Data ownership is another candidate for an attribute to be specified in the schema. When data or relations are to be added to the data base, then it should be established who has the maintenance responsibility over these data. Private and public data may be distinguished as well as source and derived data elements.

Cost Considerations. In the area of protection of privacy, there are cost considerations similar to those encountered in the discussion of reliability. Protection of privacy, even where and when desired, is relative. The more protection that is implemented to reduce accidental and deliberate incursions of privacy, the higher the cost of the system is likely to be. When the cost of protection exceeds the value of the data elements to be protected, then a limit has been reached.

The cost of providing privacy protection currently is much higher than it has to be. Large inefficiencies are introduced by the fact that computer hardware design has stressed capability and efficiency rather than protection. This emphasis has had the result that costly software mechanisms are required to implement even marginal levels of security. Many validation procedures are required to limit access between separate domains of a computer system, measurably reducing computer performance. This complexity in the protection system contributes to unreliability while still affording opportunities to compromise privacy of data.

The value of adequate protection in commercial computer systems has been estimated at 10% to 20% of basic data processing costs. However, the methods proposed by the manufacturers have not yet been shown to be satisfactory, and a higher cost can be expected eventually if protection of privacy is to be achieved. An optimistic outlook for the anticipated economics of protection of privacy is provided by hardware suppliers who project that the future cost of protection systems can be expected to fall well within the range of perceived value.

Data Base Management

In an organization where many users and programmers share the data base and use the facilities of the schema or subschemas, some joint control of the data organization is required. This control can be implemented through an automatic protection system which prevents schema changes that could cause problems for other users. Typical conflicts occur when an overall system improvement inconveniences a few users severely, or when new requirements by some users affect many other users to a slight extent. In general, a human arbiter is called upon to resolve these conflicts.

The set of control decisions regarding shared use of the data base and the data base schema is referred to as *data base management*. Both automatic and manual actions fall under this heading. In order to carry out the data base management functions, additional information may need to be appended to the schema.

When the resources of an information system are shared, then resource control through the schema is indicated. Various types of system operation allocate resources in different ways. Traditional task scheduling or *batch processing* makes all or most of the resources of the computer available to one job during the required time period. In this method only a few jobs will be able to use all of the available resources effectively as they are processed in the single job-stream batch mode.

Multiprogramming attempts to utilize resources more effectively by allowing sharing of the resources according to current requirements. In the multiprogramming method of allocating resources, there is concurrent execution of two or more programs simultaneously residing in the internal storage unit of the computer. The basic principle of multiprogramming is that the programs in internal storage share the available central processing unit's time and input-output units (Spencer, 1974, p. 362). While input-output operations of one program are being handled, the central processing unit is essentially free to perform the computational tasks required by another program.

In systems that use multiprogramming, individual jobs will take longer, but if the resources can be distributed better among the jobs, the total productivity will be greater. However, some amount of processor capability must be taken from the total amount available in order to effect the switching between jobs.

Timesharing provides a mechanism to split jobs into slices and to only allocate the critical resources to them when they can actively use them. During the inactive periods users may be "swapped out" of the primary resources and temporarily relegated to disk or drum storage. This mechanism provides more effective resource utilization when the tasks are of a type that cannot make use of resources during relatively long periods of time. These pauses frequently are related to delays in waiting for responses from users at terminals.

When jobs are naturally small, then a mode of operation referred to as *transaction processing* may be elected. In this method of resource allocation,

the jobs are started as soon as possible after a request for initiation, and they are permitted to run as long as necessary to fulfill their computational requirements. When they come to a point where they have to wait for a response from terminals or files, then they must yield to the system to enable another transaction process to be started or resumed. When a transaction is completed, the system is informed so that it can release all allocated resources.

Any form of resource sharing is associated with considerable operational and management overhead. If there is no significant benefit to the use of shared resources, then a dedicated system that is of just adequate size is to be preferred. Bisco (1970, p. 274) has offered the opinion that social science research is not yet at a point where multi-data set linkage is an important value since it requires too much interaction between the funding agency, the individual principal investigator, and the many others who would be interested in using the data collection if they had some opportunity to, say, insert their own questions in a survey. He concluded that the social science research community is only beginning to view the presently used information transfer network as a relatively expensive and inefficient system. More appropriate and more efficient methods of file organization may help to overcome this barrier to effective sharing of our information resources.

Summary

A general description of a data base encompasses a collection of mutually related information, the computer hardware that is used to store the collection, and the programs that are used to manipulate it. The technique of providing this specification for a particular application is known as data base design. A mechanism to support a data base is referred to as a data base system. The objectives of a data base system are to be able to (1) systematize the access to data elements, (2) refer to data elements without knowledge of record or file structure, (3) change record or file structure without affecting existing data base functions, (4) handle related files within one structure, and (5) describe the data base from different points of view so that it can become a communication medium between data generators and information seekers.

The basic information unit in an information storage and retrieval system is the data element, a particular value of a data entity. With each data element the attribute type has to be known to permit meaningful data processing. A record is defined as a collection of logically related data elements that can be manipulated as a single unit during computer processing. The retrieval of data elements is the central objective of a file system. A file system provides the means to fetch entire records according to defined search keys. The term search key denotes a specified attribute type and value used to retrieve records that match the search key.

A file is defined to be a collection of similar records kept on secondary computer storage devices. Not only does a file consist of similar records, it also has a consistent organization. There are six basic methods of file organization: the pile file, the sequential file, the indexed sequential file, the multi-indexed file, the direct file, and the multi-ring file.

The pile file represents an unstructured, minimal method of file organization. Data are collected in the order in which they arrive. They are not analyzed, categorized, or normalized. At best their order may be chronological. The records may be of variable length, and need not have similar sets of data elements. This method of file organization is flexible, wasteful of space when used for storage of well-structured data, easy to update, very awkward for fact finding, but amenable to exhaustive searches.

The sequential file provides two distinct structural changes relative to the pile organization: the data records are ordered into a specific sequence, and the data attributes are categorized so that the individual records contain all of the data attribute values in the same order and possibly in the same position. A single description applies to all records, and all records are structurally identical. The sequential file, containing a collection of ordered fixed records, is inflexible, efficient for the storage of well-structured data, difficult to update, awkward for fact finding, but very suitable for efficient, exhaustive searches.

The indexed sequential file has the property that the records in the file may be accessed either by the use of an index or in a sequential manner. This type of file organization is reasonably efficient for the storage of well-structured data and suitable for exhaustive processing.

A file which is accessed only through indexes is referred to as a multi-indexed file. No attribute has to be preferred, and indexes may exist for all attributes for which a search may be contemplated. A file for which complete indexes have been created is sometimes referred to as an inverted file. When all of the data in a file are indexed, then all of the available information may be contained in the index, and conceivably, the data themselves could be eliminated. This possibility is referred to as a phantom file. A transposed file offers another variation of the indexed file. This type of file is organized toward selection or evaluation of data according to a dimension which is orthogonal to the data entry or update dimension. The multi-indexed file is fairly flexible, reasonably efficient for data storage if indexes are selected with care, permits updates at a fair cost in complexity, allows convenient fact finding, but is awkward for exhaustive processing. A great deal of storage may be occupied by the indexes.

The direct file relates most closely the attribute value used in a search to the physical capabilities of direct access mechanisms. The direct access method is extremely fast since it avoids intermediate file operations, but the method forces the data to be positioned according to a single search attribute. This method of file organization is quite inflexible, has some storage overhead, and permits updates at some cost in complexity. The method may prove to be impossible for exhaustive searches.

The multi-ring file is oriented towards efficient finding of sets of records that contain some common attribute value. These subsets are explicitly chained together through the use of pointers which define some order for the members of the subsets. Redundant and empty fields are reduced, but some space is required for the linkage pointers. Updates may be performed at a moderate cost. The ring organization provides good, but not always fast

access to facts, and allows great flexibility for exhaustive or subset searches. Only the multi-indexed file and the multi-ring file provide the capability for access to data according to more than one dimension.

These six basic methods of file organization often are combined in various ways to provide access features required by a specific application. Systems using primary and secondary indexes, and tree structures are examples of the use of access methods used in combination.

The data base schema is a formally organized collection of computer-readable codes that describes the data base and guides file processes automatically. The schema is used to properly place incoming data into the files and to locate required data at a later time. The descriptions collected in the schema can be used either in generating the data base system and its processing functions, or its use can be deferred to the execution time of the processes that manipulate the data base. In a large data base system, the schema itself may be of massive proportions.

The term security is used to describe the protection of computer systems and their contents from destruction. If one wishes to secure a data base, a reliable and predictable mode of operation will need to be provided. System reliability is achieved by maximizing the probability that the system will do what it is instructed to do. Perfect reliability is attained when a computer system, both hardware and software, always produces correct results. In the absence of perfect reliability, error detection and error correction techniques are needed.

Redundancy of information is obtained when data units contain more information than is strictly necessary. The data units can be checked for internal consistency using the redundant information, and if they are found to be in error, they can be corrected. Three basic methods for effecting redundancy are parity checking, duplication of data, and error-correcting codes.

Three elements must be considered in designing an adequate system for protecting the privacy of data: (1) the accessors, (2) the type of access desired, and (3) the data objects to be accessed. Typically, the accessor is identified to the system by name, password, or a machine-readable key or badge. If seven different types of access privileges are implemented, a far greater degree of protection of privacy results. The number of data objects that are candidates for protection is the sum of the following: (1) basic data elements, (2) indexes or relations, (3) programs or procedures, and (4) schema entries and relations. Adequate protection of privacy may be achieved when data that are obtainable cannot be linked to specific individuals or cannot be used to implicate facts about them. The schema is the obvious repository of privilege information, that is, the specification of permission to access data elements.

In an organization where many users and programmers share the data base and use the facilities of the schema, some joint control of the data organization is required. The set of control decisions regarding shared use of the data base and the data base schema is referred to as data base management.

Both automatic and manual actions fall under this heading. Control can be implemented through an automatic protection system which prevents schema changes that could cause problems for other users.

When resources of an information system are shared, then resource control through the schema is indicated. Various types of system operation allocate resources in different ways. Batch processing makes all or most of the resources of the computer available to one job during the required time period. Multiprogramming allocates resources so that there is concurrent execution of two or more programs, all sharing the available central processing unit's time and input-output units. Timesharing provides a mechanism to split jobs into slices and to only allocate the critical resources of the computing system to them when they can actively use them. Transaction processing is a method of resource allocation in which jobs are started as soon as possible and permitted to run as long as necessary to fulfill their computational requirements. When they come to a point where they have to wait for a response, then they must yield to the system to enable another transaction process to be started or resumed.

ANNOTATED BIBLIOGRAPHY

1. Becker, J., & Hayes, R. M. *Information Storage and Retrieval: Tools, Elements, Theories*. New York: Wiley, 1963.

This book provides a foundation and structure within which developments in information retrieval and allied fields can be viewed for their relationship and interaction with each other.

2. Bisco, R. L. (Ed.). *Data Bases, Computers, and the Social Sciences*. New York: Wiley-Interscience, 1970.

This book is a product of the Fourth Annual Conference of the Council of Social Science Data Archives. It includes post-edited versions of some of the papers that were presented concerning the interrelations between data banks, computer technology, and the needs of the social sciences.

3. Brewer, S. Data base or data maze - An exploration of entry points. *Proceedings of the National ACM Conference*, 1968, 623-630.

An elementary review of file organization methods.

4. Climenson, W. D. File organization and search techniques. In C. A. Cuadra (Ed.), *Annual Review of Information Science and Technology* (Vol. 1). New York: Interscience Publishers (Wiley), 1966. Pp. 107-135.

This chapter considers the representations of information as symbols to be organized and manipulated within a computer system where there is no semantic content implied.

5. CODASYL Data Base Task Group. *Report of the Data Base Task Group*. New York: Association for Computing Machinery, April 1971.

This report contains revised language specifications for COBOL-oriented implementations. Very related to the Honeywell IDS implementation.

6. Collmeyer, A. J., & Shemer, J. E. Analysis of retrieval performance for selected file organization techniques. *AFIPS Conf. Proc. Fall Joint Computer Conf.*, 1970, *37*, 201-210.

This paper analyzes the retrieval performance of a selected number of file organization techniques.

7. Fiege, E. L., & Watts, H. W. Protection of privacy through microaggregation. In Bisco, R. L. (Ed.), *Data Bases, Computers, and the Social Sciences*. New York: Wiley-Interscience, 1970.

This article discusses the advantages and disadvantages of partial aggregation of data to replace microunit observations, which are typically regarded as confidential, by nonconfidential mean values for the grouped observations.

8. Hamming, R. W. Error detecting and error correcting codes. *Bell System Tech. Journal*, 1950, *29*, 147-160.

A fundamental paper discussing the use of redundant codes for detecting and correcting transmission errors in communication channels.

9. Judd, D. R. *Use of Files*. New York: American Elseviers Publishing Co., 1973.

An introduction to some commercial file methods.

10. Lancaster, F. W., & Fayen, E. G. *Information Retrieval On-Line* (a Wiley-Becker & Hayes Series book). Los Angeles: Melville Publishing Co., 1973.

This book provides a broad survey of the characteristics, capabilities, and limitations of then current information retrieval systems operated in an on-line interactive mode.

11. Martin, J. F. *Computer Data-Base Organization*. Englewood Cliffs, N.J.: Prentice Hall, 1974.

A description of data base organization methods.

12. McLaughlin, R. A. Building a data base. *Datamation*, 1972, *18*(7), 51-55.

Starting from scratch, the staff of *Datamation*'s EDP Industry Directory assembled a data base for publication. This article tells how it was done, including the problems encountered and the lessons learned.

13. Miller, A. R. Personal privacy in the computer age: The challenge of a new technology in an information-oriented society. *Michigan Law Review*, April 1969, *67*(6), 1089-1246.

An analysis of the privacy problem accompanied by a thorough bibliography.

14. Negus, A. E., & Hall, J. L. Towards an effective on-line reference retrieval system. *Information Storage Retrieval*, 1971, *7*(6), 249-270.

A discussion of the costs of a small or medium-sized, in-house, on-line retrieval system.

15. Patterson, A. C. Data base hazards. *Datamation*, 1972, *18*(7), 48-50.

A discussion of the pros and cons of installing a data base management system.

16. Plagman, B. K., & Altshuler, G. A data dictionary/directory within the context of an integrated corporate data base. *AFIPS Conf. Proc. Fall Joint Computer Conf.*, 1972, *41*, 1133-1140.

A discussion of the data base schema concept.

17. Spencer, D. D. *Introduction to Information Processing.* Columbus, Ohio: Charles E. Merrill Publishing Co., 1974.

A basic introduction to information processing concepts, terminology, techniques, and hardware. An excellent initial exposure to this area of technology for the layman or neophyte information scientist.

18. Uhrowczik, P. P. Data dictionary/directories. *IBM System Journal*, December 1973, *12*(4), 332-350.

An IBM view of data base schemas.

19. Wiederhold, G. *Data Base Structures and Schemas.* New York: McGraw-Hill, in preparation.

A textbook presenting the methods, the choices, and the principles and concepts that are focal to data base organization and management.

20. Wiederhold, G., Fries, J. F., & Weyl, S. Structured organization of clinical data bases. *AFIPS Conf. Proc. Nat'l Computer Conf.*, 1975, *44*, 479-485.

This paper describes a data base support system implemented at the Stanford University Medical Center to achieve commonality of data base use and to provide a basis for controlled future development.

21. Wilcox, R. H., & Mann, W. C. (Eds.). *Redundancy Techniques for Computing Systems.* Washington, D.C.: Spartan Books, 1962.

This book is based on a collection of papers presented at a 1962 symposium on redundancy techniques for computing systems. The objective of this symposium was to focus attention toward new ideas, research, and developments which would lead to the sound introduction of redundancy techniques into forthcoming computing systems.

22. Wos, C. M. The evaluation of file-management software packages. *Proc. Amer. Soc. Information Sci.*, 1969, *6*, 215-222.

This paper presents an evaluation procedure to be followed in selecting a file-management software package to fulfill the demands of a particular computing environment.

GLOSSARY OF KEY TERMS

Batch Processing: Traditional task scheduling in which all or most of the resources of the computer are made available to one job during the required time period.

Compiling: In the environment of the data base, compiling is equivalent to using all of the information in the data base schema when the programs are created. SEE ALSO: *Interpreting*.

Concordance: An alphabetical index of all the words in a text or corpus of texts, showing every contextual occurrence of a word.

Data Base (general description): A collection of mutually related information, the computer hardware that is used to store the collection, and the programs that are used to manipulate it.

Data Base Design: The complete specification of the data base for an information system including precise definitions of the kinds of information to be stored in the system, the relationships of the various types of data to each other, and the physical organization of the information within the computer system.

Data Base Management: The set of control decisions regarding shared use of the data base and the data base schema.

Data Base Schema: A formally organized collection of computer-readable codes that describes the data base and guides file processes automatically. The schema is used to properly place incoming data into the files and to locate required data at a later time.

Data Base System: A mechanism to support a data base which serves the function of handling relations between data elements in multiple files.

Data Element: A particular value of a data entity (e.g., 19 for Years of Age). With each data element the attribute type has to be known to permit meaningful data processing; in this case, the attribute type is "Age."

Direct File: This file relates most closely the attribute value used in a search to the physical capabilities of direct access mechanisms. The direct access method is extremely fast since it avoids intermediate file operations, but the method forces the data to be positioned according to a single search attribute.

Duplication of Data: A simple form of achieving redundancy in computing systems.

Error-Correcting Code: The addition of a number of redundant bits (check bits) to each data element. Each check bit represents different groupings of the information bits. The groups are determined in a manner that results in the detection of which bit is in error so that it can be corrected.

File: A collection of similar records kept on secondary computer storage devices.

File Directory: The header record associated with each file that contains information describing the position and the organization of the detailed records in that file.

Hamming Distance: The number of bits which are different between the coded representation of a character transmitted and the actual code received.

Hamming Weight: The number of bits of value 1 in the coded representation of a character.

Indexed Sequential File: A file that has the property that the records in the file may be accessed either by the use of an index or in a sequential manner.

Interpreting: In the environment of the data base, interpreting is equivalent to the use of a general program that, when called upon to carry out data base manipulations, looks at the data base schema to find items and determines their relationships. SEE ALSO: *Compiling*.

Inverted File: A file of text for which complete indexes have been created.

Multi-Indexed File: A file which is accessed only through indexes; indexes can exist for all attributes for which a search may be contemplated.

Multiprogramming: A method of allocating resources in which there is concurrent execution of two or more programs simultaneously residing in the internal storage unit of the computer. The basic principle of multiprogramming is that the programs in internal storage share the available central processing unit's time and input-output units.

Multi-Ring File: A file structure oriented towards efficient finding of *sets* of records that contain some common attribute value. These subsets are explicitly chained together through the use of pointers which define some order for the members of the subsets.

Parity Bit: A redundant bit added to the bit configuration for a character in order to permit automatic detection of errors in data transmission.

Phantom File: The elimination of the data themselves if all of the available information in the file is contained in the index.

Pile File: An unstructured, minimal method of file organization in which data are collected in the order in which they arrive. The records may be of variable length, and need not have similar sets of data elements.

Record: A collection of logically related data elements that can be manipulated as a single unit during computer processing.

Redundancy: When data units contain more information than is strictly necessary in order to provide a means for checking internal consistency.

Reliability: System reliability is achieved by maximizing the probability that the system will do what it is instructed to do. Perfect reliability is achieved when a computer system, both hardware and software, always produces correct results.

Search Key: A specified attribute type and value used to retrieve records that match the search key.

Security: The protection of computer systems and their contents from destruction.

Sequential File: A file organization in which the data records are ordered into a specific sequence and the data attributes are categorized so that the individual records contain all of the data attribute values in the same order and possibly in the same position.

Stop List: A list of words that excludes from processing common words that do not in themselves indicate subject matter such as pronouns, articles, prepositions, conjunctions, copula and auxiliary verbs, and quantitative adjectives.

Timesharing: A mechanism to split jobs into slices and to only allocate the critical resources of the computing system to them when they can actively use them.

Transaction Processing: A method of resource allocation in which jobs are started as soon as possible after a request for initiation, and they are permitted to run as long as necessary to fulfill their computational requirements. When they come to a point where they have to wait for a response from terminals or files, then they must yield to the system to enable another transaction process to be started or resumed.

Transposed File: A file organized toward selection or evaluation of data according to a dimension which is orthogonal to the data entry or update dimension. This method of file organization offers another variation of the indexed file.

SECTION 6. THE OPERATIONAL INTERFACE BETWEEN USERS AND A COMPUTER-BASED INFORMATION SYSTEM

Introduction

In this section the factors likely to ensure an optimal interface between users and a computer-based information system are identified. Types of interfaces are discussed as well as training approaches that should be considered to minimize the effort required to learn how to use a system. This section also includes a delineation of the variables important in determining user acceptance, and incorporates pertinent displays from the simulated attitudinal RIS that illustrate certain desirable features of an interactive system. This section should be of particular interest to those individuals concerned with maximizing the utilization of computer terminals and other peripheral devices and with taking advantage of the potential of computer-assisted instructional techniques for teaching users, via a terminal, how to interact effectively with a computer-based information system. A summary of the major concepts presented concludes this section, accompanied by an annotated bibliography and a glossary of key terms.

Types of Interfaces

In this discussion the term *interface* is defined as the common boundary between an information system and its users. The interaction that takes place between users and the system usually occurs at a remote terminal. "As far as the user is concerned, the terminal *is* the system, and its ease of operation and responsiveness will probably be the deciding factor in acceptance of an on-line information retrieval system by its clientele. The choice of terminal for a particular application is therefore of the utmost importance" (Lancaster & Fayen, 1973, p. 9). Wolfe (1971, p. 152) has noted that the type and extent of remote terminal equipment required depends upon the users' needs. Casual use would not require a cathode-ray tube (CRT) display or high-speed printer, but rather only an alphanumeric keyboard device and a data modem. However, heavy and complex usage would require the following equipment triad: a CRT display, an alphanumeric keyboard device, and a high-speed printer.

Alphanumeric Keyboard Devices. The most commonly used alphanumeric keyboard devices are still the Model 33 ASR Teletype (a teletypewriter), the IBM 2740 and 2741 terminals which use an IBM Selectric typewriter with additional electronics to support on-line data entry and reception, and other machine-readable keyboards. A large variety of competitive devices (e.g., GE Terminet, Diablo Hytype, DEC LD-30, and the Singer printing terminal) provide similar features at the same cost, but at double the printing rates (30 characters per second). With both upper- and lower-case shifting, they have a comparatively large character set as input and printing devices. They can be operated manually or driven by paper-tape readers. Their output rates are slow compared with most CRT's and high-speed printers. Their advantages are comparatively low hardware and software costs, easy availability, low maintenance cost, and ability to produce hard-copy output. Their disadvantages are low speed, noise, and a limited display formatting capability.

CRT Terminals. CRT terminals are available in a wide range of capacities, data-transfer rates, and character sets. The character set should include all upper- and lower-case alphabetic characters, numeric characters, punctuation symbols, and control symbols. "Visual factors are particularly important to the user of a CRT display....The variables that determine image quality include luminance, contrast, regeneration rate (if a CRT is not regenerated fast enough, it gives the impression of flicker), chromaticity, resolution, and size and style of characters" (Lancaster & Fayen, 1973, p. 355). "The ideal CRT display would include a full page of text, displayed clearly and legibly, and require no eye strain. Display speed should be user governable to accommodate user reading speeds and [information] comprehension rates" (Wolfe, 1971, p. 152). However, this quantity of characters (a full page of text) is not commonly available. Twenty lines of 80 characters each is the maximum display size commonly available now. Data-transfer rates and subsequent display speeds range from 30 characters per second (300 baud rate) for teletypewriters to 960 characters per second (9,600 baud rate) for CRT's. A data-transfer rate of 1,200 baud is the maximum easily available over dial-up telephone lines. Some CRT terminals that are directly connected to the computer can achieve high-speed transmission up to 50,000 baud and higher. High-speed communication facilities also are available, but more costly.

Experienced system users tend to become impatient with less than an instantaneous response and prefer at least a 9,600 baud rate of data transfer. But, speed of system response is dependent upon more than pure rate of data transfer. Other factors that also influence speed of response will be considered later in this section, and ways to expedite the speed of man-machine interaction will be discussed.

Most CRT terminal keyboards have a set of function keys which initiate particular program functions when they are depressed. The keys can be lit from the back by means of signal lights under control of the computer program, a feature which calls attention to the function being performed. In addition to these special program functions, most CRT terminals have editing and formatting features. Typical editing functions available are DELETE CHARACTER, REPLACE CHARACTER, INSERT CHARACTER, DELETE LINE, INSERT LINE, CLEAR TO END OF LINE, CLEAR TO END OF DISPLAY, and CLEAR TO END OF MEMORY. The memory usually is at least large enough to hold one screenful of data. Additional memory allows more than one screenful of data to be created and stored as a unit. This feature permits paging through the memory contents. The memory also can be used to store an image of a form to be used for data entry. The form image is locked on the screen in format mode, and the viewer completes the form by keying in the input data. Format mode is particularly suitable for entry of data according to the same format, much of which is repetitious.

One example of the use of a CRT to display a standard form for ordering a copy of a technical report based on a review of its abstract is shown in Figure 23. This figure is a printout of a display contained in the simulated attitudinal RIS. The format was displayed on the screen, and the viewer was offered the opportunity to order a report for any abstract he had seen which interested him. In this hypothetical example, the completed order form would have been sent to the library for acquisition. Upon receipt of the report, the librarian would have either notified the requester of its arrival or forwarded it to his code number.

Selection of options or actions displayed on the screen of a CRT can be accomplished by either a controllable cursor or a light pen. A cursor is an electronically generated bright signal spot on the screen of the CRT which can be positioned anywhere on the display by the CRT operator in order to select from a list of choices, options, or possible actions to be taken. Some CRT terminals use a vertical and horizontal spacer bar to position the cursor, while others use a joy stick. A joy stick uses a lever in a control box to position the cursor at the desired location on the CRT screen. The advantage of the joy stick is that the cursor spot can be moved to the desired location directly rather than moving up or down the desired number of lines and over the desired number of spaces. A joy stick is particularly suited for use with graphics terminals. Some cursors blink on and off, while others glow continuously.

A light pen is a small photocell or photomultiplier in a pen-shaped housing connected by cable to a CRT console. By touching the pen to the face of the CRT and activating the trigger, the operator may make his selection known to the computer. Arm fatigue can result from extended operation with a light pen, but it is easy to use and fast. The nature of the man-machine interface being implemented should determine whether a controllable cursor or a light pen is the preferred mode for making selections and initiating actions.

The advantages of a CRT terminal are high data presentation speeds, easily controlled display speeds, great flexibility of display formats, and silence. Their disadvantages are special software requirements and no direct hard-copy printout capability.

An alternative to the CRT terminal is the plasma display terminal. The plasma terminal incorporates a gas-filled glass panel, upon the face of which characters are generated by means of gas discharges (Stifle, 1971). The most well-known plasma terminal is the PLATO IV terminal developed at the University of Illinois for the PLATO system of computer-aided instruction. An optional feature of this system is the touch panel, a 16 by 16 grid that is superimposed on the plasma display panel. This matrix contains 256 individually addressable positions, and a user may select any data element displayed by touching it with his finger. This feature also is available on conventional displays.

High-Speed Printers. High-speed printers typically are mechanical or electrostatic devices. Mechanical printers place characters on the paper by the impact of a print wheel or belt striking against a carbon ribbon. The display shown in Figure 23 was printed by an upper-and-lower case impact printer. The legibility is good, and the character set resembles an electrical typewriter font. Electrostatic printers tend to be faster and less noisy than mechanical printers, but they do not have as legible a font since characters are represented by a dot matrix configuration. Also, they require special paper. An alternative to these two types of high-speed printers is the ink jet printer in which fine jets of ink are sprayed on the paper in a dot matrix configuration that corresponds uniquely with each character in the character set. This printer has good legibility and flexibility, but somewhat greater maintenance requirements, and for that reason it has not become popular.

If you would like to order one of the technical reports for which you have seen an abstract, please fill out the order form below:

Author: Broedling, L. A.
Title: An Assessment of a New System for Annual Administration of the Family Housing Survey
Year: 1972
Report Number: WTR 73-7
AD Number: 750-685
Any other element of document description: UNCLASSIFIED
Your Name: New Hire
Your Code: 307
Your Phone Number: 225-XXXX

Figure 23. Display from the Simulated Attitudinal RIS of a Standard Form For Ordering a Copy of a Technical Report Based on a Review of Its Abstract.

High-speed printers usually are available with 80, 120, or 132 print positions per line, and with upper-and-lower case or upper case only options. Their use is primarily as an adjunct to on-line interaction, performing the printing when large quantities of output are desired at the remote terminal. Single- or multi-form paper stock usually is available for these printers. Katter and Blankenship (1969, p. 16) list their disadvantages for direct on-line interactive use as being space requirements, cost, a more limited character set, a comparatively limited format flexibility, and no light-pen adaptability.

Modems. Modems are modulation/demodulation devices that convert the digital signals generated by computer equipment into an analog form compatible with communication facilities (e.g., telephone lines) for data transmission, or that convert the analog signals received in a data transmission back into digital form (Lancaster & Fayen, 1973, p. 7). Modems are referred to as *data sets* by the telephone company. Since the data to be transmitted over the communication lines are in the form of audible tones, an acoustic coupler may be used. In on-line interactive situations, the acoustic coupler utilizes the data tone which is transmitted by the computer to the terminal. "Any standard telephone can be used without requiring the special installation necessary for the data set. Thus, the user can go wherever a telephone is available, dial the computer, [place the telephone cradle into the acoustic coupler,] and start processing his data" (Spencer, 1974, p. 393).

Since the acoustic coupler is portable, it offers more flexibility. However, it cannot reach the transmission rates achieved by modems. "Most couplers have a maximum transmission rate of 300 bits per second, though a few can handle rates as high as 1,800. Modem transmission rates can go as high as 1,000,000 bits per second. The speed, of course, is largely dependent on the type of communications line used....A voice grade telephone line will carry up to about 9,600 bits per second, depending on whether the line is upgraded through conditioning and whether it is a private line, a leased line, or a part of a high-speed communications network" (Spencer, 1974, p. 392).

User Surrogates. Not all potential users of an information system will want to interact with the system by means of a terminal keyboard. Some individuals dislike to type or never have learned how; others may feel that typing is beneath their dignity. They may not even feel comfortable with a light pen. Some individuals may not be willing to spare the time needed to learn to use the system. Whatever the reasons may be that mitigate against willingness to interact personally with an information system, provision should be made for users who prefer to delegate their information requests to an intermediary. This surrogate for the user himself may take the form of a colleague, a less senior staff member, or perhaps a terminal operator trained especially to assist users in learning how to use the system and in achieving their information requirements. Planning for the possible need for such an intermediary at the user-system interface is discussed under training considerations later in this section and under system staffing requirements in Section 7.

Learning How To Use an Information System

Acceptance of an interactive information system will be critically affected by the effectiveness of procedures employed initially to teach potential users how to use the system. All users will not have the same needs to use the system. Therefore, not everyone who will be using the system needs to be trained in the same depth. A strong case can be made for preparing the instructional materials in several levels of complexity. At the first level the new user would be provided with the bare minimum of instruction required to log on and off the system, learn the basic repertory of commands and their names, and the system responses to be expected. "The casual or infrequent user will not want to spend a great deal of time studying a printed guide before he can get to the terminal" (Lancaster & Fayen, 1973, p. 316). Thus, with little sacrifice of his time, the neophyte user should be able to begin interacting with the sytem in a simple way. If he makes mistakes or becomes confused, assistance should be available from the system itself and/or from a training specialist. At this initial minimal level of sophistication in using the system, routine searches and explorations of information in the system could be conducted with the knowledge that help was always available if needed. Some users would never have to progress beyond this level of competence. However, other users, because of the complexity of their own information needs or perhaps because of their own growing interest in learning more about the system's capabilities, might want to increase their understanding of how to use the system more effectively. A higher level of instructional materials should be available to this class of users so that they can go on to study and learn the more sophisticated features the system has to offer. At the highest level of knowledge, procedures for training personnel who are to be the most frequent users of the system may be comprehensive, involve a substantial commitment of time, and require continuing education.

Three types of training approaches should be considered, and possibly all of them should be implemented.

1. Printed instructional manuals.

2. Personal instruction by a training specialist.

3. A tutorial display sequence presented on-line at the terminal itself.

Printed instructional manuals should be concise, simply written, and easy to understand. They should provide realistic examples of typical user-system interactions that are possible, and exercises at the terminal to become acquainted with the system's capabilities. The manuals should be self-instructional so that the learner can proceed at his own pace. A guide to what went wrong should be included for typical errors likely to be made by the neophyte system user.

Lancaster and Fayen (1973, p. 234) claim that user manuals tend to be useful as reference tools, but not very effective as training devices. They argue that there is no substitute for hands-on experience at the terminal. Personal instruction by a training specialist, while likely to be effective,

also is likely to be expensive. However, there is a clear need for a training specialist position in an information system facility. This individual should have the most comprehensive knowledge of the system's capabilities and how to exploit them. The job duties of the training specialist would be first to develop the curricula and materials for initial training of users as the system is first implemented. This person subsequently would be responsible for providing assistance to individual users, for acting as an intermediary or surrogate for individuals who do not choose to interact directly with the system, and for training new system users. Continuing education of regular users to new system features and capabilities also would be a job duty of the training specialist. Thus, it appears that there are more than enough ongoing tasks to justify the creation and support of a training specialist position.

A user can be taught how to interact effectively with the information system by means of a tutorial display sequence presented on-line at the terminal itself. Lancaster and Fayen (1973) offer the following possibilities:

"1. Use of the terminal to display a conventional set of instructions that could equally well be presented in conventional printed form.

2. Use of computer-aided instruction (CAI) techniques, either to give the user a one-time introduction to the system or to lead him by the hand in the conduct of an actual search.

3. Incorporation of explanations of specific commands or system features that the user can call up when he needs them" (p. 321).

With regard to the first possibility suggested by Lancaster and Fayen, Katter and Blankenship (1969, p. 19) recommend that instructional materials displayed on the terminal should provide page references to the same material covered in instructional manuals. Lancaster and Fayen (1973, p. 316) further suggest that outline instructions relating to the mechanics of using the system, including major system commands, be printed in large type on a card displayed prominently on or adjacent to the terminal. A fourth possibility might be to use the system as a prompter. When the system notes a failure to communicate, then appropriate suggestions could be presented to the user based on the current context of the interaction.

A number of information specialists have emphasized the desirability of using computer-assisted instructional techniques to guide users in learning how to interact with an information system. Lancaster and Fayen (1973, pp. 214 & 234) recommend the use of CAI techniques in training users on-line at the terminal, and Caruso (1970) concludes that teaching new or infrequent users how to operate an information system requires a tutorial approach. "We decided to take a frankly tutorial approach to the problem of user-system interaction, utilizing as appropriate, the current understanding of computer-aided instructional techniques---particularly those techniques which utilize simulation of the actual system to be taught" (p. 100).

Wolfe (1971) has the following to say about the characteristics and capabilities that should be included in on-line information retrieval and display systems. His observations are based on a two-year study of remote

access on-line time-shared information systems and their impact on the user and on alternative batch-mode processing systems. His study was supported by the Naval Ship Systems Command.

> "The tutorial sequence should provide enough background and instruction to train a user completely unfamiliar with the system. It should explain the origin and content of the data base, provide instruction in query formulation, and thoroughly describe the system's capabilities, language, and limitations. In addition to a reference display of sample queries, data field identification tables, and display format models, the tutorial sequence should include all the material now provided by on-going systems in the form of printed instruction manuals. For the new user it should provide a computer-aided instructional course in query formulation using a sample data base to test the new user's skill and understanding of the system before allowing him access to the full data base. The importance of the tutorial sequence cannot be overstressed. It provides the basis for user-computer interaction and by doing so determines in great part the success of potential users of the system. A good tutorial sequence will create user self-confidence, increase the efficiency with which the system is used, and greatly expand the system's marketability" (p. 149).

Lancaster and Fayen (1973) admonish that "it is imperative that on-line systems should have both the experienced and the inexperienced user in mind. Terse and verbose communications options should be available, and simplistic features designed to assist the beginner should be capable of being bypassed by the more experienced searcher" (p. 364). Others involved in the implementation and evaluation of on-line interactive information systems also conclude that an instructional mode as well as an abbreviated, short-cut mode of user-system interaction should be provided (Licklider, 1965, p. 36; Mesel & Wirtschafter, 1974, p. 8; Wolfe, 1971, p. 150).

A provision for the continuing education of users also must be made. Both Lancaster and Fayen (1973, p. 322) and Katter and Blankenship (1969, p. 36) suggest a newsletter be circulated to regular users describing system refinements, improvements, and innovations as they are implemented. Lancaster and Fayen (1973) further recommend "it is also desirable that the system have some type of NEWS command, whereby the user can discover, at the terminal itself, the latest news on system features and capabilities" (p. 322).

A final word of advice is that when a new information system is being implemented, system users should not be trained too far in advance of the system's actual installation. There have been some instances where the training program was scheduled and conducted only to find that the delivery schedules for the system hardware and software had slipped. Thus, by the time the system actually came up, users had forgotten what they had learned and needed to be retrained. In addition to resenting the loss of time required for refresher training, their initial enthusiasm and interest in the system had waned. Therefore, it is important to schedule the training program to be conducted immediately before the system is actually introduced. And the system itself should be thoroughly debugged and operationally reliable before introduction in order to avoid initial disenchantment by new users who had expected a trouble-free and dependable level of performance.

Acceptance Variables

Availability of Assistance. The acceptance of an on-line information system by users will be influenced by the availability of assistance when needed. Lancaster and Fayen (1973, p. 322) specify that the inexperienced user should be able to request assistance from a system specialist (perhaps the training specialist) when he gets into difficulties. Katter (1970, p. 10) suggests the possibility of a "helper" role also and perhaps some systematic procedures for fostering mutual-help arrangements in the form of "on-line apprenticeships." Often an informal form of mutual help develops in which more experienced users offer advice and assistance when they notice that a less-skilled user is experiencing difficulties. Or infrequent system users may ask a more experienced user how to perform certain aspects of system interaction.

A HELP command also should be available as part of the command vocabulary of the system. Thus, when a user needs help in interacting with the system, the system itself would provide this assistance in response to the user's initiation of the HELP command by providing cues and explanations.

Coverage of Information System Holdings. The coverage of holdings in an information system should be extensive if not exhaustive of the data bases of interest to the user population. If some of the major data bases are in the system and other are not, the user has to deal with two systems essentially, and communication between them is difficult if not impossible. However, the requirement of comprehensive coverage does not imply that all data bases must necessarily be maintained on-line. Only the data bases experiencing high utilization rates or those of high value or interest would need to be maintained on-line; infrequently accessed data bases could be maintained off-line on magnetic tape to be mounted on tape drives only upon request.

Huge data bases may not have to be maintained on-line unless access to them is a frequent requirement. In this latter case, they probably should be maintained on a dedicated disk pack permanently mounted. Another attractive possibility might be to maintain on-line only a representative sample of a very large data base. Thus, the system user could learn about the characteristics of the data base before actually initiating a search request that might prove to be a major undertaking. This system design feature would help to prevent unnecessary searches or searches initiated without adequate understanding of the amount of computer resources needed to accomplish the search. Perhaps a computer-calculated estimate of the cost to the user for executing a particular request for information might act as a deterrent to unnecessary information requests. Various alternatives for establishing a charge structure for services provided by an attitudinal research information system are discussed in Section 8.

Ease of Operation. The continued acceptance of an information system will depend upon whether or not the effort of learning to use it and being able to use it easily is more than offset by the benefits the system offers in terms of access to desired information. "The consideration of the effort required to use the system is particularly significant where there is direct use of computers by professional end-users such as lawyers, physicians,

scientists, or researchers. Here one has to avoid a diminution of the time that a professional has available for his primary interests. There have been a number of instances where data base systems intended for direct use by professionals have been installed and were later augmented with clerical personnel at the terminals. The costs and delays attendant in the use of intermediaries to operate the terminals seriously affects system economics" (Wiederhold, in preparation).

A number of features can be incorporated into the design of an information system to increase its ease of operation. Hansen (1971, p. 528), as one of four user engineering principles, stipulates that the requirement for memorization on the part of the user should be minimized. One important way that the system can augment the user's memory is to present him with lists (menus) from which he makes a selection rather than typing a string of characters to indicate his choice or wishes. By selecting from a list, the user is spared the burden of having to remember, for example, the repertory of commands available or the options available at different branches in the decision tree structure. Errors attributable to mistyping can be reduced dramatically by the use of displays that provide selection capability, and since a CRT terminal can display many more characters in the same time that it would take a user to type a few, the user's productivity will increase even as his error rate diminishes. Figures 4, 9, 10, 12, 16, and 17 in Section 3 of displays in the simulated attitudinal RIS all show examples of a list of choices from which the user selects the information or option that he wants to access next.

Figure 24 shows an example from the simulated attitudinal RIS where it was not practical to display all of the selections possible. In this example, the hypothetical user is exploring the possibility of searching a data base consisting of narrative summaries for 60 drug abusers treated at the Naval Drug Rehabilitation Center, Miramar using the NAVY MEDISTARS dictionary of descriptor terms. There are over 550 descriptor terms in the NAVY MEDISTARS dictionary, so instead the user is asked in this figure to specify the term that he wishes to search on, and he must type it in. The system then performs a character-string recognition by matching the word typed in to all of the legal terms in the NAVY MEDISTARS dictionary. In this example, no exact match for the word typed in was found, but two partial matches are displayed for the user's consideration. The first alternative matched out to seven character positions while the second alternative matched out to six character positions. The user then can decide whether he wants to see definitions for these two terms, or whether he wants to see the complete subject index of the NAVY MEDISTARS dictionary. If he had chosen the latter option, the system would have informed him that an extensive printout was required and might have given him the opportunity to reconsider his request since a copy of the dictionary might be easily accessible elsewhere.

Katter and Blankenship (1969, p. 30) recommend the incorporation of a variable spelling approximator in the design of an on-line information system. They view this feature as an aid to the user in minimizing the problems introduced by minor spelling errors. They suggest an approximative algorithm for matching a word typed in by a user to a file of stored words. If no exact match is found, the user is notified, and any words in the file that are close

NARRATIVE SUMMARIES OF DRUG ABUSERS TREATED AT THE
NAVAL DRUG REHABILITATION CENTER, MIRAMAR, CALIFORNIA

The narrative summaries of 60 drug abusers who sought amnesty and signed exemption papers, thus obligating them to a period of rehabilitation, were processed by the NAVY MEDical Information STorage And Retrieval System (NAVY MEDISTARS). This data base can be searched using any boolean combination of descriptor terms in the NAVY MEDISTARS dictionary. Contact Dr. E. K. Eric Gunderson, Navy Health Research Center, 225-6559, for a copy of the NAVY MEDISTARS dictionary, or see "Ramsey-Klee, Diane M. NAVY MEDical Information STorage And Retrieval System (NAVY MEDISTARS): Manual of Indexing Terms. Technical Report No. 1-71: Part Two, April 1971, R-K Research and System Design, Contract Nonr N00014-70-C-0291 - Project NR 153-316, Office of Naval Research. AD 724 306." See also DRUG ABUSE in the QUESTIONS AND QUESTIONNAIRES INDEX OF SEARCH TERMS.

The first stage of the search provides a count of how many records match the search request. If the number of records is too large, the search can be narrowed by requesting to see the hierarchical arrangement of the NAVY MEDISTARS dictionary. In the same manner a search can be broadened if too few records are retrievable in response to a search request.

Please specify the descriptor term that you wish to search on. The system will advise you if it is a legal term in the NAVY MEDISTARS dictionary.

DEFENSES

No, DEFENSES is not a legal term in the NAVY MEDISTARS dictionary. However, the dictionary contains the terms DEFENSE MECHANISMS and DEFENSIVE. There are two options available to you at this point. Indicate your choice by placing an "X" to the left of the option below that you select.

X I want to see definitions for the terms DEFENSE MECHANISMS and DEFENSIVE.

___ I want to see the complete SUBJECT INDEX of the NAVY MEDISTARS dictionary.

Figure 24. Example of a Display from the Simulated Attitudinal RIS Where It Was Not Practical To Show All of the Possible Selections.

approximations for the entered word are displayed for his consideration. The example shown in Figure 24 portrays how this feature could be used to determine the legality of a search term.

Figure 25 shows another useful feature that should be part of the design concept for an information system. Figure 25 represents the display that the user requested in Figure 24 when he indicated that he wanted to see definitions for the terms DEFENSE MECHANISMS and DEFENSIVE. Not only does the system provide the requested definitions, it also advises the user how many records are retrievable if a search on either of these two terms were to be initiated. In the example shown in Figure 25, from the number of records retrievable the user may decide that the term DEFENSE MECHANISMS is too general for his purposes, and a search on a specific defense mechanism may be more appropriate. The system has already taken this possibility into account and offers the user the opportunity to view the hierarchy of terms under the more general term DEFENSE MECHANISMS. The user should be afforded the chance to narrow the scope of his search or to broaden the scope by hierarchical expansion.

In Figure 26 the simulated attitudinal RIS displays the hierarchical arrangement of the NAVY MEDISTARS dictionary for the descriptor term DEFENSE MECHANISMS. From this display the user can determine that two specific defense mechanisms---DENIAL and EXTERNALIZATION---are prominent in the records of a sample of drug abusers undergoing treatment at the Naval Drug Rehabilitation Center, Miramar. The hypothetical user in this example decides to narrow his search to the term DENIAL. The system responds by asking the user if he wants to see a definition of this term before initiating the search. The user decides that he does, and the system displays the definition. The system also reminds the user that 23 records are retrievable in response to a search on this term, and gives him the option of seeing the first four records as a sample, or retrieving all 23 records by a search of the tape file for this data base. The user chooses to view the sample of four records, and subsequent interaction with the system would allow him the option of obtaining a hard-copy printout of the four sample records.

Error Messages. A well-designed on-line information system should minimize user effort by mechanisms for noting and compensating for common errors and by procedures for correcting errors that do occur (Lancaster & Fayen, 1973, p. 133). Hansen (1971) makes an even stronger case for the advisability of engineering for errors:

> "The system design must protect the user from both the system and himself. After he has learned to use a system, a serious user seldom commits a deliberate error. Usually he is forgetful, or pushes the wrong button without looking, or tries to do something entirely reasonable that never occurred to the system designer. The learner, on the other hand, has a powerful, and reasonable, curiosity to find out what happens when he does something wrong. A system must protect itself from all such errors and, as far as possible, protect the user from any serious consequences. The system should be engineered to make catastrophic errors difficult and to permit recovery from as many errors as possible" (p. 530).

DEFENSE MECHANISMS (87 records retrievable)

Any general statement referring to a behavior pattern, which operates unconsciously, employed by a person to seek relief from emotional conflict and to obtain freedom from anxiety.

DEFENSIVE (11 records retrievable)

Any statement referring to the characteristic trait of a person to excessively reject criticism of oneself, or to use behavior that shifts attention away from another behavior, notice of which would cause embarrassment, discomfort, or shame.

Please specify which of these two descriptor terms you want to search on.

DEFENSE MECHANISMS

Do you want to see the hierarchical arrangement of the NAVY MEDISTARS dictionary for the term DEFENSE MECHANISMS before initiating the search?

Yes

Figure 25. Example of a Display from the Simulated Attitudinal RIS Showing the Number of Records Retrievable Before a User Commits the System to an Actual Search and Retrieval.

HIERARCHICAL ARRANGEMENT OF THE NAVY MEDISTARS DICTIONARY
FOR THE DESCRIPTOR TERM 'DEFENSE MECHANISMS'

		Number of Records in Drug Abuse Sample
	DEFENSE MECHANISMS	87
	COMPENSATION	0
	CONVERSION (See also HYSTERICAL NEUROSIS, NAVY BUMED)	1
X	DENIAL	23
	DISPLACEMENT	1
	DISSOCIATION (See also HYSTERICAL NEUROSIS, NAVY BUMED)	1
	EXTERNALIZATION	30
	IDENTIFICATION	2
	INTELLECTUALIZATION	5
	INTROJECTION	0
	PROJECTION	4
	RATIONALIZATION	9
	REACTION FORMATION	1
	REGRESSION	1
	REPRESSION	1
	RESISTANCE	2
	SUBLIMATION	0
	SUBSTITUTION	0
	UNDOING	0

(NOTE: Place an "X" to the left of the descriptor term that you select for the search.)

Do you want to see a definition of the descriptor term DENIAL before initiating the search on this term?

Yes

DENIAL

Any statement referring to the defense mechanism by which there is unconscious resolving of emotional conflict and to the allaying of anxiety by denying a thought, feeling, wish, need or external reality factor which is consciously intolerable.

(NOTE: There are 23 NAVY MEDISTARS records retrievable in response to a search on the descriptor term DENIAL. You may see the first 4 records now as a sample. If you want to retrieve all 23 records, it will require initiating a search of the NAVY MEDISTARS tape containing the narrative summaries of all 60 drug abusers contained in the data base. Indicate your choice by placing an "X" to the left of the option below that you select.)

X I want to see the first 4 records now as a sample.

I want to retrieve all 23 records from the tape file.

Figure 26. Example of a Display from the Simulated Attitudinal RIS Showing How a Hypothetical User Might Narrow His Search by Selecting a More Specific Search Term.

Hansen further stipulates that the first principle in error engineering is to provide good error messages. They serve both as an invaluable training aid to the learner and as a gentle reminder to the expert.

Lancaster and Fayen (1973, p. 362) emphasize that when a system user does make an error, he must be informed immediately by the system. But it is not enough to tell a user that an error has occurred. He must be told the precise nature of the error and what he must do to correct it. Error messages should be explicit and should tell the user how to make the appropriate correction. Error-correction procedures should be simple and should disturb as little of the search as possible. Under no circumstances should an error cause an entire search to be aborted and force a user to return to an initial log-in status. Additionally, errors must not be allowed to propagate within a search.

Wiederhold (in preparation) recommends that the formats of different response messages should contain significant differences in the initial words of the message so that occasional error messages are immediately distinguishable from repetitive confirmations or other system signals. He further suggests that an audio error signal can be very helpful as an adjunct to visually displayed error messages. The use of voice-answer back devices to provide diagnostic error messages, as well as confirmatory and prompting signals, also is a possibility to consider.

Hansen (1971, p. 531) concludes that it is not enough to tell only the user of his errors. The system designer needs to know the nature of user errors too so that he can apply the principle of engineering out the common errors. In this regard, on-line monitoring operations can reveal the types of problems that occur most often. "For improved design of system interface characteristics to eliminate system-detectable errors, a program that stores records of user errors by type can be valuable, especially during the developmental stages of the system" (Katter & Blankenship, 1969, p. 30). The monitoring program could compile a log or record of errors made by users (e.g., invalid commands or incorrect actions) along with the frequency of occurrence of each type of error. Very frequent errors could be identified in this manner and steps taken to modify the system in order to eliminate them. "Clearly, such information has great potential value in identifying errors or misunderstandings that occur significantly often. Such errors will suggest ways in which the system, or instructions on how to use it, may be improved" (Lancaster & Fayen, 1973, p. 183).

Lancaster and Fayen (1973, p. 189) conclude that it seems reasonable for the system itself to monitor and collect data on the behavior of users *in the aggregate* as opposed to individually identifiable users. Such aggregate data would include statistics on how the system is used---not just how many times it is used but which files are consulted, which commands are employed, and even the frequency with which various index terms are used. A tabulation of the types of problems encountered might include, for example, the frequency of use of various types of error messages, of the HELP command and the specific types of help requested, and of the EXPLAIN command. These authors argue that this type of aggregate monitoring certainly seems justifiable and is of great importance to system managers in showing how the system is used and what might be done to improve its performance.

Another possibility might be to incorporate a command into the system by means of which the user could make suggestions for system improvement. In the SPIRES system, if a user wishes to record such a message, he keys in the command TO SPIRES and follows it with his message (Parker, 1970). It has been found that comments received in this way provide useful suggestions for system improvement or indicate ways in which instructions to users should be changed to minimize confusion.

Printout Capability. An optimal interface between the information system and users should include a printout capability. Katter and Blankenship (1969) emphasize the need for adjunctive output. "The display memory needs to be easily accessible by hard-copy reproduction equipment, so that any desired portions can be rapidly specified and output in hard copy" (p. 36).

Figures 27 and 28 demonstrate the use of this principle in the simulated attitudinal RIS. In Figure 27 the hypothetical user has requested a display of the author index for abstracts of attitudinal research reports. The system presents the user with a list of names of authors from which he may make a selection. The system also supplies the number of abstracts in the system for each author on the list. The user is given the further option of limiting the extent of abstracts retrieved by date. In this case, the user chooses to see only those abstracts written in 1966 or later. The system then displays the information shown in Figure 28. Seven abstracts are retrievable in response to the user's request, the first two of which are contained in this display. The user reads them and decides if he wants a hard-copy printout of the display. He then may continue to browse through the remaining five abstracts by calling up sequential displays, each of which can be printed or not at the user's discretion. The printout capability provides the user with a permanent record of useful information obtained during his interaction with the information system.

Response Time. Speed of response of an interactive information system will have a bearing on user acceptance and continued use. Carbonell, Elkind, and Nickerson (1968) state that "the response time is usually assumed to be the time elapsed from entering a command until its completion, the latter being characterized by the production of an output or other signal to the user, and the transfer of control to him" (p. 138). Response time in an on-line, timesharing information system will be dependent on the following factors: speed of the computer, data transfer rates, amount of core memory available, the operating system, file organization and type of file storage, the efficiency of search algorithms, the command language, number and type of users on-line at a particular time, and priority given to the user by the operating system. A mathematical analysis of response time in on-line systems for various types of file organizations has been made by Cordaro and Chien (1970) and by Wiederhold (in preparation). Higgins and Smith (1971) maintain that system response should occur within 3 to 4 seconds after a command is executed by a user. They recommend that in the case of a complex command, which involves a greater delay, the user should be informed that the amount of computing required will cause a delay to occur. A system that responds within 5 to 10 seconds is operating at the upper bound of acceptable speed of response for a single transaction returning valuable information to the user. User tolerance for response times greater than 15 seconds is low, particularly if users are not accustomed to this delay. For this reason, Lancaster and

AUTHOR INDEX FOR ABSTRACTS OF
ATTITUDINAL RESEARCH REPORTS

	AUTHOR NAME	NO. OF ABSTRACTS
	Braunstein, C.	14
X	Broedling, L. A.	7
	Farr, B.	1
	Foley, P. P.	2
	Goldsamt, M. A.	5
	Katz, A.	6
	Marshall, C. T.	4
	Mohr, E. S.	1
	Muldrow, T. W.	7
	Rafacz, B. A.	2
	Schneider, J.	2
	Somer, E. P.	11
	Ware, S. B.	5
	Wilcove, G. L.	2

(NOTE: To retrieve abstracts, place an "X" to the left of the author's name.)

You have the option now of narrowing your search to abstracts written before 1966 or those written in 1966 or later. Place an "X" to the left of your choice as indicated below. You may choose both options by placing an "X" to the left of both choices.

Abstracts written before 1966

X Abstracts written in 1966 or later

Figure 27. Display from the Simulated Attitudinal RIS of the Author Index for Abstracts of Attitudinal Research Reports.

THERE ARE SEVEN ABSTRACTS WRITTEN IN 1966 OR LATER AUTHORED BY BROEDLING, L. A.

1. Broedling, L. A. and Mohr, E. S. "Response Errors to Factual Survey Questions and Accuracy of Information in the Navy's Automated Personnel Records." WTR 73-46, June 1973, (UNCLASSIFIED).

Two types of errors present in mail surveys of naval personnel were studied in this investigation. One was the errors in the responses given to survey items on factual characteristics (e.g., pay grade, marital status, etc.). The other was the errors present in information in the Navy's automated personnel records which are used for survey sample selection and data analysis.

Response errors were assessed by obtaining two sets of responses to the factual survey items from the same group of people and also by comparing questionnaire responses to information in the personnel jackets. Errors in the automated records were assessed by comparing their information to the personnel jacket information.

The results showed a wide variance across items in the average amount of response error. Items inquiring about present and/or stable characteristics had high consistency of response, while items pertaining to past characteristics or ones subject to change generally had substantially lower consistency. Tests of statistical significance showed no difference in response consistency on the basis of sex or time interval between questionnaires but showed non-Caucasians to be less consistent than Caucasians and enlisted personnel to be less consistent than officers. More inaccurate information and more missing data was found in the automated records than in either the personnel jackets or questionnaire responses.

The implications of these results for survey research were discussed.

2. Broedling, L.A. "The Uses and Abuses of the Term Reliability in Survey Research." October 1972. A&MRDRN-372.

In survey research, the concept of reliability has been used to denote both measurement error variance and sampling error variance. The author discusses the conceptual similarities and differences between them. The research on the psychometric reliability of survey items is summarized, and the need for more practical and theoretical work on this topic is stressed.

Figure 28. Example of a Printout from the Simulated Attitudinal RIS of Selected Abstracts.

Fayen (1973, p. 350) feel that it is important for some form of *Please Stand By* message to be transmitted by the system as soon as possible after it is known that processing delays will occur. The reason for the delay should be given or knowledge of where the user is in the waiting queue. Users generally will stand by as long as they have a valid expectation that the system will respond eventually.

"Users of a time sharing system particularly dislike unpredictable response times (due to variable loads on the system). It has been observed that they usually prefer a constant delay to a possibly shorter but variable one; unpredictable conditions disturb the user and [interfere] with his efficient use of the computer. The above assertion can be interpreted by saying that if delays are long but predictable, a user can conceivably carry on some other activity instead of wasting time waiting for a result that may come now or later..." (Carbonell et al., 1968, p. 138). No matter what the length of the required delay period, Katter and Blankenship (1969) assert that users would prefer to be able to predict it accurately. "...machine response times are usually not instantaneous, so that the user has the possibility of anticipating this fact and using the delay either for rest (relaxing his attention) or for other work, whether it is related or unrelated to the on-line problem-solving process" (p. 27). Different work habits will evolve depending on the interface that the system provides (Wiederhold, in preparation).

In addition to response time, Carbonell and his associates (1968, pp. 135-137) discuss the factors affecting accessibility to a timesharing system. Monitoring of user behavior shows that typically there is a high demand to access the system in the mid-morning hours, followed by a lull over the noon hour. Another peak of activity occurs after lunch, tapering off as the work day draws to a close (e.g., see Carbonell et al., 1968, p. 136; and Mesel & Wirtschafter, 1974, pp. 27, 29-31). At particularly busy times of the day, a user may find himself waiting in line to get on a terminal. Once he signs on the system, he may experience further frustration waiting for the system to respond as it trys to handle a capacity user load. "Response times are one of the most important elements influencing users' behavior, the amount of work they are able to accomplish, and their degree of satisfaction with a time-sharing system" (Carbonell et al., 1968, p. 137).

However, a user may circumvent the queuing problem by changing his work patterns. He may come into the office early in order to guarantee access to a terminal. He may eat his lunch early or late in order to take advantage of the easier access to terminals during the regular lunch hour. Or he may stay later in the day or even come back to work in the evening in order to use the system without the pressure of competition from other terminal users. These kinds of adaptations in work patterns that users make in a very busy system environment have implications for optimal system management. For example, it may prove advantageous to schedule user time on available terminals. In effect, a user would be given a guaranteed appointment time for his system interaction. However, this practice would certainly obviate spontaneity of user-system interaction and probably would deter most users from becoming frequent or avid system users. Another management action to consider taking would be to stagger the daily working hours of professionals. This practice also is likely to meet with less than enthusiastic reception unless the pro-

fessional himself can decide which hours he prefers to work. Such severe actions to ease the problem of queuing to get on the system would only be necessary under extremely heavy workload conditions. However, information system designers should not discount the queuing problem completely, because sooner or later it will rear its ugly head, most likely during mid-morning and in the early afternoon.

User Costs and Benefits. Katter (1970, p. 7) suggests that most likely the new information system user will assume a low-risk, noncommitted, provisional stance toward his new experience. While not excessively ready to find fault, he will tend to reaffirm the positive aspects of more familiar manual procedures for obtaining needed information. Thus, he will be making an implicit if not explicit comparison between the old and new ways of obtaining information as each method benefits him or incurs personal costs. In assessing the acceptability of the new system, he will tend to weigh the costs of his training time in learning to use the system; his time spent formulating search strategies, considering appropriate analysis procedures, and describing desired output formats; his time lost waiting to use the system or waiting for the system to respond; and his actual productive time spent at the terminal against the value received by being able to achieve increased access to needed information in improved ways on a more timely basis. The cost-benefit ratio for the new system as determined by each user will depend on his perception of how the system benefits him personally weighed against the effort he must expend in order to achieve these benefits. Consequently, personal cost-benefit ratios may vary considerably over a particular user population, but in the aggregate will ultimately determine if an information system is accepted by those individuals it is intended to serve. Cost-benefit considerations for the possible implementation of an attitudinal RIS are discussed in detail in Section 8.

Future Prospects for Acceptance

It is true today that, when an on-line information system is first implemented, most of the potential users will have had little or no previous experience with any kind of interactive terminal. Some potential users will be attracted to the new technology because of its novelty and initial promise of a better way to obtain information. Other potential users may be intimidated by the "mysteries" of the operation of the system, as well as being unwilling to change from comfortable old ways of doing things. However, Lancaster and Fayen (1973) make the following optimistic prediction about the likelihood of growing user acceptance of on-line information systems in the future.

> "This entire situation is likely to change dramatically within the next decade as on-line systems are used increasingly for educational purposes in universities, colleges, and even high schools and grade schools. Within a very short time, we will encounter a breed of scientist and engineer who has been raised with on-line computers and to whom the terminal is just another tool that is readily available for exploitation. Problems of user acceptance, important now, will be virtually nonexistent in the near future" (p. 350).

Wolfe (1971) further predicts that "automatic search formulation based on on-line user feedback, statistical analysis, and heuristic optimization procedures will undoubtedly develop into the primary search formulation method for all future on-line information systems regardless of their data base content. The concept involved is extraordinarily practical in terms of potential use and personnel cost reduction. Eventually a generalized query formulation program will be developed and made adaptable to every kind of data base" (p. 151).

While these predictions bode well for the future use and acceptance of on-line information systems, the present-day system designer should be on his mettle to take into consideration all of the variables that have been discussed in this section. Attainment of the goal of achieving an optimal operational interface between users and a computer-based information system will depend on how well all of the human factors described here have been taken into account in the system design.

Summary

The term interface is defined as the common boundary between an information system and its users. The following classes of equipment were described to effect interaction between the user and the system at the interface: alphanumeric keyboard devices with ability to print, cathode-ray tube terminals, plasma display terminals, high-speed printers, and modems (modulation/demodulation devices).

Not all potential users of an information system will want to interact with the system personally. Therefore, provision should be made for users who prefer to delegate their information requests to an intermediary (user surrogate).

Acceptance of an interactive information system will be critically affected by the effectiveness of procedures employed initially to teach potential users how to use the system. All users will not have the same needs to use the system. Therefore, not everyone who will be using the system needs to be trained in the same depth. A strong case can be made for preparing the instructional materials in several levels of complexity. In addition, three types of training approaches should be considered, and possibly all of them should be implemented: (1) printed instructional manuals, (2) personal instruction by a training specialist, and (3) a tutorial display sequence presented on-line at the terminal itself. This last training approach might very well employ the use of computer-assisted instructional techniques, in particular simulation of the actual system to be taught.

Designers of on-line systems should have both the experienced and the inexperienced user in mind. An instructional mode as well as an abbreviated, short-cut mode of user-system interaction should be provided. A provision for the continuing education of users in new system features and capabilities also should be made. A final word of advice is that the program to train new users should be conducted immediately before the installation of the fully operational system.

A number of variables affect user acceptance of an on-line information system. The first of these variables is the availability of assistance, either from a system specialist or from the system itself in response to the initiation of a HELP command by the user. The comprehensiveness of data base holdings in an information system also will influence user acceptance.

A third variable affecting user acceptance is ease of system operation. A number of features can be incorporated into the design of an information system to increase its ease of operation. The requirement for memorization on the part of the user should be minimized. One important way that the system can augment the user's memory is to present him with lists from which he makes a selection rather than typing a string of characters to indicate his choice or wishes. If a user must type in his selection, minimum character-string recognition should be employed to spare him from having to type the entire word as well as provision for a variable spelling approximative algorithm. If no exact match is found for the word typed in, the user is notified, and any words in a stored file that are close approximations for the entered word are displayed for his consideration.

Another feature contributing to ease of system operation is the provision for a user to broaden or narrow his search based on knowledge of how much information is retrievable in response to a search request before the request is actually initiated.

A well-designed on-line information system should minimize user effort by mechanisms for noting and compensating for common errors and by procedures for correcting errors that do occur. When a system user does make an error, he must be informed immediately by the system. He must be told the precise nature of the error and what he must do to correct it. Under no circumstances should an error cause an entire search to be aborted and force a user to return to an initial log-in status. Additionally, errors must not be allowed to propagate within a search.

On-line monitoring operations can reveal the types of problems and errors that occur most often, and will suggest ways in which the system, or instructions in how to use it, may be improved. A form of aggregate monitoring, as opposed to monitoring of individual users, seems justifiable and of importance to system managers in showing how the system is used and what might be done to improve its performance.

Another variable influencing user acceptance of an on-line information system is whether or not provision has been made for adjunctive output from the terminal display by means of a hard-copy printer. The printout capability provides the user with a permanent record of useful information obtained during his interaction with the information system.

Speed of response of an interactive information system will have a bearing on user acceptance and continued use. Response time in an on-line, time-sharing information system will be dependent on the following factors: speed of the computer, data transfer rates, amount of core memory available, the operating system, file organization and type of file storage, the efficiency of search algorithms, the command language, number and types of users on-line

at a particular time, and priority given to the user by the operating system. A system that responds within 5 to 10 seconds is operating at the upper bound of acceptable speed of response for a single transaction returning valuable information to the user. The user should be advised to stand by if the delay is longer than 15 to 20 seconds. He should be given the reason for the delay or knowledge of where he is in the waiting queue. No matter what the length of the required delay period, users prefer to be able to predict it accurately. Thus, the anticipated delay period can be used for rest and reflection or for other work related or unrelated to the ongoing process.

Monitoring of user behavior shows that typically there is a peak of user activity during mid-morning and in the early afternoon. At particularly busy times of the day, a user may find himself waiting in line to get on a terminal or waiting for the system to respond after he does sign on. However, a user may circumvent the queuing problem at a terminal by changing his work patterns ---coming into the office earlier, changing the hour of his lunch break, staying later in the day, or even coming back to work in the evening.

In assessing the acceptability of a computer-based information system, the new user will tend to weigh the costs of his training time; his time spent formulating search strategies, considering appropriate analysis procedures, and describing desired output formats; his time lost waiting to use the system or waiting for the system to respond; and his actual productive time spent at the terminal against the benefit received by being able to achieve increased access to needed information in improved ways on a more timely basis. Personal cost-benefit ratios may vary considerably over a particular user population, but in the aggregate will ultimately determine if an information system is accepted by those individuals it is intended to serve.

It has been predicted that within a decade, with the advent of a new breed of scientists and engineers reared with on-line computers from grade school through college, the problem of user acceptance of this technology will evaporate. While this prediction bodes well for the future use and acceptance of on-line information systems, the present system designer must continue to take into account all of the human factors described here in order to ensure attainment of the goal of achieving an optimal operational interface between users and a computer-based information system.

ANNOTATED BIBLIOGRAPHY

1. Bean, J. W., Kidd, S. W., Sadowsky, G., & Sharp, B. D. *The BEAST: A User Oriented Procedural Language for Social Science Research.* Washington, D.C.: The Brookings Institution, June 1968.

This paper describes a user-oriented language and computer system designed explicitly for social science research, and reviews the assumptions underlying its development. BEAST, the Brookings Economic And Statistical Translator, attempts to integrate various steps of social science data analysis into a single uniform system within which to handle data, and attempts also to be a "high-level" language whose meaning is transparent at a single reading.

2. Bennett, J. L. The user interface in interactive systems. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 7). Washington, D.C.: American Society for Information Science, 1972. Pp. 159-196.

This review focuses on the relationship between user behavior and the design of the interface facility in interactive computer systems. Current user studies reveal that both user and designer will benefit when designers better understand the user's needs, skills, and motivations.

3. Carbonell, J. R., Elkind, J. I, & Nickerson, R. S. On the psychological importance of time in a time sharing system. *Human Factors*, 1968, *10*(2), 135-142.

This paper is concerned with the problems of access to the computer utility, response time and its effect upon conversational use of the computer, and the effects of load on the system.

4. Caruso, D. E. Tutorial programs for operation of on-line retrieval systems. *Jour. Chem. Documentation*, 1970, *10*(2), 98-105.

This article summarizes the literature on interactive computerized search systems which make provision for the inexperienced user. Novice-user interaction with these systems indicates that self-teaching systems can produce a competent user population which does not require an intermediary, human or mechanical, to create search strategies.

5. Cordaro, J. T., Jr., & Chien, R. T. *Design Considerations of On-Line Document Retrieval Systems.* Urbana, Ill.: University of Illinois, Coordinated Science Laboratory, R-456, 1970.

A mathematical analysis of response time in on-line systems for both sequential and inverted file organizations.

6. Hall, J. L., Negus, A. E., & Dancy, D. J. On-line information retrieval: A method of query formulation using a video terminal. *Program*, 1972, 6(3), 175-186.

This article describes a method of query formulation in which a matrix displayed on a CRT terminal is arranged in the form of a logical expression into which the searcher types the terms that he wishes to incorporate in his query.

7. Hansen, W. J. User engineering principles for interactive systems. *AFIPS Conf. Proc. Fall Joint Computer Conf.*, 1971, *39*, 523-532.

This paper presents a thorough discussion of engineering principles based on human factors considerations for designing interactive computer systems. In addition to the importance of tailoring the system to the user, protection against human errors and assistance in using the system are emphasized.

8. Higgins, L. D., & Smith, F. J. *The Cost and Response of an On-Line Reference Retrieval System.* Belfast: Queen's University, Computer Science Department, 1971.

This report on the QUOBIRD experimental reference retrieval system at Queen's University of Belfast discusses the emphasis that has been placed on procedures for minimizing the number of disk accesses required to complete a search, thus optimizing response time and reducing costs.

9. Katter, R. V. *On the On-Line User of Remote-Access Citation Retrieval Services* (TM-(L)-4494/000/00). Santa Monica, Cal.: System Development Corporation, 1970.

This paper considers the problem of identifying those characteristics of potential users that may prove important for successfully introducing and developing on-line bibliographic searching facilities. One aim of this paper is to create the beginnings of a check list useful for making sure that significant acceptance factors have been considered and taken account of in introducing on-line network services at various sites.

10. Katter, R. V., & Blankenship, D. A. *On-Line Interfaces for Document Information Systems; Considerations for the Biomedical Communications Network* (TM-(L)-4320). Santa Monica, Cal.: System Development Corporation, 3 June 1969.

This report is based on a state-of-the-art review of the literature concerned with man-machine interfaces for on-line file-searching activity for information retrieval. The function of the review is to help clarify and conceptualize the nature of on-line file-searching interfaces, to identify related considerations and problems, and to catalog types of provisions and features aimed at solving such problems.

11. Lancaster, F. W., & Fayen, E. G. *Information Retrieval On-Line* (a Wiley-Becker & Hayes Series book). Los Angeles: Melville Publishing Co., 1973.

This book provides a broad survey of the characteristics, capabilities, and limitations of then current information retrieval systems operated in an on-line interactive mode.

12. Licklider, J. C. R. *Libraries of the Future*. Cambridge, Mass.: MIT Press, 1965.

This book reports the findings and conclusions resulting from a study on libraries of the future conducted by Bolt Beranek and Newman Inc. for the Council on Library Resources, Inc. The first part of this book discusses various aspects of man's interaction with recorded knowledge. The second part explores the use of computers in library and procognitive functions.

13. Licklider, J. C. R. Man-computer communication. In C. A. Cuadra (Ed.), *Annual Review of Information Science and Technology* (Vol. 3). Chicago: Encyclopaedia Britannica, Inc., 1968. Pp. 201-240.

This chapter presents a review of man-computer communication and focuses sharply on on-line interaction. The major headings covered are system facilities for man-computer interaction; program packages and services; computer graphics; memory, storage, file, and data management; man-computer interaction languages; and sociotechnical issues and trends.

14. Marcus, R. S., Benenfeld, A. T., & Kugel, P. The user interface for the Intrex retrieval system. In Walker, D. E. (Ed.), *Interactive Bibliographic Search: The User/Computer Interface*. Montvale, N.J.: AFIPS Press, 1971. Pp. 159-201.

The results of an evaluation of the MIT Intrex retrieval system, including a description of the problems encountered by users and suggestions for solving or ameliorating these problems.

15. Martin, J. *Design of Man-Computer Dialogues*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1973.

A book for the designer that lays out steps that analysts should follow when putting together new man-computer interfaces. The designer should begin with a consideration of the characteristics of the task and the humans involved, only later moving to the terminal and computer requirements.

16. Martin, T. H. The user interface in interactive systems. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 8). Washington, D.C.: American Society for Information Science, 1973. Pp. 203-219.

This review describes man-computer interfaces and discusses scientists' use of the theory-literature interface to help them advance knowledge.

The need is emphasized, not only for adequate instruction of the user, but also for an interface designed in such a way that the user can learn easily in the process of use.

17. Mesel, E., & Wirtschafter, D. D. *On-Line Medicaid Billing System for Physicians' Services.* University of Alabama at Birmingham, Clinical Information Systems Group, Contract HSM 110-71-252, Health Services and Mental Health Administration, August 1974.

    A description of the design, development, and implementation of an on-line, computer-based Medicaid billing system for physicians' services in the State of Alabama.

18. Miller, R. B. Response time in man-computer conversational transactions. *AFIPS Conf. Proc. Fall Joint Computer Conf.*, 1968, *33*(1), 267-277.

    A discussion of the psychological aspects of response time in man-machine interaction. Different human purposes and actions at the on-line terminal are identified as having different acceptable or useful response times.

19. Moghdam, D. User training for on-line information retrieval systems, *Jour. Amer. Soc. Information Sci.*, 1973, *26*(3), 184-188.

    This article discusses various training media for training transient users of on-line information retrieval systems, including printed instructions, live help, audiovisual and on-line instruction, and computer-assisted instruction. The author concludes that an on-line tutorial program, built as an integral part of the information retrieval system, is the ideal alternative or back-up to live help.

20. Parker, E. Behavioral research in the development of a computer-based information system. In C. E. Nelson & D. K. Pollock (Eds.), *Communication Among Scientists and Engineers.* Lexington, Mass: Heath, 1970. Pp. 281-293.

    A description of a prototype system developed at Stanford University named SPIRES (Stanford Public Information Retrieval System) originally restricted to the subject of physics but now being expanded to other subject areas.

21. Rosenberg, V. Technique for monitoring user behavior at the computer terminal interface. *Jour. Amer. Soc. Information Sci.*, 1973, *24*(1), 71.

    In this technique the user-system dialogue is monitored and later printed out. At the same time, the searcher uses a tape recorder to articulate his thoughts as he interacts with the system. The taped account of the search is then compared with the printout in order to recognize difficulties and problems.

22. Sass, M. A., & Wilkinson, W. D. (Eds.). *Computer Augmentation of Human Reasoning*. Washington, D.C.: Spartan Books, Inc., 1965.

This book contains a series of papers presented in 1964 at a symposium on computer augmentation of human reasoning sponsored by the Office of Naval Research and the Bunker-Ramo Corporation. The main emphasis was on reasoning and problem solving, methods of approach, and implementation. A special feature of this book is a glimpse into the future through the medium of a dialogue on "Potential Implementation."

23. Stifle, J. *A Plasma Display Terminal*. Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, 1971.

A description of the design objectives and characteristics of the plasma display terminal developed at the University of Illinois for the PLATO system of computer-aided instruction.

24. Thompson, D. A. Interface design for an interactive information retrieval system: A literature survey and a research system description. *Jour. Amer. Soc. Information Sci.*, 1971, *22*(6), 361-373.

This article focuses on the human interaction characteristics of an information retrieval system, suggests some design considerations to improve man-machine cooperation, and describes a research system at Stanford University that is exploring some of these techniques.

25. Van Dam, A., & Rice, D. C. On-line text editing: A survey. *Computing Surveys*, 1971, *3*(3), 93-114.

A review of on-line text editing procedures for restructuring and rearranging text as well as for correcting errors.

26. Wiederhold, G. *Data Base Structures and Schemas*. New York: McGraw-Hill, in preparation.

A textbook presenting the methods, the choices, and the principles and concepts that are focal to data base organization and management.

27. Wolfe, T. Suggestions for exploiting the potential of on-line remote access information retrieval and display systems. *Jour. Amer. Soc. Information Sci.*, 1971, *22*(3), 149-152.

This article builds the case for the eventuality of the ideal on-line information retrieval system in which statistical inference analysis and heuristic optimization programs would be offered as aids to query formulation.

GLOSSARY OF KEY TERMS

Acoustic Coupler: A modulation/demodulation device that converts the digital signals generated by computer equipment into an analog form compatible with telephone lines for data transmission, or converts the analog signals received in a data transmission back into digital form using the telephone receiver. Since the data to be transmitted over the communication lines are in the form of audible tones, an acoustic coupler may be used. SEE ALSO: *Modem*.

Baud: A unit of signaling speed derived from the duration of the shortest code element. Speed in bauds is the number of code elements per second.

CAI: Computer-Assisted Instruction. A concept that applies computers and specialized input-output display terminals directly to individualized student instruction.

Controllable Cursor: An electronically generated bright signal spot on the screen of a CRT which can be positioned anywhere on the display by the CRT operator in order to select from a list of choices, options, or possible actions to be taken. Some CRT terminals use a vertical and horizontal spacer bar to position the cursor, while others use a joy stick. SEE ALSO: *Light Pen*.

CRT: Cathode-Ray Tube. An electronic vacuum tube with a screen for visual display of output information in graphical or alphanumeric form.

Data Set SEE *Modem*

Error Message: A message generated by a computer system to advise the terminal operator that he has made an error in the man-machine interaction. The message may be diagnostic, that is, it may tell the operator the exact nature of his error and how to correct it.

Interface: The common boundary between an information system and its users.

Joy Stick: A joy stick uses a lever in a control box to position the cursor at the desired location on the CRT screen. The advantage of the joy stick is that the cursor spot can be moved to the desired location directly rather than moving up or down the desired number of lines and over the desired number of spaces. A joy stick is particularly suited for use with graphics terminals.

Light Pen: A small photocell or photomultiplier in a pen-shaped housing connected by cable to a CRT console. By touching the pen to the face of the CRT and activating the trigger, the operator may make selections or initiate actions. SEE ALSO: *Controllable Cursor*.

Modem: A modulation/demodulation device that converts the digital signals generated by computer equipment into an analog form compatible with communications facilities (e.g., telephone lines) for data transmission,

or converts the analog signals received in a data transmission back into digital form. Modems are referred to as *data sets* by the telephone company. SEE ALSO: *Acoustic Coupler*.

Response Time: Usually assumed to be the time elapsed from entering a command to a computer system until its completion, the latter being characterized by the production of an output or other signal to the user, and the transfer of control to him.

Teletypewriter: A keyboard printing unit that is used to enter information into a computer and to accept output from a computer.

SECTION 7. SYSTEM DESIGN AND IMPLEMENTATION REQUIREMENTS FOR AN ATTITUDINAL RESEARCH INFORMATION SYSTEM

Introduction

The answer to the question "Is it technically feasible to implement a computer-based attitudinal research information system?" is an unqualified yes. This section delineates the system design and implementation requirements for an attitudinal RIS. Aspects considered in this section are information indexing requirements, file organization and estimated volume, file storage requirements, communication requirements, projected user load, processing required for disk and computational references, system and programming support facilities required, and operating system facilities required. System specifications are provided, and three possible computer hardware and software configurations fulfilling these requirements are enumerated. The manpower required to operate and maintain the system also is projected.

Lancaster and Fayen (1973, pp. 238 & 413), in reviewing the historical development of computer-based information storage and retrieval systems, have noted that systems implemented in the late 1950's and 1960's were largely more mechanized, more powerful, more efficient versions of searching processes that could have been implemented manually. A decade ago there was a strong tendency to mechanize existing systems without giving too much consideration to the possibility of incorporating new approaches to indexing, searching, and vocabulary control. Now that interactive on-line systems have come of age, there has been a tendency to take an existing batch-processing system and merely provide access to the file(s) by means of on-line terminals. New design approaches have tended to be neglected and the true interactive, heuristic, exploratory, and instructional capabilities of on-line systems in most cases have not been fully exploited. The designer and developer of an on-line information system should have freedom to incorporate new design concepts and new operating philosophies. An organization moving from an existing batch system may be operating under a handicap inasmuch as the natural tendency is to disturb existing procedures as little as possible. The organization without any existing batch system perhaps is in a happier position, since it is less likely to have preconceived notions or to be philosophically committed to particular processing methods.

In this section it is assumed that the designer and developer of an attitudinal RIS would not be constrained by having to adhere to old and inadequate methods of file organization and data base structure. Rather the premise underlying the system specifications is that the design and implementation of an attitudinal RIS should incorporate all of the desirable features that current theory and technology have to offer.

Such a system would not come into being overnight. Its implementation would take place over the course of several years, with the system growing asymptotically in the process. The major effort in the beginning would be to establish the data base design and organization of the various files, to prepare important data bases for data entry, and to enter these data bases into the system. However, by the fifth year it is expected that the system would have approached its maximum size and that removal of obsolete data bases

would equal the addition of new data bases. If this latter assumption proves to be incorrect, then the system holdings would continue to grow but at a much slower rate than in its early years.

Using an asymptotic growth curve, projections of anticipated file volume and use have been made for Years 1, 2, 3, and 5 in order to arrive at estimates of the computer hardware and software required and the manpower needed to operate and maintain the system. These estimates may prove to be in error as actual experience is gained in implementing the system. However, it is important that the estimates have been made and recorded. A revision of the projections can only be made if the original assumptions have been made explicit.

Information Indexing Requirements

A special condition of the Navy personnel research community is that typically the generators or collectors of research information are also the users of the information. This condition leads to a homogeneity of the universe of information that would have to be dealt with. The subject matter and concepts expressed in the information is in coincidence with the professional terminology and discourse of potential users of the information. The property of homogeneity of the total environment minimizes the indexing requirements for an attitudinal RIS since the subject area encompassed by the array of data bases that would be included in the system is relatively narrow. The expectation is that the indexing vocabulary required to control the storage and retrieval of data elements would be small in comparison to, for example, the indexing vocabulary of large bibliographic retrieval systems such as MEDLARS, DDC, or NASA's RECON.

From the smallness of the indexing vocabulary required, a number of additional benefits result. Thesaurus construction would be simplified. The indexing terms to be included in the thesaurus can be chosen from the array of data bases themselves. Thus, the indexing vocabulary would be as specific as the data bases that it describes, resulting in high precision. Because many of the data fields are already succinctly named or can be described using the survey question content, there would be no difficulty in optimizing exhaustivity of indexing. The need to show a multitude of term relationships would be minimized since related terms would tend to be displayed in close proximity to each other due to the smallness of the indexing vocabulary size. Because of the homogeneity of the information universe and the user community, the possibility of many synonyms for a concept is reduced; consequently, the entry vocabulary would be small.

The features of smallness of the indexing vocabulary, the synonym tables, and the entry vocabulary lead to a thesaurus that would not be expensive to construct. It would be relatively inexpensive in time and effort to use and to update, and indexers would not have to be highly skilled personnel. Aids to information searchers at the on-line terminal, including entry vocabularies, synonym tables, and term hierarchies, typically are very cost effective. Although there may be some expense involved in their initial construction and maintenance, they are likely to repay their cost many times over by saving subsequent time expended in preparing search statements.

The objective in formulating an indexing vocabulary for Navy personnel survey data should be to provide a means for exhaustive retrieval of all data pertinent to a search request. At the same time, the vocabulary should be easy to apply both at the time of indexing and searching. It is expected that the indexing burden would be manageable and well within the resources available to allocate to this task. The main portion of the indexing burden would occur in the first year or two of system implementation as data bases are prepared for entry into the system. After this initial expenditure of effort, the addition of new data bases would proceed at a much reduced flow, requiring minimal indexing and file maintenance.

The search performance of the system is expected to be good. Recall should be high since system users would be well versed in the subject area covered by the data base holdings. Thus, they should be able to formulate precise search statements to meet their information needs. There is no need for the use of sophisticated precision devices such as role indicators and links since boolean search expressions would serve the same purpose. Furthermore, the additional burden of assigning role indicators or links would be averted in the indexing process.

For all of the reasons discussed above, the recommended choice of an indexing strategy is the controlled vocabulary approach. The information content of anticipated data base holdings does not lend itself particularly well to a natural language or automatic indexing approach. Since the indexing burden is expected to be minimal, there should be no indexing backlog once the initial holdings are entered into the system. The requirement to enter new data bases into the system as quickly as possible, which would help justify an automatic indexing approach, does not appear to present a problem. Therefore, the advantage of a more beneficial search capability resulting from the controlled vocabulary approach opts in favor of this indexing alternative.

The development of a controlled indexing vocabulary for a manual archive of existing and future Navy personnel survey data will be undertaken during 1975-1976 by R-K Research and System Design under contract to the Navy Personnel Research and Development Center in San Diego. This work will lay the foundation for an indexing methodology that will apply to a manual data archive as well as to a possible future implementation of a computer-based attitudinal RIS.

File Types and Organization Required

The analytical processes to be supported by an attitudinal RIS are based on the following model of research activity:

- The researcher conceptualizes the problem and the data element types relevant to conjectures of relationship.
- He uses the RIS index to search a general schema for available data element types that match his information needs.

- He evaluates the selected data element types for their suitability, using such information as their source, dates of coverage, population characteristics and size, range of values, reliability, and completeness.

- He tabulates some sample data to verify that his selection is appropriate. Tabulation could include the production of totals, counts, averages, and standard deviations of variables.

- He checks out the analysis procedure that he has chosen to use on the sample data set.

- He accesses the source holdings in order to process all of the relevant data, using the appropriate analysis procedure.

The files needed to support the research functions enumerated above will be summarized, and then their attributes will be described from the point of view of the specifications required for the design and evaluation of an attitudinal RIS. In these considerations, the design will be oriented toward achieving an upper size for the system holdings in about five years. Up to this time, growth is expected to be asymptotic; beyond that point, removal of obsolete data bases is expected to equal the addition of new data bases. If this assumption proves to be incorrect, then the system holdings would continue to grow but at a much slower rate.

The data components for the attitudinal RIS are the following categories:

Catalog

Schema

Access Processes

Samples

Programs

Files of Source Holdings

A definition of each of these categories is provided below.

Catalog. A *catalog* of data base holdings would describe the source files that are available and have potential use in Navy personnel research.

Attributes of the files to be described in this catalog would include the following:

Name or names: as commonly used by personnel researchers

Reference name: chosen to be succinct and unique

Source: of the data gathered, including the size and type of population

Dates: of the data collection period

Code or individual responsible: for the maintenance of the data

Size: of the data collection, typically both the number of records and the size of the records

Location: of data collection (e.g., NPRDC, NHRC, NEL, BuPers, MARDAC,...)

Type and computer format: of the data (i.e., magnetic tape or punched cards, printing format, etc.)

File name: for computer-based data files, the name encoded in the file header

Access process: for the data

Sample data name: the name given to a small representative subset of the file contents

List of data types contained within the file records (file schema) with (1) encoding methods for individual data elements, (2) original question if the data element is a questionnaire response, and (3) as described below in the description of the schema

Schema. A *schema* describes the variables contained in all of the source holdings. This list would be prepared from the catalog to the files of source holdings and would provide the primary entry point for researchers wishing to verify the existence of relevant data. It would bring together similar data elements found in distinct holdings. It might be indexed to reflect conceptual relations between distinct data element types.

Attributes for data element descriptions stored in this file would include the following:

Name: of the data element type

Occurrence number: so that identical data items in different data files could be managed

Description of variable: background to provide information for selection

Reference to the source file: a name listed in the catalog

Reference to the access process: for the file of source holdings specific for this variable (i.e., record selection)

Position and size: place and extent for this data element type in the file of source holdings

Encoding method used: characters, multiple-choice response possibilities, numeric range

Population size: that this question was asked of or that the data file is based on

Percentage: of population that answered this question

Estimate: if any, of the reliability and psychometric characteristics of the data elements

Only short, coded information would actually be kept in this inverted form. Extensive data would be obtained by reference to the description in the catalog for each file of source holdings.

Access Processes. Since it is beyond the scope of an attitudinal RIS to attempt to standardize all of the data base holdings of interest to Navy personnel researchers, a large variety of access mechanisms might be required to carry out data fetching for analysis. Both routine and specific manual operations might be required to make data accessible for actual processing. Procedural descriptions might be required frequently. Even if data are located at the processing center, data abstraction programs might be required. This interface to the data also could provide data privacy and security to the extent required.

Contents of the files containing access processes would include the following:

Name: of the access process

Listing: of manual actions required

Job control cards: with condition-specifying prefixes for selection and modification

Programs: to carry out required selections and modifications based on interaction with the system users

Samples. So that adequate information would be available during the user's selection process, samples of the data files should be made available without going through the actual selection process. These sample sets would be created once during the process of cataloging the source holdings.

Use of these sample data also would permit the debugging of analysis programs to be carried out in a rapid and efficacious manner.

The number of samples to be selected might amount to approximately five percent of the number of records in the source files, subject to a lower limit of, say, 10 samples and an upper limit of, say, 50 samples. Random selection of records would be the simplest selection method, although a representative sample might be more informative. Human post-editing of randomly selected samples might add considerably to the information value of the sample; however, some bias could be introduced by the editing process.

The creation of sample files would be a by-product of the verification of the access procedures developed. The sample files themselves would be accessed by a formal simplication of the access processes described above. The format of the records would be identical for the sample and the source holding, but the fetching of the records could be done directly from local storage rather than remotely.

Programs. Routine utility and statistical programs also would be maintained within the attitudinal RIS. These programs would be in a form that is applicable to the needs of Navy personnel researchers and the access facili-

ties provided. Less research set-up time would be needed, and the step from sample analysis to the full data analysis would be simplified and less subject to errors.

The access procedures discussed above would transform source holdings into a standardized format. This format should be suitable for processing with minimal further transformation by the BMD, SPSS, or OSIRIS programs maintained in the program library. The standardized files could be catalogued as secondary holdings if further use were anticipated. The program file would be upgraded over time through the collection of analysis procedures used in conducting personnel research.

Files of Source Holdings. Last but not least are the source data for research analysis processes. These master data holdings would be maintained as much as feasible in their existing form and would constitute a passive component of the research information system. Abstract files could be produced periodically from certain very large files to make access more economical or more convenient. This practice would result in multiple entries for variables in the schema files. The abstract files would be kept in standardized form to simplify access.

Estimated Volume of Files

In order to assess the magnitude of the hardware and software required to implement an attitudinal RIS of the type envisioned, it is necessary to make some initial qualitative assumptions. These assumptions should be successively refined as the design process takes place. The setting out of the initial estimated values in a tabular form also provides a basis for discussion between potential system users and system implementers.

Catalog.

Number of holdings: 1,000 files

Growth curve:
Year 1 - 100 Year 2 - 250 Year 3 - 500 Year 5 - 1,000

Average number of variables per holding: 100

Space for description of holding: 1,000 characters (1 page)

Space for description of each variable: 100 characters

Schema.

Number of variables: 100,000 (product of number of files times variables per file)

The growth curve would be proportional to linear catalog growth.

Year 1 - 10,000 Year 2 - 25,000 Year 3 - 50,000 Year 5 - 100,000

Access Processes.

Number of distinct access processes: 100

Description of access process (ea. 100 lines): 10,000 characters

The growth curve would be linear due to combined logarithmic and exponential effects. Initially, growth would consist of many simple processes, extending eventually to fewer, complex ones.

Year 1 - 25 Year 2 - 50 Year 3 - 75 Year 5 - 100

Samples.

For 1,000 holdings, estimate an average of 25 records of 100 variables, each of four characters in size. This estimate leads to a total size of 10,000,000 characters.

The growth curve again would be proportional to linear catalog growth.

Year 1 - 1,000,000 Year 2 - 2,500,000 Year 3 - 5,000,000 Year 5 - 10,000,000

Programs. The current program library index at NPRDC indicates that there are about 450 utility and statistical programs on file at this institution. The total number of programs required will be based on this count. It is improbable that the full number would ever be actually used, since programs from separate statistical packages may duplicate each other, and also since certain procedures (e.g., BMD survival statistics) may not be applicable to the kinds of analyses performed by personnel researchers. On the other hand, it can be expected that with increased data processing activity, more new programs would be written and obtained.

An estimate is that the utility and statistical programs average 600 80-column cards each, thus eventually requiring 21,600,000 characters of storage. The growth curve would be sharp because programs would be acquired at a rapid rate initially.

Year 1 - 250 Year 2 - 350 Year 3 - 450 Year 5 - 450

These projections yield an eventual file size for the attitudinal RIS of the following number of characters:

Catalog	11,000,000
Schema	10,000,000
Access Processes	1,000,000
Samples	10,000,000
Programs	21,600,000
TOTAL	53,600,000

Taking the expected growth curves into account, the following projections are possible:

	Year 1	Year 2	Year 3	Year 5
Catalog	1,100,000 char.	2,750,000 char.	5,500,000 char.	11,000,000 char.
Schema	1,000,000 char.	2,500,000 char.	5,000,000 char.	10,000,000 char.
Access Processes	250,000 char.	500,000 char.	750,000 char.	1,000,000 char.
Samples	1,000,000 char.	2,500,000 char.	5,000,000 char.	10,000,000 char.
Programs	12,000,000 char.	16,800,000 char.	21,600,000 char.	21,600,000 char.
TOTALS	15,350,000 char.	25,050,000 char.	37,850,000 char.	53,600,000 char.

File Storage Requirements

Main Files. The projected file size indicates that the required data up until the final projected volume could be kept on one double-density 20-surface disk pack (50,000,000 characters); these disk packs are commonly available. This amount of file storage is equivalent in capacity to two standard single-density IBM 2314 units. If a considerable number of indexes and dictionaries are to be kept in the system, then additional file storage would be required. The quantity of additional storage would be a function of the access design, but it is doubtful that a careful design would add more than 25 percent to the total file storage. In this case, two double-density or three single-density drives might be required.

Drives for these units can be mounted on a large variety of computers, from medium-size minicomputers to the largest ones made. It should be realized that the full file size requirements would not be reached until the system has been in regular use for some time. Consequently, single-density units or fewer double-density units may be considered initially.

Auxiliary Files. To provide for loading, back-up, and maintenance, adequate tape capability for auxiliary files would be required. If the system being considered uses only the one disk required for the main file, then two tape drives might be desired for operational viability.

Communication Requirements

It is foreseen that after suitable data have been identified, indexed, and catalogued, both test and production jobs would have to be run. For this function, communication facilities would have to exist between the catalog system and the execution system. Communication would be easy if both processes reside in the same machine. If they do not, then certain bandwidth needs for communication would be required (i.e., if the computer executing the jobs were to be located remotely).

It is estimated that a typical test job would consist of the following:

100 job control cards × 80 columns	8,000 char.
600 statistical program cards × 80 columns	48,000 char.
25 × 100 × 4 characters of data for testing (sample data subset)	10,000 char.
TOTAL	66,000 char.

A production job would not include the test data since it instead would access the main source holdings. However, the access procedures involved would be larger.

It is desirable that the submission of a job should not take more than a couple of minutes at most. If 66,000 characters have to be transmitted in two minutes, then the required transmission data rate is

$$66{,}000 \div (2 \times 60) = 550 \text{ chars./sec.}$$

Using an asynchronous transmission protocol, 10 bits per character and 30 percent overhead for line turnaround are needed, so that the required communication transmission rate would be

$$550 \times 10 \times 1.30 = 7{,}150 \text{ baud.}$$

This requirement could be satisfied by 7,200-baud or 9,600-baud transmission lines. The 9,600-baud speed is more commonly available. Both rates require leased-line connections with telephone-company provided conditioning, if the distance exceeds a few miles.

The use of slower speed lines, as available through dial-up service, would provide only 1,200 baud, and the submission time for a job would be

$$66{,}000 \div (1{,}200 \div 10) \times 1.30 = 715 \text{ seconds or } 12 \text{ minutes.}$$

Regular use of such slow transmission implies the use of background processing in the system, which might add to the aggregate system cost.

If the attitudinal RIS were to reside in the same computer that carries out execution of the programs and has the source holdings available, then the communication process would be simplified for the submittal of a job to the execution partition. On IBM systems this function can be carried out through the HASP input/output monitor or similar facilities. Most other large computers also have techniques for background job submittal from an interactive partition. Experience has shown that such a submittal can proceed at about 10,000 characters per second. In this case, job submittal would take

$$66{,}000 \div 10{,}000 = 6.6 \text{ seconds.}$$

Projected User Load

User demand is, of course, the most difficult factor to project, since the system would provide an interface with the information store that would be new to Navy personnel researchers. However, the following assumptions can be made from observations in other research environments:

- A researcher cycles through processes of problem formulation, hypothesis definition, model building, data collection, data analysis, and result presentation. When using secondary data, the data collection phase---normally the longest step---is drastically reduced.

- Data preparation and data analysis would be facilitated by the use of an attitudinal RIS.

- A user session at the terminal on the average would be less than one hour. An active user would want about four sessions per week during which he would be actively exploring the data. However, these active periods would occupy only one-quarter of his time during any particular year. The rest of the time he would be engaged in other aspects of the research process, such as problem formulation, experimental design, and data collection.

Assuming that the attitudinal RIS were to be housed at NPRDC, there are about 75 full-time professionals at this institution who would be using the system. There also are about 30 half-time student research assistants who would be expected to use the system two hours per week each. If it is postulated that 50 out of the 75 professionals might actively use the system, then a total user population of $50 + 0.5 \times 30 = 65$ equivalent full-time researchers can be estimated as being active users of the system. The required availability for these users then is

$$1 \text{ hour} \times 4 \times 1/4 \times 65 = 65 \text{ hours per week.}$$

If queuing is to be minimized, a terminal can only be productively used up to 30 hours per week. Training and data exploration should be carried out on separate terminals to avoid conflicts of demand for terminal usage for activities perceived to be of different values.

Time also would be required for file maintenance and updating of the data contained in the source holdings, program editing, scheduling, and preparation of usage statistics. A terminal dedicated to these functions could be used more extensively since these operations can be scheduled.

Thus, it is possible to project that initially a single user station would be adequate, to be augmented by a second and third station when user demand warrants. One, possibly simpler, station could be dedicated to training. One station, located conveniently to the data maintenance group, could carry out the update and maintenance function.

The operations carried out by the researcher-users would be predominantly READ operations. Some writing would take place when a job had to be assembled prior to submission. The operations carried out by the system maintenance-users would involve considerably more writing. It should not be necessary to inhibit researcher-users from reading files while they are being updated, since few problems could be caused during exploratory data analysis. During job assembly, however, the files should be locked to prevent concurrent update. This requirement implies support in the computer system of a multiple-reader/one-writer file interlock system.

Processing Required

Disk References. A majority of the processing capacity would be used up by disk references. A sequential search through a schema of 5,000,000 characters in size would require a double-density 2314 ISS or AMPEX disk given one revolution and one seek per track. Therefore,

$$5{,}000{,}000 \div 7{,}000 \times (37 \text{ msec per revolution} + 25 \text{ msec per seek})$$

$$= 44{,}286 \text{ milliseconds or } 44 \text{ seconds.}$$

This response time is beyond the range adequate for a good man-machine interface, and a more sophisticated access method would have to be used. For most systems, indexed or direct access methods are available based on key values. Since a variety of criteria can be considered for access to the schema, one system requirement would be that file organization must provide for multiple indexed or multiple direct access to a file.

Computational References. Between 1,000 and 10,000 processing cycles typically are required to process one record. On minicomputers one cycle can do less, so that more cycles are required to carry out a task. On large computers sophisticated access routines tend to use more instructions, even though each instruction can do more. The following conservative estimate is based on 10,000 cycles per record fetch. On the type of disk system considered under *File Storage Requirements*, a record can be fetched in a maximum time of 37 + 25 milliseconds or a minimal time of about 10 milliseconds. Since it is desirable that computation be completed well within the disk fetch time, the processing cycle speed required would be

$$10 \text{ milliseconds} \div 10{,}000 = 1 \text{ microsecond.}$$

This estimate is a reasonably high, but available speed for both large and small processors.

System and Programming Support Facilities Required

In order to provide the data manipulation required, it would be desirable that a higher level language, supporting file activities and string processing, be available. It is envisioned that only minor numeric computations would be needed, and any programmable language system would provide these.

Adequate languages for these services would be COBOL or subsets thereof, or a system supporting the BASIC language with adequate file extensions. A FORTRAN-only facility is likely to be inadequate. Some pre-programmed data management and retrieval systems exist which could carry out many of the tasks indicated. A need for the linkage to procedural subroutines would be required to carry out minor, but essential analyses in the area of data exploration.

The operations which would have to be carried out in the analysis area with the support of the system are the following:

1. Search through a dictionary to select appropriate variables.

2. Search through an ordered schema to locate appropriate data.

3. Determine location of the source file and status for these data.

4. Retrieve and tabulate a sample of selected data elements to assess the value of these data.

5. Assemble a job to be submitted for processing.

The operations which would have to be carried out to support the maintenance functions include the following:

1. The creation of files from card or tape input.

2. The updating of files from terminal and card or tape input. Updating would require extension of files, insertion of records into files, replacement of records in files, and the changing of fields within files.

3. The selective listing of file contents, both by key and by data field contents.

Since the files may be largely in character or textual format, a powerful text-edit facility might satisfy the maintenance requirements.

The means to achieve these operations can vary considerably from system to system. It would prejudice the system selection process to specify in detail *how* these operations are to be performed. The choice would be between (1) a higher level language programming system with file capability and strong manipulation capability, or (2) a data base management system with program insertion capability. In either case, an edit capability for the files would be required.

Operating System Facilities Required

In order to support the operation envisioned, the following categories of operating system services would be required:

1. Multiple-user access to the system.

2. Individual processes for each user extending over complex search transactions. This requirement implies conversational processing rather than simple transaction processing. Timesharing is the most frequently used means to achieve this goal.

3. Simultaneous access to files using a multiple-reader/one-writer access constraint to avoid deadlock.

4. Higher level language facilities supporting all of these functions.

Interaction at the rate foreseen would require CRT terminal interaction. It also would be necessary to provide paper (hard-copy) output in order that the system user could take reference material from the terminal site to his office. If the attitudinal RIS terminal(s) were to be remote from the execution computer, then the station(s) also would function as a remote batch output station. This configuration would require the addition of a remote moderate-speed printer and possibly a plotting device. To allow the necessary operational flexibility, both printing and plotting output should be spooled to disk at the execution computer and be produced on demand. It might be desirable to allow the researcher to sample print and plot output before incurring the expense of volume printing and plotting.

Copier-like hard-copy devices are now available which permit the reproduction of a CRT graphic image on paper, and one of these devices would be desirable. The number of terminals required would depend on whether or not the user community was located in close physical proximity. A more spread-out operation would necessitate additional devices. The following complement of terminals is suggested for the case where the attitudinal RIS were to be located at NPRDC.

3 interactive text CRT's

2 interactive text and graphics CRT's

1 printing/plotting attachment for same

2 high-speed printing terminals (120 chars./sec.)

For data input and control:

1 card reader

1 console terminal

And if remote operation is required, add:

1 700-line-per-minute printer

1 graphics plotter

1 card punch

System Specifications

Summarizing all of the requirements delineated above, the following system specifications emerge:

- A processor with a cycle speed of one microsecond or better.
- A disk system with a capability to mount more than 60 million characters and a standard access speed of 37 msec per revolution and 25 msec per seek.
- A file system allowing key-based random access to files.
- A file system supporting at least multiple record sizes; if possible, variable sizes.
- Communication capability to drive at least five terminals at up to 9,600 baud.
- A spooling facility for remote output operation.
- An operating system with multiple-user access to shared files in a timesharing or equivalent mode, allowing at least one user write privileges.
- A programming language to support these facilities, or a data base management system that provides the services needed.
- A text-editing system for the maintenance of the files.

Possible Systems Fulfilling These Requirements

In order to obtain a feeling for the hardware and software costs of an attitudinal RIS, three sample configurations will be sketched. The prices listed are rounded figures taken from recent quotations. They do not represent any commitments from a manufacturer. It is advised to go out on competitive bid and evaluate the offerings carefully. Prices and facilities change rapidly in this field, and within a few months of this writing, new choices may be available.

Configuration 1: Intelligent Terminal System (e.g., Four Phase Model IV/70 Intelligent Terminal System)

6 text CRT's @ $1,500	$ 9,000
48K processor	30,000
1 Centronix line printer (300 lpm)	4,000
2 double-density disks @ $20,000	40,000
1 tape unit (45 ips, 800 bpi)	12,000

1 slow-speed card reader	7,200	
1 9,600-baud communications interface	2,000	
COBOL software	1,000	(estimate)
TOTAL	$105,200	

NOTE: This configuration offers no graphics capability or spooling for remote output. Data-base support software would have to be written. Monthly maintenance is estimated at .6% or $631 per month.

Configuration 2: Full, Dedicated System (e.g., Hewlett Packard 3000 System, Type 300CX)

CPU with 128K bytes	
1 magnetic tape drive (45 ips, 800 bpi)	
1 2M-byte fixed-head disk	
1 47M-byte double-density disk	
1 1,250-lpm line printer	
1 reader/punch subsystem	
	$203,500
1 synchronous communications controller	1,240
1 additional 47M-byte double-density disk	20,000
1 additional magnetic tape drive (1,600 bpi)	13,500
1 Calcomp plotter interface	1,030
1 Calcomp plotter (.01 inch increments)	3,000
3 HP 2640 display terminals @ $3,200	9,600
2 Tektronix text and graphics terminals @ $4,500	9,000
1 hard-copy unit	3,500
1 COBOL compiler	4,500
1 editor	1,000
1 data-base management package	10,000
1 query system	1,000
TOTAL	$280,870

NOTE: Monthly maintenance for this system configuration is estimated at .6% or $1,685 per month.

Configuration 3: Shared, Large Computer System

The cost estimates provided below are based on full utilization projections. It should be realized that these projections probably would not be reached until Year 5. While a shared system is apt to provide very comprehensive software support, it is not likely to be stable over a 5-year horizon due to requirements for improvements placed on the system by other users. The figures cited below are typical academic center charges.

5 × 8 terminal hours per day × 200 days @ $5		$ 40,000
CPU utilization at 1 minute per terminal hour @ $720 per hour		96,000
File storage for 60M bytes @ $0.50 per track of 15,000 bytes per month		24,000
Printing at $.01 per page, 1,000 pages per day		2,000
Rental:	3 CRT terminals @ $125 per month	4,500
	2 CRT text and graphics terminals @ $175 per month	4,200
	1 hard-copy unit @ $125 per month	1,500
	2 1,200-baud printers @ $135 per month	3,240
	TOTAL	$175,440 per year

NOTE: Maintenance, tape mounting, and other ancillary charges are included.

For all three configurations described, additional costs to be incurred would be for supplies and for telephone line costs and modems if the execution computer were located remotely from the RIS user community.

Manpower Required To Operate and Maintain the RIS

The staff required to perform each of the following tasks is considered below:

- Indexing of the data base holdings.
- Creation and maintenance of the catalog of source holdings.
- Creation and maintenance of the schema.
- Development, verification, and maintenance of access processes.
- Creation, verification, and maintenance of samples.

- Obtaining, adjusting, verifying, and maintaining of programs.
- Development and maintenance of system functions.

The following considerations apply to each of these areas.

Indexing of the Data Base Holdings. It is estimated that on the average it would require three man-days of effort to index a data base using a controlled vocabulary. This estimate may prove to be high; however, it is better to over-estimate this requirement so that adequate manpower for this task may be planned. The first year of experience will indicate if this estimate is realistic. The incremental volume of files to be created from the indexed source holdings would be as follows:

Year 1	Year 2	Year 3	Year 4	Year 5
100	150	250	250	250

Assuming an indexing rate of three man-days per data base, the following 5-year projections for man-days of indexing effort required can be made:

Year 1	Year 2	Year 3	Year 4	Year 5
300	450	750	750	750

Using 250 man-days as the standard annual number of days of work for a full-time employee, one indexer would be required initially, to be augmented by a second indexer near the end of the first year. These two indexers between them could handle the indexing load in the second year. A third indexer would be required from Year 3 on.

Creation and Maintenance of the Catalog of Source Holdings. Entry of data into the system for the catalog would require keyboarding of the catalog entries. The incremental volume of characters estimated for the catalog for each of the five years projected would be the following:

Year 1	Year 2	Year 3	Year 4	Year 5
1,100,000	1,650,000	2,750,00	2,750,000	2,750,000

Assuming a data entry rate approximately equal to that of a typical keypunch operator (80 cards per hour - all 80 columns punched), a daily output of 50,000 characters would require the following number of man-days:

Year 1	Year 2	Year 3	Year 4	Year 5
22	33	55	55	55

Creation and Maintenance of the Schema. The schema is largely an automated abstract of the catalog. It is expected that one man-day would be required per source holding to verify and cross check schema entries. This estimate for each of the five years amounts to the following number of man-days:

Year 1	Year 2	Year 3	Year 4	Year 5
100	150	250	250	250

Development, Verification, and Maintenance of Access Processes. The access processes required would be very similar for similar files on similar equipment using similar program packages. It is assumed that ten basic access processes would provide the standard features needed, and that the remainder would be derived from this basic experience. To develop an access process, one man-month or 20 man-days would be required; to adapt a process is assumed to take one day. The estimates below assume that five basic processes would be developed in each of the initial two years. The number of man-days required, then, in each of the five years would be the following:

Year 1	Year 2	Year 3	Year 4	Year 5
100 + 25	100 + 25	25	13	12

Entry of access processes would require keyboarding into the computer system. The estimate for this effort is conservatively based on the yearly volume of access processes and would require the following man-days of effort:

Year 1	Year 2	Year 3	Year 4	Year 5
5	5	5	3	2

Creation, Verification, and Maintenance of Samples. The obtaining of samples is estimated here as a by-product of verification of the access process on each new source holding. Manual selection to optimize the information value of the samples is not included in these estimates. One half of a man-day is the estimate for the test of the access process and the visual scanning of the obtained sample for validity. In each of the five years, the following number of man-days would be required:

Year 1	Year 2	Year 3	Year 4	Year 5
50	63	125	125	125

Obtaining, Adjusting, Verifying, and Maintaining of Programs. The programs to be used would be obtained largely from existing libraries. It is desirable, however, that their correct operation be verified. This verification might involve considerable effort for the initial program tested from one specific source library; however, subsequent programs from this same library should require relatively little effort. If it is assumed that programs would be obtained from ten different libraries, and that for each new library acquired one month is required for verification of the initial program, then ten man-months or 200 man-days in the initial two years would be occupied with this task. One additional man-day of effort for each subsequent member of a program library after the first would be required, so that the total estimate of man-days required for each of the five years would be the following:

Year 1	Year 2	Year 3	Year 4	Year 5
100 + 250	100 + 100	100	0	0

Development and Maintenance of System Functions. The work required for this task would precede in part work on any of the above tasks. The type of effort required would depend greatly on the capability of the system at the particular point that is reached in system implementation. The system envisioned would require the development of functions which are largely serially dependent on each other. Given the total scope of work, it is estimated that one half year to one year would be required for pre-implementation development work, given that the design phase is completed and has produced a correct operational specification.

The amount of manpower required for the pre-implementation phase would be between one and three individuals, so that the amount of effort to be expended would range from six man-months to 36 man-months.

Some keyboarding assistance also would be required during this initial phase. An approximation of support requirements places this effort at one hour per professional man-day, or 2.5 days per man-month, or 15 to 90 man-days.

Total Estimated Manpower Required. The total estimated manpower required by personnel category is as follows:

Indexing Personnel:

Year 1	300 man-days	=	15.0 man-months
Year 2	450 man-days	=	22.5 man-months
Year 3	750 man-days	=	37.5 man-months
Year 4	750 man-days	=	37.5 man-months
Year 5	750 man-days	=	37.5 man-months

RIS Personnel:

Year 0	6 to 36 man-months
Year 1	100 + 125 + 50 + 350 = 625 man-days = 31.3 man-months
Year 2	150 + 125 + 63 + 200 = 538 man-days = 26.9 man-months
Year 3	250 + 25 + 125 + 100 = 500 man-days = 25.0 man-months
Year 4	250 + 13 + 125 + 0 = 388 man-days = 19.4 man-months
Year 5	250 + 12 + 125 + 0 = 387 man-days = 19.4 man-months

Data Entry Personnel (Keyboarding):

Year 0	15 to 90 man-days = .75 to 4.5 man-months
Year 1	22 + 5 man-days = 27 man-days = 1.4 man-months
Year 2	33 + 5 man-days = 38 man-days = 1.9 man-months
Year 3	55 + 5 man-days = 60 man-days = 3.0 man-months
Year 4	55 + 3 man-days = 58 man-days = 2.9 man-months
Year 5	55 + 2 man-days = 57 man-days = 2.9 man-months

Given the above projections of manpower required to operate and maintain an attitudinal RIS, it becomes possible to attempt to conceptualize what a table of organization to man this facility might look like. Figure 29 portrays one possible way of organizing the manpower functions needed to operate and maintain an attitudinal RIS. Three major categories of functions appear in this table of organization: data base acquisition and maintenance, system operations, and user services and training. This latter category was not included in the projections to operate and maintain the system. These user support services could be performed by one to several full-time professionals, depending on the demand for their services and the economic feasibility of using professionals in this capacity. Also, if the attitudinal RIS were to be implemented on a shared, large computer system rather than on a dedicated computer system, the Manager of System Operations and the Computer Operator job positions would drop out, but many of the functions inherent in this line of responsibility would have to be provided for in different ways.

In considering Figure 29, it should be kept in mind that the job positions shown are purely hypothetical and that the job duties attributed to several positions could be combined into one billet. For example, one individual might be assigned to provide all user services and training, or the Manager of System Operations might also be responsible for systems analysis. However, to keep the various manpower functions required as distinct and clear as possible, the organization shown in Figure 29 is provided. The number of billets needed for each job position would be a function of system design, growth, and user demand. What are portrayed in Figure 29 are the types of job positions to be considered in setting up an actual table of organization for an attitudinal RIS. To assist in this consideration, the following tentative job descriptions are provided for each job position.

Manager of the Attitudinal RIS. Plans, organizes, and controls the overall activities of the attitudinal RIS, including data base acquisition and maintenance, the tape library, system operations, and user services and training. Consults with, advises, and coordinates between his division and other departments. Reports to higher management on objectives, plans, projects, performance, and other matters related to the operations of the RIS.

Manager of Data Base Acquisition and Maintenance. Plans, coordinates, and directs all activities involved in data base acquisition and maintenance, including liaison with other organizations providing magnetic tape files, data base indexing, data base preparation and cleaning, and data base maintenance. Maintains the thesaurus, with its controlled vocabulary, and the entry vocabulary. Reports to the Manager of the Attitudinal RIS.

Data Base Indexer. Indexes the contents of data base holdings preparatory to input to the RIS using the controlled vocabulary. Reports to the Manager of Data Base Acquisition and Maintenance.

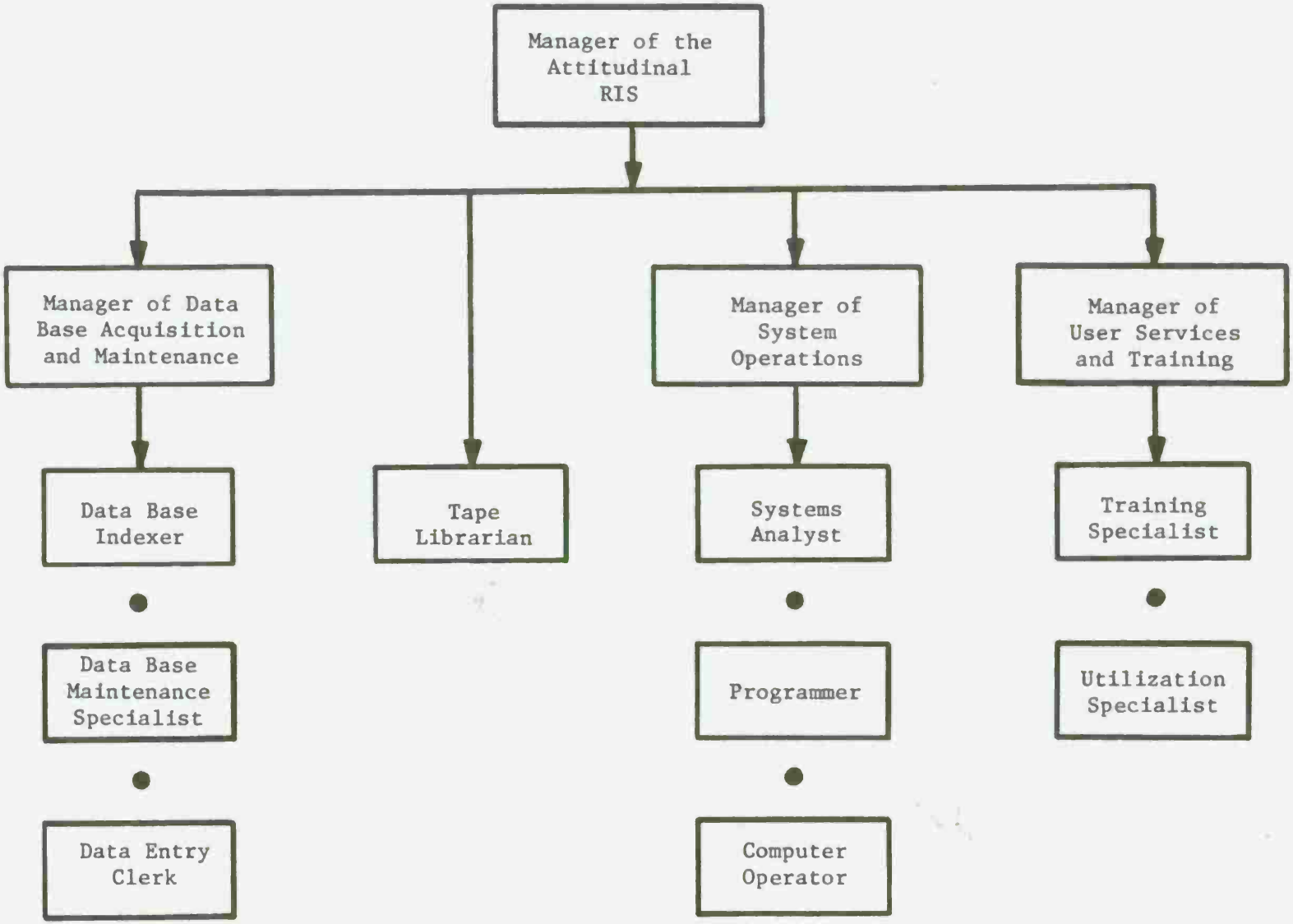


Figure 29. A Possible Table of Organization for an Attitudinal Research Information System.

Data Base Maintenance Specialist. Creates and maintains the catalog of source holdings, the data base schema, and the samples of source holdings. Reports to the Manager of Data Base Acquisition and Maintenance.

Data Entry Clerk. Keys the data into machine-processable form using a console terminal keyboard. Also provides keyboarding support for maintenance of the indexes, the catalog, the data base schema, the sample of source holdings, file access processes, and programs. Reports to the Manager of Data Base Acquisition and Maintenance.

Tape Librarian. Maintains the library of magnetic tape reels and their documentation. Classifies, catalogs, and stores reels. Maintains charge-out records. Inspects tape for wear or damage. Reports to the Manager of the Attitudinal RIS.

Manager of System Operations. Plans, coordinates, and directs all activities involved in implementing the ongoing operations of the RIS, including systems analysis, programming, and computer operation. Supervises developmental work to improve the system as well as scheduling day-to-day system operation. Responsible for conducting feasibility studies of new applications and system innovations. Prepares system usage reports for management. Reports to the Manager of the Attitudinal RIS.

Systems Analyst. Plans, develops, and maintains required file access processes and system functions, depending on the capability of the available computer hardware and software. Participates in feasibility studies of new applications and system innovations. Reports to the Manager of System Operations.

Programmer. Responsible for obtaining, adjusting, editing, verifying, and maintaining programs used by the RIS. Programs any specialized functions not available through the standard system software. Reports to the Manager of System Operations.

Computer Operator. Operates the central console and reruns job steps to recover from machine error or program error, consulting with technical staff where necessary. Maintains machine performance and utilization records. Reports to the Manager of System Operations.

Manager of User Services and Training. Plans, coordinates, and directs all activities involved in training users and assisting users in utilizing the RIS. Develops training approaches and implements training curricula including continuing education. Serves as advisor and consultant in ways to utilize the RIS to advantage. Reports to the Manager of the Attitudinal RIS.

Training Specialist. Assists in the development and implementation of training curricula and materials. Helps train new system users and provides continuing education to regular system users. Acts in the role of a user surrogate for individuals preferring not to interact directly with the RIS. Reports to the Manager of User Services and Training.

Utilization Specialist. Assists users in achieving a better utilization of the capabilities of the RIS, including consultation in experimental design, statistical methodology, and data exploration and analysis procedures available through the RIS. Reports to the Manager of User Services and Training.

The salaries appropriate for each of the job positions described above are more difficult to assess. However, some guidance in this area is provided by a recent article appearing in *Datamation* which presents the results of a nationwide survey of the salaries of data processing personnel (McLaughlin, 1975). Job descriptions are provided as well as nationwide average monthly salaries by job title. Average salaries by job title for 17 major cities in the United States also are included since there sometimes are substantial discrepancies in pay for equivalent jobs in different parts of the country. Three West Coast cities are included among the 17 listed---Los Angeles, San Francisco, and Seattle. If the attitudinal RIS were to be implemented at NPRDC, the salary structure for Los Angeles probably would provide the best guidelines for comparable salaries in San Diego.

ANNOTATED BIBLIOGRAPHY

1. Lancaster, F. W., & Fayen, E. G. *Information Retrieval On-Line* (a Wiley-Becker & Hayes Series book). Los Angeles: Melville Publishing Co., 1973.

   This book provides a broad survey of the characteristics, capabilities, and limitations of then current information retrieval systems operated in an on-line interactive mode.

2. McLaughlin, R. A. DP salary survey. *Datamation*, 1975, *21*(1), 40-46.

   A nationwide survey of the salaries of data processing personnel including over 110,000 employees. Job descriptions are provided as well as nationwide average monthly salaries by job title. Average salaries by job title for 17 major cities in the United States also are included.

SECTION 8. COST-BENEFIT CONSIDERATIONS*

Introduction

The purpose of this section is to explore cost-benefit considerations as they might apply to the possible implementation of an attitudinal research information system. The identification and measurement of cost categories is considered as well as the much more difficult task of identifying anticipated benefits and placing a value on these benefits. An analytical model for cost-benefit analysis in this context is offered, and alternatives for establishing a charge structure for services provided by an attitudinal research information system are discussed.

The decision to conduct this feasibility study of an attitudinal RIS grew out of an awareness of the need for designing and developing a computer tool to increase the efficiency of personnel research by standardizing data collection, storage, and retrieval procedures; by making complex research designs and analytic techniques readily available to individual researchers; and by building upon information already available from Navy personnel files and from past attitudinal and opinion studies as well as from previous research on organizational behavior. The economic advantages of secondary analysis of existing data bases have been cited by a number of authors. Hyman (1972, pp. 6-7) has conservatively estimated that the phases of a survey essential to producing data in a processed form ready to be analyzed consume about 40 percent of the total budget. The remainder of the budget is consumed in very much the same fashion whether the investigator is the primary or secondary analyst of the data. Hyman (1972) notes that, "In 1970 in the United States, the bill for the services necessary simply to produce new data from a national sample of 1500 to 2000 cases could easily be $60,000" (p. 7). Bisco (1970) makes an even higher estimate of the cost of acquiring a data collection. "A new data collection may cost $75,000 or more; a copy of an appropriate computer-processable data collection may cost as little as $5....The cost of acquiring a copy of a data set can be as little as one fifteen-thousandth the cost of conducting a new data-collection operation" (p. 2). Bisco (1970) concludes that "data banks are a necessary development in the provision of facilities because social science data collections are very expensive, and therefore taxpayers benefit if these data collections can be made accessible to researchers who can use copies for the same or other analytic purposes" (p. 2).

Emery (1974, p. 967) reports that annual worldwide expenditures for computer-based information systems easily exceed $50 billion. As large as this figure is, he regards it as small in comparison to the costs of nonautomated information processing. The inefficiency of nonsystematic and uncoordinated manual methods for processing information is becoming less and less tolerable

* The contributions of Ingeborg M. Kuhn, Graduate School of Business, Stanford University, to this section are gratefully acknowledged.

in a society with dwindling national resources. Emery (1974) concludes, "As long as we have limited resources and unsatisfied requirements, we cannot escape from the economic consequences of a decision to allocate resources to develop an information system" (p. 967). Wilson (1972, p. 48) notes that resources are rarely, if ever, equal to information needs. Therefore, management typically is faced with the decision of how to allocate scarce resources ---both money and personnel---among competing needs and requirements, only one of which is the need for a better information processing capability.

King and Bryant (1971) make the important observation that the objectives of an information system may be viewed from a number of levels by management. "A higher-level objective is to provide informational support in meeting organizational goals; a lower-level objective is to satisfy users' information needs, requests, or wants. Regardless of the level, however, [information] systems play a support role and are of value only insofar as they assist organizations or individuals in meeting their goals" (p. 218).

"Once the objectives of a system are established, cost-benefit analysis can be employed in aiding allocation decisions. The methodology involves comparison of the cost of a project with its benefit over time" (Cooper, 1973, p. 17). Carlson (1974, p. 61) has noted that cost-benefit analysis is more often used in *feasibility studies* [italics ours] of information systems than in evaluations of these systems after they are implemented. From a theoretical and traditional point of view, cost-benefit analysis has been used to evaluate proposed projects. Historically, the basis premise for cost-benefit analysis has been to provide a tool for decision making in the public sector with respect to the allocation of scarce resources and to aid the decision-making process regarding the initiation of new public projects. Basically, cost-benefit analysis is an attempt to compare the costs incurred by undertaking an activity with the potential benefits to be derived. It requires a systematic and disciplined analysis of both the costs and the benefits beyond that which is likely to be undertaken for the direct requirements of fiscal control (Ramsey-Klee, Ed., 1970, p. 285). Simply stated, a cost-benefit analysis is a systematic identification, measurement, and placing a value on all costs and benefits over time associated with a project that is designed to achieve specific goals. In contrast, cost-effectiveness analysis is designed primarily to compare the economic efficiency of alternative systems for utilizing resources which are directed at the same objective. The purpose of a cost-effectiveness analysis is to indicate whether or not the output of one or another system is likely to require fewer resources to attain the desired degree of accomplishment, or, alternatively, whether the same dedication of resources can provide greater output (Ramsey-Klee, Ed., 1970, p. 284). At this early stage in considering the merits of proceeding with the design and development of an attitudinal RIS, the question is not so much one of which system implementation might be the most cost effective, but rather one of whether the project under consideration is worth implementing at all, since the resources required would have to be drawn from other uses. Therefore, the ensuing discussion will concern itself with how to identify and measure both costs and anticipated benefits and with the development of an analytical model for cost-benefit analysis.

Identification and Measurement of Costs

Wilson (1972) points out that "...no manager can expect to find a cost system that he can apply 'as is' to his own operation" (p. 41). In arriving at estimates of the costs of an information system, two cost categories need to be taken into account---investment costs and operating costs. The investment costs are the one-time costs required initially to design, develop, and implement the system, while the operating costs are the recurring costs of operating and maintaining the system year by year. The operating costs may also be subdivided into fixed costs and costs that vary by frequency of use and level of use (King & Bryant, 1971, p. 219). *Fixed* costs are those costs invariant to the quantity of services provided and not related to optional services. They are those costs required to run the basic system regardless of volume and to meet the primary operating objectives. There may be two types of *variable* costs. One type relates to the volume of services or transactions. The other type of variable cost relates to optional services that are not considered an essential component of a set of services needed to meet the basic system operating objectives.

The accounting procedures used by most Federal agencies regard computer equipment as a capital acquisition that appears as a line item in the budget for the year in which the expenditure is made. Capital acquisitions typically are not depreciated or amortized over future years of operation, as these costs usually are in the private sector. If the computer selected to support the RIS operation were to be purchased, then it would be regarded as an investment cost since it would be acquired prior to the first day of operation. If the computer equipment were to be leased, its cost instead would be regarded as an operating cost for the year in which this cost was incurred. This same line of reasoning applies to computer software. If the software were to be purchased, or developed locally or by a contractor prior to the initial operation of the system, it would be regarded as an investment cost. Software developed subsequent to the initital day of system operation or leased software would be regarded as an operating cost for the year in which this cost was incurred.

The following list of possible cost categories is offered as an appropriate starting point for arriving at an estimate of the total cost of designing, developing, and implementing an attitudinal RIS. Cost categories should be included only if they represent costs that would not be incurred if the system were not implemented.

Investment Costs. This category includes one-time expenditures required initially to establish the RIS operation.

Labor. This category reflects total direct labor costs of RIS personnel prior to the time that the facility becomes routinely operational. It includes the cost of indexing the data bases to be contained in the system and preparing them in machine-readable form for data entry.

Contractors. This category includes the costs incurred if contractors are used to perform or assist in planning, designing, developing, and installing the RIS.

Training. This category includes the cost of the various types of training associated with the first group of people to staff the RIS operation and the initial training of potential system users.

Documentation. This category includes the cost of writing and the initial printing of such documents as public relations brochures, procedural handbooks, training manuals, system documentation, and other documents pertaining to the operation of the RIS facility.

Supplies. This category includes the costs incurred initially for printed forms, business and administrative supplies, bookkeeping material, administrative forms, and other office supplies.

Computer Hardware. This category includes the cost of the computer equipment selected to support the RIS operation if this equipment is purchased prior to the initial day of system operation.

Computer Software. This category includes the cost of any system software purchased, or developed locally or by contractor prior to the initial day of system operation.

Protection of Privacy. This category includes any mechanisms or protection systems that are added to the basic operating system hardware and software in order to protect the privacy of data. These costs have been estimated at 10 to 20% of the basic data processing costs, but may go as high as 50%, depending on the complexity and adequacy of the protection system employed.

Other System Equipment. This category includes the cost of other system-related equipment such as computer terminals, remote printers, graphics plotters, card readers, and card punch machines purchased prior to the initial day of system operation.

Storage of Data Bases. This category reflects the initial cost of both on-line and off-line storage media for storing the data base files and includes dedicated disk packs and magnetic tapes.

Office Equipment. This category includes the costs associated with typewriters, calculators, adding machines, copy machines, and any other office equipment that is to be used in the day-to-day operation of the RIS office acquired prior to the initial day of system operation.

Furniture. This category includes the cost of tables, desks, chairs, filing cabinets, and other furniture to be used in the RIS facility acquired prior to the initial day of system operation.

Facilities. This category includes the cost of architects, engineering, land, and construction or renovation of major facilities associated with the RIS operation along with real property installed equipment such as air conditioning, false floors, special electrical cabling, lighting, and telephone/telegraph wiring.

Other. This category includes costs that cannot be identified specifically to line entries under *Investment Costs*, such as travel, utilities, and overhead incurred prior to the operational phase.

Operating Costs. Under operating costs are listed the categories in which the day-to-day costs of operating the RIS facility are stated, including additional system development and system support activities. These categories represent costs that recur from reporting period to reporting period. The operations phase begins with the first day of actual operation of the RIS facility.

Labor. This category reflects the total direct labor cost of operating an RIS, including the salaries and wages of such personnel as facility manager, computer operators/programmers, technicians, tape librarian, data input specialists, steno/clerks, and other personnel. Also to be specified in this category is the payroll burden, i.e., the employer's costs associated with employees' vacations, sick leave, retirement, unemployment insurance, health insurance, and other fringe benefits.

Replacement Training and Continuing Education. This category includes the costs associated with training of new RIS personnel and of staff retraining in the RIS facility after it becomes operational. Included is the portion of the RIS supervisor's time that is devoted to assisting new RIS personnel in learning their tasks. Also included in this category is the cost of training new users of the RIS, continuing education, and the salary of the training specialist.

User Time Preparing To Use the RIS. This category includes the time and effort spent by system users in order to formulate search strategies, to consider appropriate analysis procedures, and to describe the desired format for system output.

User Time Waiting To Use or Actually Using the RIS. This category includes the time and effort spent by system users in waiting to use the system, waiting for the system to respond, and their actual productive time spent at the terminal.

Maintenance of the RIS's Data Base Holdings. This category includes the costs associated with maintaining the currency, comprehensiveness, and accuracy of the data bases contained in the system and the indexes to these holdings. It also includes the cost of documenting, indexing, and preparing new data bases for entry into the RIS.

Storage of Data Bases. This category reflects the ongoing cost of both on-line and off-line storage media for storing the data base files and includes dedicated disk packs and magnetic tapes.

System Operating Costs. This category includes the total costs associated with system operation, including maintenance, equipment rental, the development of new software or the lease of software packages, and purchase of add-on equipment.

Computer Hardware Maintenance. This category includes costs of personnel and materials or maintenance service associated with maintaining the RIS computer equipment.

Maintenance of System and Programming Support Facilities. This category includes maintenance of the catalog of data base holdings, the data base schema, the access processes, sample subsets, programs, and system functions.

Computer Hardware Rental. If the computer equipment is leased instead of purchased, this category should show the leasing cost. This cost typically includes equipment maintenance. If the computer is shared with other applications, show a pro rata share for the RIS operation. Also included are computing services purchased from a time-sharing vendor.

Computer Software. This category includes the cost of leased software or software developed subsequent to the initial day of system operation.

Other System Equipment. This category includes the leasing cost of other system-related equipment---such as computer terminals, remote printers, graphics plotters, card readers, and card punch machines---if these pieces of equipment were not purchased initially. Also included here should be the cost of these devices if they are purchased as add-ons to the system after it becomes operational.

System Communications. This category includes the cost of modems and telephone, telegraph, WATS, and/or any other communications equipment and services used exclusively in the RIS system operation.

Telephone. This category includes the cost of telephone services related to administrative functions of the RIS.

Supplies. This category includes the costs of stationery, postage, paper, file folders, forms, and other office supplies used in the day-to-day operation of the RIS facility.

Facility Rent. This category includes the rent paid for use of the facility in which the RIS is housed.

Consultants. This category includes fees and retainers of all professional consultants who support the RIS operation.

Travel. This category includes the cost of travel for all personnel and consultants associated with the RIS operation when they are on official business for the facility.

Other. This category includes costs that cannot be identified specifically to line entries under *Operating Costs*, but which are incurred as part of the RIS operation, such as utilities and overhead.

Some of the cost categories specified above require additional comment. Under *Investment Costs*, the labor cost entailed in indexing and preparing the data bases for entry into the system is often overlooked, and this cost may be substantial. In some cases, it may be possible to glean at least a portion of the system holdings from data collections already in machine-readable form. Most likely, though, they will require some "cleaning" and improved documentation, and they may also entail the necessity of reorganizing the data base contents in order to fit into the overall data base system. "A large portion of the information in data bases still is entered manually, either indirectly via punched cards, magnetic tapes, or optically read documents or directly via computer terminals" (Wiederhold, in preparation). If the input data are keyed in by means of a terminal, it may be possible to incorporate input verification procedures and error detection mechanisms. Wiederhold (in preparation) points out that in assessing data entry costs, it is not proper to consider only how much it costs to enter a data element. The proper question to ask is, How much does it cost to enter a data element correctly? Wiederhold (in preparation) concludes that "...the cost of detection and correction of an element entered wrongly, as well as the cost of the havoc the erroneous value might have wrought while it was part of the data base, has to be added to the data entry cost. Errors during data entry can be classified into errors occurring before, during, or after the transcription of the data into the computer. If an error can be detected while the source of data is still available, then the cost of correction may only be a few times the cost of the original data entry. If errors are found much later, then considerable time may be wasted in the retrieval of the erroneous data element or in a call back to the supplier of the data." These considerations also apply to the cost category, Maintenance of the RIS's Data Base Holdings, under *Operating Costs*.

Under the cost categories, User Time Preparing To Use the RIS and User Time Waiting To Use or Actually Using the RIS, the effort required to use a system may be underestimated. "Direct components of effort to use the system are the time required to formulate queries, the time spent in order to describe the layout of desired reports, as well as the time needed to type or otherwise enter the search parameters into the system. Another component of the effort required to use a system is the amount of training needed to interact with the system effectively. The additional effort involved in designing a data base system so that its use requires little indoctrination can be well justified in terms of overall system cost" (Wiederhold, in preparation).

Wiederhold (in preparation) also reminds us that "...the cost of storage facilities is a major component of file system cost. There are many instances where the potential value of data is less than the minimal cost of storage. Consequently, we will want to verify that we can afford to store the data that we wish to process." The sobering reality of this trade-off should serve as a guideline in deciding which data bases are valuable enough to be included in an attitudinal RIS.

Identification and Measurement of Benefits

Cooper (1973, p. 32) has remarked that there is little agreement as to what constitutes the benefits of an information service or how to measure benefits. The benefits of having information available arise from the uses to which the information is put. Wilson (1972) observes that "until we can better define and trace the uses and effects of information---its value, price, utility, whatever we want to call them---the idea of benefit, whether measured in dollars or other units, is at least forcing us to search out the purposes that information serves in our society" (p. 55).

Lancaster (1971) lists the following criteria for measuring the benefits of an information storage and retrieval system: actual hard dollar cost savings; loss of productivity if the information is not available; improved decision making (decisions that could not be made or would be made wrongly or badly without the appropriate information); avoidance of duplication of effort; and stimulation of invention. If what may be counted as a system benefit can be measured, then the problem becomes one of valuation. "*Value* is usually defined as that property of a thing which makes it esteemed, desirable, or useful, or the degree to which this property is possessed" (Gregory & Van Horn, 1974, p. 473). The effect of a manager's knowing or not knowing some piece of information and the actions that follow such knowledge are important determinants of its value. But "the benefits and value received from information are not, many insist, necessarily quantifiable in money terms; other considerations have to be taken into account. But many benefits can be and should be tagged with dollars, and because resources are usually stated in dollars, stating benefits in dollars helps relate the two" (Wilson, 1972, p. 52). Another approach to placing a value on benefits is to ask the user what he would be willing to pay for the services of the information system. The benefits that can be measured and valued in monetary terms are considered to be the *tangible* benefits that the system has to offer.

Benefits that are extremely difficult or impossible to measure and place a value on generally are referred to as *intangible* benefits. For example, Emery (1974) maintains that "it is normally impossible to express in monetary terms the effects of improved information. Quantification would require determination of the following:

- The 'surprise' content of the new information---that is, what it tells the decision maker that he would not otherwise know.
- The way in which each potential 'surprise' will alter decisions.
- The effect on the organization's goals brought about by changes in decisions" (p. 969).

Emery (1974, p. 969) further opines that when benefits cannot be translated practically into monetary estimates, subjective judgments inevitably must be relied upon. However, he cautions that these judgments should be supported with appropriate analysis. In this respect, he suggests a number of useful approaches. The ones pertinent to this discussion are the following:

- Quantify intangible benefits in non-monetary terms if possible, for example, reduced time to accomplish a task.
- Make boundary estimates, sometimes called 'best case' or 'worst case' analyses. Even though it may not be possible to determine an objective estimate of a benefit, it sometimes is possible to get managers to give an estimate of an upper and lower bound.
- Determine cost of the lowest cost alternative. If an intangible benefit must be achieved, it is legitimate to ascribe as the value of the benefit the cost of obtaining it using the best available alternative.
- Establish a break-even point. If the value of an intangible benefit cannot be estimated, it is often useful to translate estimated costs into the minimum level of improvement that would justify going ahead with the implementation of the system.

From the preceding discussion, it can be seen how truly difficult it is to identify and measure benefits plus the added obstacle of finding an appropriate way to value benefits once they are identified and measured. Difficult measurement problems can be encountered with both tangible and intangible benefits, but attempts should be made to identify and measure both. The identification and measurement of the benefits accruing from an attitudinal RIS should first be approached from the system objectives. If the objectives are well stated in measurable terms, then the measurement of realized benefits becomes easier. The benefits derived from meeting system objectives can be considered as the primary benefits, which does not preclude the realization

of other secondary benefits associated with the existence of a new system. For example, if the major objective of the RIS is to provide an information support system for Navy personnel researchers, then primary benefits would be those relating to research activities, such as improved documentation of existing data bases, availability of information, and a reduction in the need for additional surveys. However, along with the advent of the information support system, secondary benefits might also be realized, such as reduction of errors, more timely availability of information, and possible improved decision making attributable to the availability of information or new analytical tools.

It should be noted that the benefits may not always be positive necessarily. Negative benefits or disbenefits could result from the implementation of the attitudinal RIS as well. A common example is system user morale, which might be negative in the early days of system implementation. Of course, the benefit and its magnitude may change over time, from negative to positive, or vice versa.

This exposition of the subject does not pretend to provide a solution to the problems that have been posed; however, an attempt is made below to identify both the tangible and intangible benefits to be gained if an attitudinal RIS were to be implemented.

Tangible Benefits (TB). Tangible benefits are the benefits accruing from the RIS that can be measured and valued in monetary terms.

System Cost Savings. This benefit reflects the cost savings to be achieved by replacing the inadequate nonsystem of research information management with a computer-based attitudinal RIS. Expectations of future benefits, however, are often based on anticipated cost containment. This assumption implies that while there may be no net saving when the system is first introduced, the new system eventually will allow growth in needed information services at a smaller increase in cost than can be foreseen otherwise. It is obvious that such reasoning requires even more careful benefit analysis, since the rate of expected growth is always uncertain and affects, when achieved, many aspects of an organization beyond the area of computing management (Wiederhold, in preparation).

Manpower Cost Savings (MCS). This benefit involves the estimated time to be saved by Navy personnel researchers in realizing their information needs and requirements. Affected will be both managers and researchers. Also included here is the better utilization of clerical and support personnel reflected in previous clerical tasks replaced by automation, improved work patterns, and increased efficiency.

Secondary Analysis Cost Savings (SACS). This benefit comprises the cost of collecting survey data and preparing it in a machine-processable form that can be avoided by the secondary data analyst.

Intangible Benefits (IB). Intangible benefits are the benefits accruing from the RIS that are difficult or impossible to measure and value in monetary terms.

Provision of Information. This benefit reflects the availability of needed information in the RIS in ways that are not possible without the implementation of the system.

Exhaustive Coverage of Information (ECI). This benefit consists of a more exhaustive coverage of the data collections deemed important by the Navy personnel research community.

Improved Documentation of Data Bases (IDDB). This benefit reflects the improved documentation of data bases necessitated by their inclusion in the RIS.

Accuracy of Information (AI). This benefit comprises the increased accuracy of information if maintained in the RIS.

Flexible Presentation of Information (FPI). This benefit involves the ability of the RIS to analyze data and present results in the form desired by the information user, a measure of system flexibility.

Currency of Information (CI). This benefit involves the up-to-the-minute nature of data base holdings in the RIS, a measure of system currency.

Timely Presentation of Information (TPI). This benefit reflects the ability of the RIS to locate and present required information on a more timely basis, a measure of system responsiveness.

Availability of New Analytical Tools (ANAT). This benefit includes user access to new analytical tools not available before the advent of the RIS.

System User Morale (SUM). This benefit reflects system user satisfaction with and acceptance of the RIS in preference to a working environment without the RIS.

Information for Management Response (IMR). This benefit consists of the kinds of information readily available to management from the RIS to aid them in responding to information requests from outside sources.

Other Potential Benefits (OPB). This benefit category includes any outcome benefits from having the results of research available sooner upon which some action may be taken.

The above list of tangible and intangible benefits is not intended to be definitive. Rather, it represents a first approximation to an enumeration of benefits that can be expected to accrue from the installation of an attitudinal RIS. The list should be subject to continual review and modification if a cost-benefit analysis is actually undertaken.

Furthermore, with respect to the measurement of costs and benefits associated with the secondary analysis of existing data, care will have to be taken to include all relevant elements before and after the implementation of the RIS to get some measure of benefits. For example, costs associated with the design, implementation, and analysis of survey data need to be identified and compared to the costs associated with obtaining similar data via the RIS. It seems likely that there may be some disadvantages (or negative benefits) to the secondary analysis of existing data, as well as the positive benefits that have been mentioned. Questions that should be considered include the following: (1) Will the nature of the data bases to be assembled for secondary analysis be adequate, both in terms of breadth and depth, to meet future research needs, particularly future research projects that have yet to be defined?, (2) To what extent will future research projects be constrained by the use of the existing data bases?, and (3) To what extent will future research projects need to supplement the data available with additional survey information? The fundamental question appears to be, Will the need for new surveys to gain additional data be eliminated or at least reduced enough to make the implementation of the new information system worthwhile? Considering the high costs of collecting survey data and preparing these data in computer-processable form, the positive benefits to be gained through secondary analysis of existing data bases would appear to markedly outweigh the possible negative benefits cited above. In fact, the positive benefits accruing from the costs to be avoided in secondary data analysis may be large enough in themselves to justify proceeding with the development and implementation of an attitudinal RIS.

An Analytical Model for Cost-Benefit Analysis

Generally, a proposed project being evaluated by cost-benefit analysis has an expected life, and associated benefits will be realized during that time period but may also be realized after the project is completed or terminated. Traditionally, a cost-benefit analysis covers the entire period during which costs are incurred and benefits are realized, and measurements are in terms of the present value of projected costs and benefits to be realized over time. If the analytical model includes a determination of the present value of future dollar streams (the projected costs and benefits), then an appropriate discount rate will have to be selected (Wildavsky, 1966, p. 373). This decision requires a judgment as there is no predetermined rate for this purpose. "In practice, the agencies of the Federal government have employed a wide range of discount rates, usually without giving a reason. Nevertheless, the consequences of choosing a high or a low discount rate are clear. A low discount rate favors projects or programs with benefits accruing in the distant future; a high rate favors projects with costs in the distant future.... When a project or program is short lived, with both benefits and costs concentrated in the near future, the choice of a discount rate is of minor or no consequence. Indeed, for a short-lived program discounting may be dispensed with" (Klarman, 1973, p. 174).

If the costs and benefits of the RIS are expected to have a relatively smooth pattern or relationship over time with little variance, an alternative approach may be to look at a representative operating period, such as a year.

With respect to the RIS, if its objectives are defined in terms of improvement to ongoing research activities and operations rather than the attainment of some ultimate outcome, this emphasis may provide the justification for looking at a representative period rather than the total life of the project. If the representative period approach is chosen, care should be taken not to simply add investment costs to the operating costs for the cost-benefit comparison. The investment costs should be allocated over a reasonable time period (say, five years) so that only a portion of these costs is included in the cost-benefit comparison.

The identification and measurement of costs over time is the traditional approach, however, and the more comprehensive one. If a cost-benefit analysis of an attitudinal RIS were to be conducted, a determination of which was the more appropriate approach might be made after an initial assessment of costs and benefits. In the presentation of the analytical model for cost-benefit analysis provided below, the traditional approach has been employed.

The role of a feasibility study of a hypothetical information system is to *assess* the technical feasibility of implementing a system with certain design goals. In the design and development phases of an information system, the primary objective is to *demonstrate* the technical feasibility of the system. "However, when management considers the operational feasibility [of an information system], the primary objective should shift toward demonstrating economic feasibility, and performance measures should reflect both costs and benefits" (King & Bryant, 1971, p. 217).

Expanding on King and Bryant's line of reasoning (1971, p. 220), for an information system operating at a given performance level, C and B may be used to denote the costs and the benefits, respectively, provided that both benefits and costs can be expressed in the same units. Three measures which are readily derived from B and C are the cost-benefit ratio, the net benefit (or profit), and the net benefit cost ratio. These three measures may be computed as follows:

Cost-Benefit Ratio	$CBR = B \div C$
Net Benefit	$NB = B - C$
Net Benefit Cost Ratio	$NBCR = NB \div C$

In order for an information system to be judged economically feasible, one or both of the following two conditions should hold:

$$CBR \geq 1$$

$$NBCR \geq 0$$

For a situation where adequate resources are available for the selection of one of several alternative systems, the generally accepted measure of relative economic worth is the net benefit, i.e., the benefit minus the cost. In a commercial enterprise, this choice is equivalent to maximizing the profit.

The cost and benefit terms in the equations shown in this section represent the present value of projected costs and benefits to be realized over the expected life of the RIS. The three cost-benefit comparisons defined above may be related to the system costs and benefits described earlier in this section as follows:

Costs will consist of the total investment costs to design, develop, and establish the RIS plus the costs to operate the system on a day-to-day basis, including additional system development and system support activities, as delineated in the discussion of the *Identification and Measurement of Costs*.

Benefits will consist of the following:

B = TB + IB where:

TB = Tangible Benefits defined to be

= MCS + SACS where:

MCS = Manpower Cost Savings (Manpower costs to be eliminated by the greater efficiency of the RIS)

SACS = Secondary Analysis Cost Savings (Data collection and computer preparation costs to be avoided in secondary analysis)

IB = Intangible Benefits defined to be

= ECI + IDDB + AI + FRI + CI + TPI + ANAT + SUM + IMR + OPB where:

ECI = Exhaustive Coverage of Information (Increased coverage of important data collections by the RIS)

IDDB = Improved Documentation of Data Bases (Increased level of data base documentation)

AI = Accuracy of Information (Increased accuracy of information in the RIS)

FPI = Flexible Presentation of Information (Increased flexibility of the RIS to analyze data and present results)

CI = Currency of Information (Decreased lag time required to input new data bases into an information system)

TPI = Timely Presentation of Information (Increased responsiveness of the RIS to short-term information requests)

ANAT = Availability of New Analytical Tools (Increased access to desired analytical tools)

SUM = System User Morale (Preference for the RIS to a working environment without the RIS)

IMR = Information for Management Response (Increased availability of information to management for responding to information requests from outside sources)

OPB = Other Potential Benefits (Outcome benefits associated with having the results of research available sooner)

If the costs and the benefits of an attitudinal RIS can be expressed in monetary terms using the investment and operating cost categories and the tangible and intangible benefits that have been defined in this section, then a value for *C* and *B* can be determined. These values then can be substituted in the equations for calculating the Cost-Benefit Ratio and the Net Benefit. The Net Benefit Cost Ratio can be calculated once the Net Benefit is known. If a cost-benefit analysis is actually undertaken for the RIS, it probably will be very difficult if not impossible to arrive at dollar values for *C* and *B*. Therefore, the final management decision of whether or not to proceed with the system's implementation most likely will have to be based on subjective estimates of the system's costs and benefits. However, Lancaster and Fayen (1973, p. 401) caution that an organization funding an information system must feel that the benefits of the services outweigh the costs of providing these services.

Alternatives for Establishing a Charge Structure for Services Provided by an Attitudinal Research Information System

The history of information services reveals that early information storage and retrieval systems typically were subsidized during their developmental and testing phases. During this time the consumers of these services did not pay for them. Since the costs of information services either were buried in the overhead item for the organization or were paid for by a source of funds received from outside the organization, these services tended to be regarded as free goods by the consumers. However, both system managers and consumers pay greater attention to the efficient operation and utilization of an information system when it is evident that the services provided cost money (Becker & Hayes, 1963, pp. 239-240).

As information systems have come of age and are proliferating in our society, the requirement to begin paying their own way has become evident. In a nonprofit environment, the management of an institution providing information services may be willing to work within a charge structure that will support system costs including equipment replacement. If the institution is a subsidiary of a larger parent organization, a line item in the budget of the subsidiary institution may amount to a partial subsidy of the information system by the parent organization, with the balance of the system's cost being recovered through user charges. For example, were an attitudinal RIS to be developed at NPRDC, a possible mechanism for funding its development and operation might be an allocation by the Chief of Naval Material (NAVMAT) for the support of RIS activities, with the remainder of the support coming from user charges apportioned to the budgets of individual code numbers according to actual usage of information services. Such a support mechanism would involve management decisions at several levels concerning the allocation of scarce resources among competing projects.

In a profit-making environment, such as a service bureau that provides information services, fee for service is the usual method of charging customers. The fee is set to include the recovery of actual costs, including development costs, for the system plus a profit margin. If the costs to operate the system increase, this increase is passed on to the consumer in the form of a hike in fees.

The method for establishing a fee for information services in either a profit or a nonprofit environment may be simple or fairly complex, depending on a number of trade-offs. "An elaborate pricing mechanism is only justified if users have considerable control over their use of information system services and if the financial control mechanism is such that they are motivated to improve their performance. But in a large, relatively decentralized organization with strong financial incentives, the improvement in efficiency and effectiveness brought about by a rational pricing scheme is well worth its added complexity and administrative difficulties" (cited in Emery, 1974, p. 971). A number of organizations apparently have concluded that the complexity and administrative burden entailed is offset by the fairness of the pricing scheme. Complex algorithms for calculating user fees based on utilization of individual computer components are in use at many computer facilities. These algorithms typically make a summation of costs across such individual computer operations as terminal connect time, time used by the central processing unit, number of input/output operations, use of various storage media, disk mounts, tape mounts, and number of pages printed. In most instances, the actual charge to the user is calculated by the computer as part of the job run. The Naval Electronics Laboratory Center and the Naval Undersea Center both use this type of algorithm in their computer billing procedures. The NELC charge structure for computer services makes a low charge for utilization of core memory, but penalizes users with high input/output utilization.

Cooper (1973) has commented as follows on the problems associated with establishing a fair charge structure for computer service. "The increased sophistication of computer systems has often led to increased productivity but has also caused numerous problems in the determination of equitable prices for

computer service. Most of the problems stem from the introduction of multiprogramming, multiprocessing, and time-sharing where tasks are competing for computer resources and where there is no 'standard' or fixed time within which a process or program will terminate. One approach to solving the accounting aspect of this problem has been suggested by Rettus & Smith [1972]. They report on an accounting system for a computer center that accumulates operations, managerial support, systems support, and programming support costs into an extensive series of accounts. Then, employing past usage statistics of the computer facility, they develop standard costs. These costs serve as the basis for charging an individual user for computer center services (including computer time). Variances from the standard costs are accumulated and adjustments in the standard rates are made periodically" (p. 29). Emery (1974) also comments on this issue. "The determination of costs for a given job run in a multiprogramming environment requires a number of cost allocations that are largely arbitrary, nonreproducible, and involve factors beyond the user's control. Furthermore, they bear no direct relation to the quality of the service provided the user, such as turnaround time or the time of day at which service is provided. Many of these problems can be eliminated by establishing prices charged to users that bear no direct relation to costs as determined by conventional cost accounting techniques. Instead, prices are set in a way that rations available capacity and motivates users to use the system efficiently and effectively....To the extent possible, the objective should be to charge users for information in a way that fosters the efficient and effective use of information processing resources" (p. 970).

It would appear that if an attitudinal RIS were to be implemented, the charge structure for paying for the information service should be devised to encourage system efficiency and effectiveness as recommended by Emery. The charge structure probably should be tied to the amount of time that a user spends at a terminal (i.e., terminal connect time) in order to simplify the accounting and billing procedures. The amount of the charge per unit of terminal usage could be related to the actual operating costs of the system so that the funds needed to balance the budget for the system's operation would be recovered. If part of the system's operation were subsidized, then charges to individual users could be reduced by the proportion of funds contributed by the subsidizing organization. Periodic adjustments to the standard charging scheme could be made at regular intervals to reflect the actual cost experience of operating the RIS.

ANNOTATED BIBLIOGRAPHY

1. Becker, J., & Hayes, R. M. *Information Storage and Retrieval: Tools, Elements, Theories*. New York: Wiley, 1963.

This book provides a foundation and structure within which developments in information retrieval and allied fields can be viewed for their relationship and interaction with each other.

2. Bisco, R. L. (Ed.). *Data Bases, Computers, and the Social Sciences*. New York: Wiley-Interscience, 1970.

This book is a product of the Fourth Annual Conference of the Council of Social Science Data Archives. It includes post-edited versions of some of the papers that were presented concerning the interrelations between data banks, computer technology, and the needs of the social sciences.

3. Carlson, E. D. Evaluating the impact of information systems, *Management Informatics*, 1974, *3*(2), 57-67.

This paper discusses techniques for evaluating the impact of an information system on the organization(s) it serves. Six evaluation methods are outlined: event logging, attitude surveys, rating and weighting, system measurement, system analysis, and cost-benefit analysis.

4. Cooper, M. D. The economics of information. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 8). Washington, D.C.: American Society for Information Science, 1973. Pp. 5-40.

The analysis of the economics of information is concerned with the resources available to promote the process and the constraints that limit it. This review begins with a brief presentation on the micro-economics of information and includes a discussion of the tools and methodologies available for making resource allocation decisions.

5. Emery, J. C. *Cost/Benefit Analysis of Information Systems*. Chicago: Society for Management Information Systems, SMIS Report No. 1, 1971.

One of the few papers that directly addresses the issue of cost-benefit analysis of information systems. The major emphasis is directed toward an analysis of various system features rather than a system as a whole.

6. Emery, J. C. Costs and benefits of information systems. In *Information Processing 74*. Amsterdam: North-Holland Publishing Co., 1974. Pp. 967-971.

This article takes the position that cost-benefit analysis of information systems does not offer a panacea, but it can provide valuable information to decision makers faced with the question of how to allocate resources among competing projects. The paper deals with such issues as efficiency vs. effectiveness, determination of information requirements, the analysis of benefits, and the treatment of costs.

7. Gregory, R. H., & Van Horn, R. L. Value and cost of information. In J. D. Couger & R. W. Knapp (Eds.), *System Analysis Techniques*. New York: Wiley, 1974. Pp. 473-489.

The costs of operating an information system depend on many factors, the most important of which are identified in this chapter. The authors discuss the quality, quantity, and timeliness of information in respect to its relevance to the decision-making process. Other factors touched on briefly are system capacity, flexibility, communications, processing schemes, and rate of transition.

8. Hyman, H. H. *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities*. New York: Wiley, 1972.

This book contains convenient lists of research designs for secondary analysis of survey data, problems amenable to study, archives, and other sources of data. It includes detailed case studies of the research process and of discovery and productivity in secondary analysis. In addition, there is an exhaustive bibliography and a detailed, analytic table of contents for easy reference.

9. King, D. W., & Bryant, E. C. *The Evaluation of Information Services and Products*. Washington, D.C.: Information Resources Press, 1971.

This book is an excellent distillation of the experience of Westat Research Inc. in the area of evaluation of information services. A construction of a detailed system model depicts the relations among system features, costs, performance, and benefits. Guidance is given as to what to measure, how to measure, and how to interpret results. Also contains brief and lucid primers on user surveys, statistics, sampling methods, and experimental design.

10. King, D. W., & Caldwell, N. W. *Cost Effectiveness of Retrospective Search Systems*. Washington, D.C.: American Psychological Association, 1971. (ED 051 837)

A comparison of 36 retrospective search subsystems for the American Psychological Association: three search modes at four levels of recall, and six screening alternatives.

11. Klarman, H. E. Application of cost-benefit analysis to health systems technology. In M. F. Collen (Ed.), *Technology and Health Care Systems in the 1980's*, DHEW Publication (HSM) No. 73-3016, 1973.

This article does not specifically refer to information systems, but it contains a very good description of cost-benefit analysis and theoretical economic issues involved in cost-benefit analysis.

12. Lancaster, F. W. Cost-effectiveness analysis of information retrieval and dissemination systems. *Jour. Amer. Soc. Information Sci.*, 1971, *22*(1), 12-27.

An article dealing with cost-effectiveness and cost-benefit models as applied to information systems in analyzing the relations among system features, costs, performance, and benefits.

13. Mishan, E. J. *Cost-Benefit Analysis: An Introduction*. New York: Praeger Publishers, 1971.

A classic work on cost-benefit analysis. Referenced by many authors.

14. Prest, A. R., & Turvey, R. Cost-benefit analysis: A survey. *The Economic Journal*, December 1965, *75*, 683-735. Reprinted in *Surveys of Economic Theory* (Vol. III). London: McMillan and Co., Ltd., 1966. Pp. 155-207.

A classic work on cost-benefit analysis. Referenced by many authors.

15. Ramsey-Klee, D. M. (Ed.). *Provisional Guidelines for Automated Multiphasic Health Testing and Services* (Vol. 3). Rockville, Md.: U.S. Department of Health, Education, and Welfare, Health Services and Mental Health Administration, January 1970.

This volume contains the proceedings of an invitational conference on multiphasic health testing and services (AMHTS). Part One discusses the use and purposes of AMHTS in health care. Part Two deals with technology in AMHTS, and Part Three is concerned with human factors in AMHTS. Part Four is devoted to cost and cost analysis in AMHTS.

16. Rettus, R. C., & Smith, R. A. Accounting control of data processing, *IBM Systems Journal*, 1972, *11*(1), 74-92.

Description of an accounting system for a computer center that accumulates operations, managerial support, systems support, and programming support costs into an extensive series of accounts. Then, employing past usage statistics of the computer facility, standard costs are developed. These costs serve as the basis for charging an individual user for computer services.

17. Sharpe, W. F. *The Economics of Computers*. New York: Columbia University Press, 1969.

The first part of this book describes the tools of micro-economics. The second part of the book applies this methodology to computer problems. The size and growth of the computer industry is explored, the issue of purchase vs. lease of computer equipment is analyzed, the allocation of computing resources is discussed, methodologies for measuring system performance are presented, component costs are discussed, and various pricing schemes for rationing the resources supporting the computer facility are investigated.

18. Wildavsky, A. The political economy of efficiency: Cost-benefit analysis, systems analysis, and program budgeting, *Public Administration Review*, December 1966, *26*(4), 292-310.

The major purpose of this paper is to take the most recently popular modes of achieving efficiency---cost-benefit analysis, systems analysis, and program budgeting---and show how much more is involved than mere economizing. Another purpose is to describe the characteristic features of these three techniques for decision making: the aids to calculation designed to get around the vast areas of uncertainty where quantitative analysis leaves off and judgment begins.

19. Wilson, J. H., Jr. Costs, budgeting, and economics of information processing. In C. A. Cuadra & A. W. Luke (Eds.), *Annual Review of Information Science and Technology* (Vol. 7). Washington, D.C.: American Society for Information Science, 1972. Pp. 39-67.

An examination of the approaches and trends in the area of costs, budgeting, and the economics of information systems. The author reviews reported progress in cost analysis and reporting; cost-effectiveness analysis; programming, planning and budgeting systems; and cost benefits and marketing.

SECTION 9. CONCLUSIONS AND RECOMMENDATIONS

The need for a computer-based attitudinal research information system (RIS) to support the activities of the Navy personnel research community was documented in Section 3. Navy personnel researchers do not have any systematic way of knowing what data bases exist that might be relevant to their research interests. Inadequate or nonexistent documentation of existing data bases is a chronic problem. In addition to the problem caused by poor or inadequate documentation of magnetic tape files created by Navy personnel researchers themselves, another significant problem revolves around coding or format changes made over time by the Navy Bureau of Personnel that introduce incompatibilities in data if one wants to analyze or select from historical data sets. The present inability to search files in a dynamic manner in order to provide quick responses to inquiries from Navy management is a direct result of the lack of file documentation and the absence of a data base design. Any request to search a file on an unindexed attribute or to relate multiple attributes for an individual retrieved from several files is impossible now without special programming since there is no overall data base design that provides for these information processing requirements. Thus, the available computer support, as compared against an ideal standard, is hardware and software limited, inflexible, and often unresponsive.

The conclusion reached by this study team is that it would be technically feasible to implement an attitudinal RIS that could go a long way toward helping to solve the research information management problems described above. The unique contribution that an attitudinal RIS would make in this area would be to systematize the way in which attitudinal research data bases are documented, indexed, and stored. The system design should provide for interactive, on-line data exploration and analysis via CRT terminals that have printing capability. A graphics terminal with copier-like hard-copy output capability also is recommended.

The benefits to be gained from the use of an attitudinal RIS would be more exhaustive and current coverage of data base holdings, increased accuracy of information, improved documentation of data bases, more flexibility in methods for analysis and presentation of information, and ability to locate and present required information on a more timely basis. Additional benefits might be the availability of new analytical tools, outcome benefits associated with having the results of research available sooner, and increased researcher morale. It also is expected that substantial manpower costs would be eliminated by the greater efficiency of the RIS. However, the largest benefit expected to result from the advent of an attitudinal RIS might be in the area of secondary analysis. The positive benefits accruing from the costs of collecting survey data and preparing these data in computer-processable form to be avoided in secondary data analysis could be large enough in themselves to justify proceeding with the development and implementation of an attitudinal RIS.

Acceptance of an interactive, on-line attitudinal RIS would be critically affected by the effectiveness of procedures employed initially to teach potential users how to use the system. All users would not have the same needs to

use the system. Therefore, not everyone who might be using the system would need to be trained in the same depth. A strong case can be made for preparing instructional materials in several levels of complexity. In addition, three types of training approaches should be considered, and possibly all of them should be implemented: (1) printed instructional manuals, (2) personal instruction by a training specialist, and (3) a tutorial display sequence presented on-line at the terminal itself. This last training approach might very well employ the use of computer-assisted instructional techniques, in particular simulation of the actual system to be taught. It is advised that the program to train new users should be conducted immediately before the installation of the fully operational system.

An instructional mode as well as an abbreviated, short-cut mode of user-system interaction should be provided. A provision for the continuing education of users in new system features and capabilities also should be made. Not all potential users of an attitudinal RIS would want to interact with the system personally. Therefore, provision should be made for users who prefer to delegate their information requests to an intermediary (user surrogate). The user surrogate probably should be the training specialist.

The system design should take into account all of the acceptance variables discussed in Section 6, and as many desirable features as possible mentioned in this section should be implemented. The system designer and developer should not be constrained by having to adhere to old and inadequate methods of file organization and data base structure. Rather the premise underlying the system specifications is that the design and implementation of an attitudinal RIS should incorporate all of the desirable features that current theory and technology have to offer.

The following specific system specifications are recommended:

- A controlled vocabulary approach to information indexing.
- A processor with a cycle speed of one microsecond or better.
- A disk system with a capability to mount more than 60 million characters and a standard access speed of 37 msec per revolution and 25 msec per seek.
- A file system allowing key-based random access to files.
- A file system supporting at least multiple record sizes; if possible, variable sizes.
- Communication capability to drive at least five terminals at up to 9,600 baud.
- A spooling facility for remote output operation.
- An operating system with multiple-user access to shared files in a timesharing or equivalent mode, allowing at least one user write privileges.

- A programming language to support these facilities, or a data base management system that provides the services needed.

- A text-editing system for the maintenance of the files.

Since one of the largest groups of Navy personnel researchers is concentrated at the Navy Personnel Research and Development Center (NPRDC), it would be logical to establish the attitudinal RIS at this facility, with communication links to other installations where personnel research is conducted. To protect the software and educational investment in the system, purchased hardware is recommended rather than implementing the system on a shared, large computer. The choice of purchased equipment would provide stability over time and also permit adaptations to the system as resources and conditions might change, protecting Navy personnel researchers from the constant system transients characteristic of a large computer system shared by many users. A log of user activity would provide the management data needed to make intelligent decisions about which aspects of the system should be emphasized and upgraded, and which aspects should receive a lower priority of attention.

The attitudinal RIS should operate under some kind of written charter that would establish its organization and its mission. This organization should assign responsibility for the operation and maintenance of the system and for user training and continuing education. The charter also should define how the system is to be supported. If user charges are determined to be the best way to financially support the system, the charge structure should be tied to the amount of time that a user spends at the terminal in order to simplify accounting and billing procedures.

A final recommendation concerns the advisability of establishing a point of coordination for all of the information support systems that are under design and development currently at NPRDC, such as the STAR (Status of Program Report) system, CENDEX, RDCS (Research and Development Coordination System), and CCOPS (Control and Coordination of Personnel Surveys). If computer hardware were to be purchased to implement an attitudinal RIS, the additional requirements of these other information support systems should be taken into consideration so that one computing facility at NPRDC could handle all of this Center's computational requirements for the foreseeable future.

APPENDIX A

A SELECTED LISTING AND DESCRIPTION OF SOCIAL SCIENCE DATA ARCHIVES

AMERICAN INSTITUTES FOR RESEARCH (AIR)

Center for Research in Social Systems (CRESS)
Kensington, Maryland

AIR/CRESS comprises two main information divisions: The Information Systems Branch, an automated data bank integrated with a conventional reference library; and the Information Analysis Division which provides bibliographic services, analysis, referral, and consulting services.

BLACK RESEARCH INFORMATION COORDINATING SERVICES, INC. (BRICS)

Black Research Information Coordinating Services, Inc. (BRICS)
Tallahassee, Florida

The collection of books, periodicals, audio-visual material, microforms, and input from consultant specialists in the field covers research and scholarship in the Black American experience, with future coverage planned of other minority groups (Mexican, Puerto Rican, and Indian American). BRICS is a nonprofit corporation and services are primarily intended for Black Americans and other minority groups.

BOWLING GREEN STATE UNIVERSITY

Social Science Information Center
Bowling Green, Ohio

The collection contains the Center's research in quantitative historical studies, and surveys on demography, fertility, education, organizations, and alienation.

CARLETON UNIVERSITY

Department of Sociology and Anthropology
Social Science Data Archive (SSDA)
Ottawa, Canada

A collection of social science data on Canada, including data on education, ethnic groups, community studies, and Canadian public opinion surveys. In addition, SSDA has extensive listings of data held at most data banks and archives in North America. Services are provided to researchers at outside institutions on an individual basis.

CITIZEN'S RESEARCH FOUNDATION (CRF)

Citizen's Research Foundation (CRF)
Princeton, New Jersey

CRF is a nonpartisan, nonprofit organization dedicated to the study of significant aspects of contemporary political finance. CRF serves as a

clearinghouse of ideas and information and conducts research itself. The file collection includes extensive data on contributions and expenditures in presidential campaigns since 1952.

CITY UNIVERSITY OF NEW YORK

Brooklyn College
Center for Migration Studies (CMS)
Brooklyn, New York

CMS focuses on the study and dissemination of information concerning the migration process, both international and internal, and its effect on intergroup relations. Archives, consisting of governmental, organizational, and individual records, biographies, autobiographies, interviews, oral histories, and questionnaires, are presently being developed by the Center. A book collection currently exists.

Hunter College
Political Data Center
New York, New York

The Center provides the students and faculty of the City University of New York with training in and access to quantitative political data. The data archive maintains tapes of student-conducted surveys that serve as a teaching and training tool, and a 1965 survey study of Columbus, Ohio.

COLUMBIA UNIVERSITY

Bureau of Applied Social Research (BASR)
New York, New York

The collection consists of data from surveys, books, and journals which relate to mass communications, educational systems, manpower and demography, comparative international research, sociology of science, and problems of research utilization and methodology. Services are available without restrictions.

School of Public Health
Sociomedical Research Archives
New York, New York

The Research Archives is a repository of published and unpublished reports, questionnaires, and of machine-readable data in the fields of sociomedical research and health care administration. Many of the studies' samples consists of entire households. Qualified scholars connected with academic and research institutions may receive data after the approval of the particular study's original investigator has been obtained.

CONGRESSIONAL INFORMATION SERVICE (CIS)

Congressional Information Service (CIS)
Washington, D. C.

CIS provides access to all Congressional publications (except the Congressional Record), all relevant statistical publications of all federal agencies and subagencies, Congressional committees, judicial sources, and other statistics-producing programs as input for the American Statistic Index (ASI). The scope of CIS is all subjects covered in Congressional documents and data covered in the statistical publications of the U.S. Government - ASI. Broad categories covered by ASI include social, socioeconomic, economic, and environmental statistics. Services are available on a subscription or demand basis.

EUROPEAN CONSORTIUM FOR POLITICAL RESEARCH (ECPR)

ECPR Data Information Service
Bergen, Norway

ECPR collects and disseminates information on the activities of data archives and research centers in Europe for European social scientists. Research projects currently being collected and disseminated by the Service include a statistical history of Norway and a time-series analysis of modernization which is being conducted in Germany.

FLORIDA STATE UNIVERSITY (FSU)

Institute for Social Research
Tallahassee, Florida

The Institute presently consists of five centers: the Survey Data Center, the Center for the Study of Education, the Community Mental Health Research Center, the Urban Research Center, and the Center for Policy Analysis. The Census Processing Center also is housed within the Institute. The subject coverage is a wide range of governmental and societal problems, ranging from the response of local government to rapid growth in Florida to national policies in the area of vocational rehabilitation and education. The Institute's information and consultation services are available to citizens' groups, government officials, other educational institutions, and individuals. The Census Processing Center tries to restrict its services to nonprofit institutions.

GENERAL SERVICES ADMINISTRATION

National Archives and Records Service
Washington, D. C.

The National Archives and Records Service maintains a Machine-Readable Archive of noncurrent machine-readable records of permanent value produced by any agency of the Federal Government.

INDIANA UNIVERSITY

Department of Political Science
Political Science Laboratory and Data Archive
Bloomington, Indiana

The collection of survey and aggregate data includes Indiana electoral and legislative data, several cross-national socioeconomic and stability data sets, American state legislative general election returns, world trade data, and a large number of data sets relating to international organizations. The Archive supports the researchers in political science at Indiana University.

Institute of Social Research
Survey Data Bank
Bloomington, Indiana

Data on tape from approximately 15 sociological surveys are maintained covering sociological and related social researchers and noncommercial organizations.

INTERNATIONAL CITY MANAGEMENT ASSOCIATION (ICMA)

Urban Data Service (UDS)
Washington, D. C.

UDS provides concise, analytical, and timely information from data collected through year-round surveys of local jurisdictions, including all incorporated cities over 2,500 population, all city manager cities, all counties, and all regional councils. All topics of interest to professional municipal management are covered, including use of computers by police; salaries of municipal officers; budgeting, planning, and evaluation of programs; municipal finance; city councils; regional councils; land use; drug treatment; services for the elderly; and unionization. Services are available only to subscribers and to members of the International City Management Association.

Inter-University Consortium for Political Research
see UNIVERSITY OF MICHIGAN

IOWA STATE UNIVERSITY

Department of Sociology and Anthropology
Social Indicators Project
Ames, Iowa

The Project is involved in developing social indicators of human resource and community development, and methodologies for less-developed countries to assess their social conditions. The Project has published several selected bibliographies in the social indicator field and has established a data bank on Iowa health statistics (mortality).

KENT STATE UNIVERSITY

Center for Urban Regionalism (CUR)
Library and Information Services
Kent, Ohio

The subject coverage includes environmental studies, urban problems, civil rights movements, and policy research from original Center data, data received from the Inter-University Consortium for Political Research, literature, and government documents. Services are available without restrictions.

MANPOWER RESEARCH AND DATA ANALYSIS CENTER (MARDAC)

Manpower Research and Data Analysis Center (MARDAC)
Alexandria, Virginia

The Survey Data Bank is currently in operation with 24 surveys on file which include 14 DoD-wide surveys, two surveys of civilian personnel, two Air Force surveys, and seven Army surveys. When completed the data bank will include all major military surveys dating back to July 1970. The data stored for each survey include texts of all items, title, abstract, demographics, and information about the location of response data tapes. The responses themselves are not part of the on-line system.

MICHIGAN STATE UNIVERSITY

Department of Political Science
Political Data Archive
East Lansing, Michigan

The Political Data Archive maintains data archives in the social sciences.

NATIONAL DATA USE AND ACCESS LABORATORIES, INC. (DUALabs)

National Data Use and Access Laboratories, Inc. (DUALabs)
Rosslyn, Virginia

This nonprofit corporation was established specifically to assist census data users in obtaining access to census data on magnetic tape and to computer programs. DUALabs services the START (Summary Tape Assistance, Research, and Training) community as well as the general public.

NATIONAL LEAGUE OF CITIES (NLC)

Urban Observatory Program
Urban Research and Metropolitan Community Service Project
Washington, D. C.

The collection includes published and unpublished reports as well as raw data on urban problems. Services are available to the public and technical assistance is provided to Project participants only.

NATIONAL OPINION RESEARCH CENTER (NORC)

National Opinion Research Center (NORC)
Chicago, Illinois

NORC is affiliated with the University of Chicago. NORC and its division, Survey Research Service, are engaged in the use of the survey method to study problems of society. A collection of surveys on health and welfare, occupations and professional communications, economics and business, political science, education, community affairs and problems, and intergroup or race relations provides information on survey methodology and about public opinion research. Services are for scholarly purposes only and there is a charge.

NATIONAL PLANNING DATA CORPORATION (NPDC)

National Planning Data Corporation (NPDC)
Rochester, New York

The collection includes census data and other sources of demographic data with important socioeconomic variables, historical perspectives, and future projections. All services are available on a purchase basis.

NAVY OCCUPATIONAL TASK ANALYSIS PROGRAM

Navy Occupational Task Analysis Program
Bolling Air Force Base, Washington, D.C.

A large data bank of survey information derived from a questionnaire devoted to the task analysis of Navy jobs.

NORTHWESTERN UNIVERSITY

Council for Intersocietal Studies
Evanston, Illinois

Data archives are maintained in intersocial and comparative studies in the social sciences.

Department of Political Science
International Comparative Political Parties Project (ICPP)
Evanston, Illinois

ICPP's data set pertains to 153 political parties in 52 countries, chosen at randon from ten cultural/geographical areas of the world, and scored on nearly 100 variables relative to the parties' internal organization and their external relations with society. Services are available to outside scholars without restrictions.

OHIO STATE UNIVERSITY

Center for Human Resource Research
Columbus, Ohio

The Center has the National Longitudinal Surveys of Labor Market Experience Files (NLS) that are a revised version of files originally prepared by the U.S. Census Bureau. They contain demographic characteristics, work and educational experience, and attitudes towards work and school for each of four age/sex groups surveyed annually for six consecutive years beginning in 1966-1968.

PRINCETON UNIVERSITY

Princeton University Computer Center
Princeton, New Jersey

The Computer Center provides a variety of data and programming services including the Princeton-Rutgers Census Data Project, a regional information clearinghouse for U.S. Bureau of Census data and its usage; and the Social Science User Services (SSUS), an in-house program providing access to machine-readable data files in the social sciences. SSUS files include three data files from the Inter-University Consortium for Political Research, Survey Research Archives, and International Relations Archive. All data from Inter-University Consortium for Political Research are restricted to academic users; other SSUS programs and data are available to members of Princeton University and by special arrangement to other academic users.

R. L. POLK AND COMPANY

Profiles of Change
Detroit, Michigan

The data consisting of annual door-to-door city directory surveys are on population, housing, labor force, and business. Services are available with various options and prices.

ROPER PUBLIC OPINION RESEARCH CENTER

Roper Public Opinion Research Center
Williams College
Williamstown, Massachusetts

A collection of raw or basic data from sample surveys conducted by research organizations located throughout the world. The data bank currently totals approximately 9,000 studies, with an annual input of some 500 additional surveys. Services include computerized data searching and reference service available for scholarly research.

ROYAL NETHERLANDS ACADEMY OF ARTS AND SCIENCES
(Koninklijke Nederlandse Akademie van Wetenschappen - SWIDOC)

Social Science Information and Documentation Centre
(Sociaal-Wetenschappelijk Informatie- en Documentatiecentrum)
Amsterdam-C, the Netherlands

The Centre acts as the central information service for research in the social sciences in the Netherlands, and as a referral centre with regard to social science literature in general. The Centre includes a lending library for social science report literature and a data archive, the Steinmetz Archives, for secondary analysis of social science research material. The subject coverage is current research in the social sciences in the Netherlands and libraries and documentation centers concerned with the social sciences.

SMITHSONIAN INSTITUTION

Smithsonian Science Information Exchange, Inc. (SSIE)
Washington, D. C.

Since 1949 SSIE has been maintaining a data base of basic and applied research in progress and recently completed research in twelve major fields. SATRA (Science and Technology Research Abstracts), published by G. K. Hall & Co., Boston, Massachusetts, currently has available in the social sciences 7,000 to 9,000 records covering such specific areas as government, manpower, race and ethnic relations, and urban affairs. In the behavioral science field, the coverage is in such specific areas as communication, education and training, learning and retention, and social psychology. Data are available to subscribers. Access to the reports is by subject, by investigator, by funding organization, and by research organization. The information is drawn from some 1,300 input sources ranging from the U.S. Government departments and agencies to private foundations, including reports from foreign sources.

SOCIAL SCIENCE RESEARCH COUNCIL (SSRC)

SSRC Survey Archive
Housed by the University of Essex
Essex, England

The Survey Archive exists to preserve and make available for secondary analysis machine-readable data of interest to social scientists. Data sets are received from individual depositors and from other social science archives all over the world. The Archive is investigating exchange agreements with the Inter-University Consortium for Political Research and with the York University Data Bank. Services are available free of cost to academic researchers, subject only to strictures applied by the depositor.

Steinmetz Archives *see* ROYAL NETHERLANDS ACADEMY OF ARTS AND SCIENCES

SURVEY RESEARCH CENTER

International Data Library and Reference Service (IDLRS)
Berkeley, California

The Service was formerly a unit of the University of California, Berkeley. The collection consists of survey and aggregate data from sample surveys conducted in Asia, Latin America, and the United States covering public opinion; family planning; education; political attitudes, beliefs, and behavior; and other data sets, with emphasis on data from developing nations. Some data sets are "permission required," some are "closed" file, and the remainder are unrestricted.

U.S. BUREAU OF THE CENSUS

International Statistical Program Center (ISPC)
International Demographic Data Directory (IDDD)
Washington, D. C.

The collection includes surveys, official statistical publications, census reports, and data from international organizations. The subject coverage is world-wide demographic and family planning data, with emphasis on the developing countries. Services are available by contract.

Social and Economic Statistics Administration (SESA)
Data User Services Office (DUSO)
Washington, D. C.

The DUSO devises, tests, and applies techniques for improving access to the U.S. Bureau of Census statistical data base. It serves as a focal point for the coordination of requests for data and maps, and prepares general-purpose statistical compendia. Among its numerous functions, it

also conducts studies to identify new or changing requirements of data users. The data input are from censuses and surveys of households, business and industrial firms, state and local governments, and farms; and administrative records for other federal agencies. Coverage also includes foreign trade statistics.

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

Social and Rehabilitation Service (SRS)
Division of Research Utilization
SRS Research Information System (RIS)
Washington, D. C.

Input consists largely of SRS-sponsored research reports covering a wide variety of social and rehabilitation problems and solutions. Services and publications are available without restrictions.

U.S. DEPARTMENT OF LABOR

Bureau of Labor Statistics (BLS)
Information System and Data Bank
Washington, D. C.

The collection consists of surveys and other data generated by the U.S. Bureau of Labor Statistics, including approximately 30,000 monthly time series, as well as several cross-sectional data sets. The subject coverage is historical and current labor and manpower statistics for the nation, the states, and geographic areas. Release of unpublished data is determined on a case-by-case basis.

U.S. OFFICE OF EDUCATION

National Center for Educational Statistics (NCES)
Washington, D. C.

The collection includes reports, surveys, and published literature with statistical data on elementary and secondary education, higher education, and adult and vocational education. Services are available to all, with appropriate caution in releasing information about individually named entities.

U.S. PUBLIC HEALTH SERVICE

Health Resources Administration
National Center for Health Statistics (NCHS)
Rockville, Maryland

The collection includes surveys conducted by the Center; vital statistics encompassing births, deaths, marriages, and divorces; and health

statistics including psychological measurements. Services are available contingent on staff time, and most material is free of charge.

UNITED NATIONS EDUCATIONAL SCIENTIFIC AND CULTURAL ORGANIZATION (UNESCO)

Social Science Documentation Centre (SSDC)
Paris, France

SSDC serves as a clearinghouse of UNESCO documentation in the social sciences and includes social science data of interest to UNESCO and its member states. Use restrictions have not yet been established.

UNIVERSITY OF BALTIMORE

Baltimore Regional Institutional Studies Center (BRISC)
Baltimore, Maryland

The collection consists of institutional research, urban studies, voluntary associations, city planning, social science research, and archival descriptions. Services are available to authorized researchers, with restrictions placed on some materials.

UNIVERSITY OF BRITISH COLUMBIA

University of British Columbia Data Library
Vancouver, Canada

This data base is composed of 160 files of machine-readable nonbibliographic data on magnetic tape including public opinion surveys, Canadian census, and statistical and financial data.

UNIVERSITY OF CALIFORNIA, BERKELEY

Institute of Governmental Studies
State Data Program
Berkeley, California

The collection of state-level data currently includes California Public Opinion Surveys; Legislative Evaluation Study; California Registration and Voting Summaries; State Regulation of Marriage, Separation and Divorce of all fifty states; and Legislative Partisanship. Services include the answering of specific questions of varying degrees of complexity ranging from simple frequency counts or cross-tabulations to parametric and nonparametric multivariate analytic techniques.

University of California, Berkeley *see also* SURVEY RESEARCH CENTER

UNIVERSITY OF CALIFORNIA, LOS ANGELES

Databank of Program Evaluations (DOPE)
Los Angeles, California

The main purpose of DOPE is to accumulate and analyze reported evaluations of programs in the broad mental health and social action fields. Information from the Databank is available to users who have access to the Advanced Research Projects Agency (ARPA) network of the Department of Defense (DoD). Besides the status of program success, the Databank will provide other information about each evaluation report, such as age, sex, race, and income characteristics of the sample population; sample size; condition and treatment method; characteristics of the study design; how and what is measured; and how data are collected.

Institute of Government and Public Affairs
Los Angeles, California

The Institute maintains a library and data archives in government programs and social problems.

Survey Research Center
Social Science Data Archive
Los Angeles, California

The Social Science Data Archive has instituted a program of survey resources development. The resources include studies conducted by the Center in a variety of social science fields, with a well-developed collection of community studies and health surveys.

University of Chicago *see* NATIONAL OPINION RESEARCH CENTER

UNIVERSITY OF CONNECTICUT

Social Science Data Center (SSDC)
Storrs, Connecticut

The Center's membership in the Inter-University Consortium for Political Research and the Roper Public Opinion Research Center and contributions of individual faculty researchers and sponsors of research projects provide researchers with access to rich sources of aggregate and survey data. The SSDC has also secured major surveys sponsored by the media, by other private business corporations, and by such private research organizations as the Carnegie Commission on Higher Education. The Center has been designated by the U.S. Bureau of the Census as a Summary Tape Processing Center. SSDC serves students and faculty in the social sciences.

University of Essex *see* SOCIAL SCIENCE RESEARCH COUNCIL (SSRC)

UNIVERSITY OF FLORIDA

Center for Latin American Studies
Latin American Data Bank (LADB)
Gainesville, Florida

LADB is an interdisciplinary social science data archive which responds to requests from academic users for social, economic, and political data concerning Latin America. Data investigations, time series, and surveys are carried out throughout Latin America.

UNIVERSITY OF HAWAII

Social Science Research Institute
Honolulu, Hawaii

The Social Science Research Institute maintains data archives in its fields of interest. Census services also are available at the University of Hawaii.

UNIVERSITY OF ILLINOIS

Survey Research Laboratory
Data Analysis Section (DAS)
Chicago, Illinois

DAS maintains a Social Science Data Archive which supplies files for secondary analysis to researchers in sociology, political science, economics, and consumer behavior. The collection contains the 1970 census summary data for Illinois and Indiana and 1970 public use samples for the entire country. Data are obtained from U.S. Bureau of Census, Clearinghouse and Laboratory for Census Data (DUALabs), National Opinion Research Center, U.S. Department of Labor, and the Survey Research Center of the University of Michigan. All data and services are available without restrictions.

UNIVERSITY OF IOWA

Laboratory for Political Research
Social Science Data Archive (SSDA)
Iowa City, Iowa

The collection includes survey data on voting and attitudes, as well as studies on party leaders and lobbyists; aggregate data; roll call votes; and background information on Soviet and other East European elites. Census data are available as well as data in the area of political science, sociology, economics, history, geography, and education. Raw data are received from other institutions and from various private and governmental sources. SSDA maintains both restricted and unrestricted files.

UNIVERSITY OF LOUVAIN

Belgian Archives for the Social Sciences (BASS)
Louvain, Belgium

Currently, the Archives are a distinct unit of the Faculty of Economic, Social, and Political Sciences. By collaboration with the various research centers of the University of Louvain and of other Belgian and foreign universities, the Archives make available a collection of data sets for comparative research and secondary analysis in the social sciences.

UNIVERSITY OF MICHIGAN

Institute for Social Research (ISR)
Center for Political Studies
Inter-University Consortium for Political Research (ICPR)
Data Archives
Ann Arbor, Michigan

There are three main Data Archives of broad substantive interest: the Survey Research Archive (SRA) containing national election surveys for the United States and other countries, studies of mass political behavior, socialization, comparative politics, race relations, urban problems, organizational behavior, judicial behavior, and elite studies; the Historical Archives (HI) containing U.S. census, election, and roll call data from 1790 through 1972, various other U.S. aggregate data, and extensive non-U.S. aggregate political, social, and demographic data; and the International Relations Archive (IRA) which holds data on international events, conflicts, dyads, international organizations, alliances, and international systems. ICPR is organized as a partnership between the Center for Political Studies and approximately 190 universities, colleges, and nonprofit research organizations in the United States and abroad. The computer services include the ISR-developed OSIRIS computer program package specifications. Individual faculty members and students of member institutions have open access to ICPR resources, without charge in most instances. Services are available to individuals belonging to nonmember institutions on a charge basis.

UNIVERSITY OF MINNESOTA

Minnesota Family Study Center
Inventory of Published Research on Marriage and Family Behavior
Minneapolis, Minnesota

The Inventory maintains bibliographic control of all publications related to marriage and the family, listing every research item published since 1900 in which some manifestation of marriage or the family exists. Research includes any work which reports empirical data as well as illustrative material.

UNIVERSITY OF MONTREAL

Survey Research Center
Survey Data Bank
Montreal, Canada

Raw data are collected by University of Montreal's Survey Research Center and by York University's Survey Research Center. Studies to date include professional groups such as pharmacists and teachers, nonmedical use of drugs, political and voting behavior, urban sociology, and use of languages in multilingual milieus. Services are intended for University of Montreal researchers, but data also will be made available to appropriate public bodies.

UNIVERSITY OF NORTH CAROLINA

Institute for Research in Social Science
Social Science Data Library (SSDL)
Chapel Hill, North Carolina

The data collections and sources include survey data in political science, anthropology, social welfare, sociology, and urban affairs; Harris Public Opinion Polls; Human Relations Area Files; U.S. Census Data; Southeastern Regional Surveys; Comparative State Election Census Projects; Inter-University Consortium for Political Research; Roper Public Opinion Research Center; International Data Library and Reference Service (University of California, Berkeley); Knight Newspapers, Inc.; and a number of data sets obtained from contributing University of North Carolina scholars. Services are available to users inside the University of North Carolina and to outside researchers for a fee.

UNIVERSITY OF PENNSYLVANIA

Social Science Data Center
Philadelphia, Pennsylvania

The Social Science Data Center collects data and maintains archives in the social sciences.

UNIVERSITY OF PITTSBURGH

Social Science Information Center
Pittsburgh, Pennsylvania

The center maintains a data archive in the social sciences and provides services such as tape copying and consultation. Because of limited staff and budget for census processing, the Center is not well prepared to do processing for organizations outside the University of Pittsburgh.

University Center for International Studies (UCIS)
Social Sciences Information Utilization Laboratory (SSIUL)
Pittsburgh, Pennsylvania

All aspects of social sciences information are included in the collection. Input sources consist of commercially produced abstracts and data bases, an internally developed file of descriptions of publicly available social sciences data sets, and a file of resumes of social scientists in Pennsylvania. Services are available to social scientists without cost or restrictions.

UNIVERSITY OF WISCONSIN

National Program Library and Central Program Inventory Service for the Social Sciences
Madison, Wisconsin

The National Program Library and Central Program Inventory Service for the Social Sciences maintains a program library and data archives in its fields of interest.

Social Science Data and Computation Center (DACC)
Madison, Wisconsin

The Center consists of two divisions, Programming and Computation Service (PACS) and Data and Program Library Service (DPLS). The collection consists of economic, political, and social data from surveys, research reports, and publications. The Center's scope is quantitative research in the social sciences and data processing programs to support such research. Service is provided for social science research and graduate instruction at the University of Wisconsin and elsewhere.

WESTAT RESEARCH, INC.

Westat Research, Inc.
Rockville, Maryland

Data covering information systems in general and social, economic, and demographic statistics, in particular, are supplied by customers, the U.S. Census Bureau, the U.S. Internal Revenue Service, various federal agencies, and sample surveys. Services are available on a contract basis.

Williams College *see* ROPER PUBLIC OPINION RESEARCH CENTER

YALE UNIVERSITY

Social Science Library
Social Science Data Archive (SSDA)
New Haven, Connecticut

SSDA provides access to machine-readable data held at Yale University. Input to the system is derived from the Inter-University Consortium for

Political Research, the National Opinion Research Center, the U.S. Bureau of the Census, and the Roper Public Opinion Research Center.

World Data Analysis Program
New Haven, Connecticut

The Program collects information and quantitative data in international politics, domestic political developments and background conditions of foreign policy in different countries, with emphasis on social conditions conducive to political change, including studies of vital statistics, government expenditures and employment, military personnel and expenditures, behavior in international organizations, national income, political stability, political parties and national elections, trade, communications, and distribution of land and income.

YORK UNIVERSITY

Institute for Behavioural Research (IBR)
Data Bank
Downsview, Canada

The collection includes bibliographic material, and numeric data obtained from individual researchers, governmental agencies, and polls. The collection focuses on sociology, political science, and social psychology. Data are available to users affiliated with York University free of charge, and at moderate charges to others. Access to certain data are restricted.

APPENDIX B

A PARTIAL INVENTORY OF DATA BASES AT NPRDC THAT ARE CANDIDATES FOR INCLUSION IN AN ATTITUDINAL RESEARCH INFORMATION SYSTEM

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Maintain files for the NPRDC research community and facilitate the use and exploitation of these files.	The officer files are of good quality, but they are used infrequently by NPRDC researchers.	Officer Cumulative Attrition File		Annual cumulative received from BuPers as of 30 June.	Indiv.		Mag tape	YES
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Maintain files for the NPRDC research community and facilitate the use and exploitation of these files.	The officer files are of good quality, but they are used infrequently by NPRDC researchers.	Officer Current History or Master File		Entire population; received semi-annually from BuPers.	Indiv.		Mag tape	YES
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Maintain files for the NPRDC research community and facilitate the use and exploitation of these files.	The officer files are of good quality, but they are used infrequently by NPRDC researchers.	Officer Extract of Master File		Entire population; received monthly from BuPers.	Indiv.		Mag tape	YES
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Maintain files for the NPRDC research community and facilitate the use and exploitation of these files.	Changes in format of file over time; supposed to be received 30 June and 31 December.	Enlisted Master File	550,000 records	Entire population; scheduled for semi-annually.	Indiv.		Mag tape (32 reels at 1600 bpi)	YES
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Maintain files for the NPRDC research community and facilitate the use and exploitation of these files.		Enlisted Extract File		Entire population; received monthly.	Indiv.		Mag tape (16 reels)	YES
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Study of NROTC effectiveness.		NROTC Master File to Officer Candidate File			Indiv.		Not yet available	NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Aggregate strength data from enlisted change tapes.		Enlisted Change File	Average of 300,000 changes per month	Entire population; received monthly from BuPers. (Navy Health Research Center has files back to 1965 in a single format.)	Indiv.		Mag tape	NO
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Creation of a reference file to describe Navy activities.		List of Activities with Allowance, Mobilization, and Wartime Complement	5,000 to 6,000	Entire population; received from BuPers activity processing.	One per defined activity			NO
Suiter, Roy; Copeland, Al; Balaban, John; Warrington, Jim	NPRDC 220	Maintain addresses for the various Navy activities.		Activity Addresses		Entire population; received from PAMIPAC, PAMILANT, & CONUS. Current addresses only. Received monthly.	Work Unit - One address per defined activity			NO
Bowser, Sam	NPRDC 301	Organizational criteria - unit performance standards.		Content Analysis of Structured Interviews To Derive Unit Performance Standards		Structured interviews with management personnel of 4 organizational units - Total N=87.	Indiv.	YES	Mag tape	NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Bretton, Gene (as told by Ross Vickers)	NPRDC 301	Identify, measure, & evaluate the time-series changes in basic structure and processes of a large, high-technology, and administratively complex military organization as it evolved from commissioning to fully operational state.	Problems relate to the F14 Aircraft Systems Program at NAS Miramar: (1) Research funded by ONR to principal investigators at UC, Berkeley. Must get official "RELEASE" from all parties in order to store data in external source(s) - RELEASE is doubtful; & (2) missing data in final survey very extensive, for multiple reasons.	F14-ONR	≈500	Entire population.	Indiv., Work Unit, Squadron		On reduced data decks at UC, Berkeley	NO
Dockstader, Steve (as told by Ross Vickers)	NPRDC 301	Investigate the specificity of feedback on performance (using a card sorting task).		Performance Feedback	200 reserves, officers, & enlisted men	Grab sample.	Indiv.		Punched cards and a disk file	NO, but 125 characters per person
Doherty, Linda (as told by Ross Vickers)	NPRDC 301	Investigate (1) if there is a value differential between Navy officers and civilians, and (2) the effect of length and type of service on values.	A 1% sample of officers in the Navy was sent an officer values questionnaire; only 40% responded. These were followed up with a questionnaire collecting information on demographic characteristics, work billets they had held, and their educational history. Down to 12-14% of the original 1% sample, so nature of final sample is unknown.	Officer Values Questionnaire	→→→→→	Two samples: (1) N=105 - responded twice; (2) N=150 NPG students. The NPG students also did pairwise comparisons of the values for the purpose of multidimensional scaling.	Indiv.	NO	Punched cards	NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Holzbach, Bob	NPRDC 301	Analyze effects of manning level and distribution on ship performance during refresher training.	Obtaining 1080-14 data (ship readiness - Naval forces status) from the BuPers microfilm file in Washington, D.C.	Ship Performance During Refresher Training	149+ ships (final sample will include ≈30 additional ships)	Ship performance scores by department during Refresher Training exercises since 1 January 1972 and manning level by rate from nearest quarterly 1080-14 report.	Ship		Punched cards	Not complete as of 12/74
Nebeker, Del (as told by Ross Vickers)	NPRDC 301	Investigate incentive expectancy approaches to motivation by use of extensive questionnaires.		Incentive Expectancy Questionnaire	+++++	Four samples ranging from 60 to 150 cases (NPG people).	Indiv.	NO	Punched cards	NO
Boller, Bob	NPRDC 303	Manpower planning.		Manpower Planning Models		Bulk flow models; no individualization.				
DiGialleonardo, Frank (not interviewed)	NPRDC 303 Wash. Branch Office	Perform a systems analysis of the Navy's current manpower planning processes.	Considerable data collection (10-page questionnaire). A special methodology to do the systems analysis was necessary involving approximately 2,000 "communications," each 150 characters long in coded form.	MARRCS - TISA	150 individuals involved in the manpower planning process	Entire population of functions was surveyed. However, a sample of individuals responded with regard to their respective function.	Indiv. by function	YES	Disk	YES

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Lonsdale, Norm (Bob Boller also is interested in these data)	NPRDC 303	Manpower Systems. Development of LOS by pay grade matrices for each rating. Provides data for manpower modeling and input to the FAST model which projects personnel needs and training requirements for five years in advance.	There are 200 magnetic tapes at the Computer Sciences Dept. in Lonsdale's name. He maintains his own log of the status and whereabouts of these tapes since CSD does not.	LOS by Pay Grade		Data needed are extracted from the enlisted master file which is in sequence by SSN. Entire population since 1965.	Indiv. aggregated by rating		On dedicated disk packs at NELC (with mag tape back-up)	NO
Lonsdale, Norm	NPRDC 303	Manpower Systems. By request they will prepare a GCT breakdown by rating and pay grade.								
Silverman, Joe (not interviewed)	NPRDC 303	Manpower Modeling. Aggregate data from the change tapes to provide summary information on Navy personnel.	Due to computer system changes at BuPers, there have been no change tapes received for some time.	Matrix of Personnel Changes		This is not a source tape.	A matrix aggregated by type of change, rating, etc.			
Brock, John	NPRDC 306	Improvement of training and shipboard manpower utilization; job definition by task analysis.	The engineering and hull rating data are at NPRDC. The air ratings are at Millington, Tenn. The electronics and electrician ratings are either at the Naval Training Center in San Diego or in Memphis. One cannot find out where a job is being trained; a file of just this information alone is needed.	Job Definition by Task Analysis		Entire population.	Description of job duties for each rating		Mag tape	

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Smith, John	NPRDC 306	Training evaluation feedback system - proposed. Development of a Navy Instructor System for the 1980 Decade.		No data bases now.						
Braunstein, Claude	NPRDC 307	Retention of MD's and DDS's.		Attitudes of Medical Personnel Toward Retention in the Navy	4,000 MD's; 1,700 DDS's	Entire population (85% return).	Indiv.	NO	Mag tape	NO
Braunstein, Claude	NPRDC 307	Retention.	Individual identification is voluntary.	Survey of Navy Personnel Attitudes Toward Retention	20,000	Stratified sample by SVIB group; 10% officers and 5% enlisted; annually since 1964.	Indiv.	NO		NO
Braunstein, Claude	NPRDC 307	Attitudes toward bonus (of nuclear power personnel who have left the Navy).		Attitudes Toward Bonus of Nuclear Power Personnel	≈500	Entire population; all of those who left the Navy.	Indiv.	NO		NO
Broedling, Laurie	NPRDC 307	To relate tendency to return mail questionnaires to the characteristics and attitudes of naval personnel. Four separate questionnaires were used. Demographic information is available on all sample members. Questionnaire items cover a wide variety of topics.	One portion of data was lost during a move; however, remaining data are all usable.	Tendency To Respond to Mail Surveys	1,260 enlisted males	Disproportionately stratified by enlistment & race. Comprised of five subsamples. Approximately 860 individuals were sent each questionnaire.	Indiv.	YES	Mag tape	YES

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Broedling, Laurie	NPRDC 307	To determine the relationship of perceptions of Internal-External Control to the work motivation of naval personnel using an expectancy theory framework.		Relationship of Internal-External Control to Work Motivation	207 officers and enlisted personnel	All pay grades. General cross-section, but disproportionately weighted with officers.	Indiv.	YES	Mag tape	YES
Somer, E. P.	NPRDC 307	Study of changes in time of various attitudes and adjustment toward the Navy.	Problem in tracking sample members when they were transferred.	Prince Surveys		Longitudinal study on first-term enlisted personnel; mental categories I-III.	Indiv.		Mag tapes	
Somer, E. P.	NPRDC 307	Study of changes in time of various attitudes and adjustment toward the Navy.	Problem in tracking sample members when they were transferred.	Pauper Surveys		Longitudinal study on first-term enlisted personnel; mental category IV.	Indiv.		Mag tapes	
Somer, E. P., Attitude & Motivational Research Division, NPRDL, Washington, D.C.	NPRDC 307	Survey Data (Questionnaires)	No documentation exists to determine precisely how to read the mag files or what data fall in which fields.	An historical collection of 117 reels of magnetic tape containing the individual responses to survey questionnaires.						

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Stumpf, Sue	NPRDC 307	To determine the preferences of military personnel and their spouses for military vs. civilian housing and the relationship of satisfaction with housing to overall quality of military life.		Family Housing Preferences Survey	15,000 military personnel and their spouses (30,000 total)	About 1/3 each from Army, Air Force, & Navy/M.C. All pay grades. Cluster sample; within each installation, SAMS sampling technique was used.	Indiv.	YES	Mag tape	YES
Stumpf, Sue	NPRDC 307	To determine the preferences of military personnel for military vs. civilian housing and the implications of these preferences for the cost-effectiveness of building military housing.		1973 OSD Family Housing Survey	12,000 military personnel	About 1/3 each from Army, Air Force, & Navy/M.C. 17 installations surveyed; within each installation, SAMS sampling technique was used. All pay grades.	Indiv.	YES	Mag tape	Will be
Wilcove, Gerry (not interviewed)	NPRDC 307	Obtain questionnaire data on need and job satisfaction among junior enlisted men and junior officers as a basis for offering recommendations to Navy management.	Problem interfacing with computer support personnel who are not familiar with statistics. Problem finding first obliger officers in master tape file.	Questionnaire on Attractiveness of Navy Life and Policy Changes			Indiv.	YES	Mag tapes (6 reels)	Data dump
Crawford, Kent; Durning, Kathy	NPRDC 308	Organizational climate and peer relationships as a function of (1) demographic information, (2) enlistment an individual is in, (3) age, and (4) advancement rate.	Some items on the questionnaire are locally determined. Individuals are anonymous; if there is only one individual in a cell, that cell is omitted from the analysis.	Organizational Climate (part of the Navy Human Resources Management Survey analysis - see next entry)	75,000 individuals but summarized	Original data on individuals summarized by ship (50-60) X level (4) X department.	Indiv. but summarized	NO	Mag tapes	NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Thomas, Ed; Crawford, Kent; Malone, John; Dodson, Jan	NPRDC 308	Central repository for response data to the Navy Human Resource Management Survey. This survey serves as a vehicle to provide individual unit commanding officers with feedback data concerning internal lines of communication, race relations, drug and alcohol abuse, informal (vice formal) lines of organization, and overseas diplomacy. NPRDC is charged with the responsibility of providing a measure of the impact of the Human Resource Management Program.	1. Confidentiality of individual units must be maintained and yet criterion data relating to each individual unit must be merged with survey data. 2. Several forms of the survey are in existence. Therefore, strict input controls had to be devised. 3. Data tapes from 3-5 processing centers. The recording technique used by one center is not compatible with NPRDC facilities and must be translated "bit by bit."	Navy Human Resource Management Survey	35 to 5,000 per unit surveyed	(as of 6/30/75) 80,000 individual responses from 400 individual units consisting of approximately 200 variables (increasing at the rate of approximately 12,000 individuals and 40 units per month). Individual and aggregate unit data files are maintained.	Indiv.	YES (data only in special cases)	Disk with mag tape back-up	YES
Royle, Marjorie	NPRDC 309	Improvement of the performance evaluation and advancement system; removing leniency and other confounding factors in evaluation. Develop new performance evaluation forms and provide support to personnel selection boards.	Getting clean and complete performance data; then standardizing it to temper the effect of "gun decking" (giving all high marks).	E5 to E9 OCR Performance Evaluations	≈100,000 (E5 & E6); ≈50,000 (E7 to E9)	Entire population ideally, but some missing data because of difficulty in reading OCR forms. Received yearly since 1967. Current source - BuPers.	Indiv.		Mag tapes	YES

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Royle, Marjorie (Joe Silverman also uses these data but for different purposes)	NPRDC 309	Improvement of the performance evaluation and advancement system.		Enlisted Advancement Records		Entire population of those taking advancement exams. Data for Pay Grades E4-E6; some data for Pay Grades E7-E9.	Indiv.		Mag tape in CSD in two formats	YES
Royle, Marjorie (Norm Abrahams also is interested in this data base)	NPRDC 309	Variables include the SVIB, Edwards Personal Preference, a biographical index, Comrey Personality Schedule, Guilford Social Intelligence, and criterion data including a self-rating of performance.		Interpersonal Effectiveness Study	≈600	Enlisted men in face-to-face billets - a representative sample.	Indiv.		Mag tape	YES
Royle, Marjorie; Ward, Sam (Norm Abrahams also is interested in this data base)	NPRDC 309	Variables include the SVIB, Comrey Personality Schedule, part of the Guilford Social Intelligence test, a biographical index, and criterion data.		Career Counselors Study	≈600	Includes 1 to 8 paired counselor/counselee responses.	Indiv.		Mag tape in preliminary form	YES
Abrahams, Norm; Neumann, Idell (also mentioned by Suiter et al.)	NPRDC 310	Recruiting/ Retention		NROTC Master Tape File (OCARS will replace this file)		Entire population; entering data on NROTC students thru college to commissioning since 1964.	Indiv.		Mag tape	NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	Stored Now?	File Format Made Available?
Abrahams, Norm; Neumann, Idell (also mentioned by Suitar et al.)	NPRDC 310	Recruiting/ Retention.	Performance data field is blank on tape received from BuPers. Mag tapes received from the Naval Academy are hard to read.	Officer Master File		Entire population; longitudinal data.	Indiv.		Mag tape	YES
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention.		SVIB's on Naval Academy Students	7,500-8,000 annually	Naval Academy applicants and/or midshipmen.	Indiv.		Mag tape	NO
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention.		NVII Profiles on Enlisted Recruits	50,000-60,000	Entire population for 1969 and 1970.	Indiv.		Mag tape (<20 reels)	NO
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention.		Background Questionnaire (part of SVIB)		Collected since 1967; received from NEC.	Indiv.	NO	Mag tape	NO
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention.		Unobtrusive Bias Study	20,000 records	Enlisted classification records.	Indiv.		One mag tape	NO
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention.		Officer Candidate School Applicants	10,000-20,000 per year since 1968	SVIB response tape.	Indiv.			NO
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention.		SVIB Responses - Naval Enlisted Scientific Education Program (NESEP)	400		Indiv.			NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Abrahams, Norm; Neumann, Idell	NPRDC 310	Recruiting/ Retention		SVIB on Naval Academy Officers		1961-1964 Naval Academy graduates.	Indiv.			NO
Neffson, Nancy (Ed Thomas also is interested in this data base)	NPRDC 310	Prediction of performance from experimental test data on enlisted men and recruits in order to recommend best men for re-enlistment.	Poor cooperation from the supervisors in returning the questionnaire; only a 50% return.	Prediction of Performance			Indiv.	YES		NO
Rafacz, Bernard	NPRDC 310	Feasibility of granting 1st-term USMC recruits their job preference option. Feasibility was determined on the basis of whether there was a significant difference in job satisfaction and performance between those recruits who were granted their preferences and those who were not.	From an initial sample of 14,000 men, the sample size reduced to 7,542 and 2,480 men at the remaining two data collection periods. Samples were further reduced in size because not all men reported complete data.	TAPE 900490	14,000 originally, but markedly reduced on follow-up	Longitudinal study.	Indiv.		Mag tape	YES
Sands, William A.	NPRDC 310	Improved selection of Navy enlisted applicants.	Obtaining and organizing data collected by researchers at the Navy Health Research Center.	Original Odds for Effectiveness Data	11,008	Recruits from all entry conditions; four sampling periods from 1960-1961.	Indiv.		Mag tape	YES
Sands, William A.	NPRDC 310	Improved selection of Navy enlisted applicants.	Obtaining, organizing, and editing data collected by researchers at the Navy Health Research Center.	Original Odds for Effectivness Data - Normal Entry Standards	3,649	Recruits from normal entry conditions; four sampling periods from 1960-1961.	Indiv.		Mag tape	YES

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Sands, William A.	NPRDC 310	Improved selection of Navy enlisted applicants.	Editing data obtained from Navy Recruiting Command.	New Odds for Effectiveness Data on Accessions	165,376	Recruits entering Navy between 7/72 and 7/74.	Indiv.		Mag tape	YES
Sands, William A.	NPRDC 310	Improved selection of Navy enlisted applicants.	Obtaining data on attritions of all types.	New Odds for Effectiveness Data on Attritions		Recruits leaving service prematurely.	Indiv.			
Swanson, Len	NPRDC 310	Validation of Navy Basic Test Battery, Form 7 and Form 8, against a criterion of Class "A" school performance (final grades and pass-fail status code).	Incomplete data from many schools. Very little or no data for dropped students from some schools.	Class A School Criterion Data and BTB Scores	97,000-143,000; each sample covers a 1-2 year time period	Navy Class "A" school students in most schools.	Indiv.		Received on punch cards; then screened & put on mag tape	NO
Swanson, Len	NPRDC 310	Validation of Army Classification Battery (ACB-61) used by the Marine Corps for assigning enlisted personnel to schools.		Marine Class A School Criterion Data	5,700 at present; expect to increase to 20,000 by 3/75	All graduates and drops from A schools (1974 to Jan. 1975). Data from over 100 schools; 50 schools with N>100.	Indiv.; OMR school data form available		Mag tape	YES
Swanson, Len	NPRDC 310	Validation of the Armed Forces Vocational Aptitude Battery, Forms 2 and 5. The criterion is performance in Navy Class "A" schools.		Validation of the Armed Forces Vocational Aptitude Battery for the Navy	150 to 300 per school	22 Class "A" schools.	Indiv.		Data are being collected	NO

(Continued)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Swanson, Len; Thomas, Ed; Cory, Chuck	NPRDC 310	Minority group research with marginal personnel. Data collected: noncognitive, cognitive, motivation, and background information.		Phase 1 - Fall '67; Phase 2 - Spring '68; Phase 3 - '68/'69; Phase 4 - '69.	4,200-10,000	All entering recruits at the Naval Training Center, San Diego during specified time periods.	Indiv.			NO
Yellen, Ted (not interviewed)	NPRDC 310	Analyze SURVICE* results for feedback to commands and ICR teams. * Survey of In-Country Experiences	Changes to the questionnaire required redesign of the op scan form and also modifications to the data analysis program.	Overseas Homeporting	8,000	Aggregated response - administered to 8,000.	Indiv.	YES (being revised)	Mag tape	OP Scanner Record Layout
Harris, Bob	NPRDC 311	Human Engineering. Standardization of human factors as a basis for experimental studies. Human factors in performance data are culled from the literature.	Questionable quality.	Human Factors in Performance Data	N=100	Proposed pilot study - stratified sample.	List of human factors			

OTHER INFORMATION SYSTEMS AT NPRDC AS OPPOSED TO ATTITUDINAL RESEARCH DATA BASES

Person Interviewed	Code	Name and Description of Information System
Ramras, Eugene; Blanchard, Robert	NPRDC 02	*RDCS (Research and Development Coordination System).* NPRDC is charged with coordinating the Navy's personnel research and development activities as well as relating them to similar projects in the Army and the Air Force. The objective of the RDCS is to strengthen links, eliminate overlap, and/or fill in gaps.
Sjoholm, Al	NPRDC 201	*STAR (Status of Program Report).* A computer-based management information system reporting on the progress of NPRDC research projects.
Shumate, Chan	NPRDC 301	*CENDEX.* An inventory of the talents and interests of the NPRDC staff. Part of a larger effort to establish an intra-Center information-communication system to make more efficient use of in-house talents.
Magnusson, Paul	NPRDC 307	*CCOPS (Control and Coordination of Personnel Surveys).* A system to control and coordinate all personnel survey administration in the Navy in order to eliminate duplication of effort and to prevent over-surveying.

DATA BASES AT THE NAVY HEALTH RESEARCH CENTER OF USE IN PERSONNEL RESEARCH

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Gunderson, E.K.E.; Dean, Larry; Pugh, Bill	NHRC	To determine the influences of environment and organizational climate on individual performance and health; to ascertain the interrelationships among environmental factors, individual characteristics, and organizational effectiveness.	Several sources of data: questionnaires, ships' records, special forms, and BuPers and BuMed computer tapes. Merging these data proved to be difficult. After enormous effort, consistency was achieved.	Study of Habitability	→→→→→→	5,000 men on 20 ships plus 3 shore stations.	Indiv.+ some data reflecting ship, environment, division, and department characteristics	YES	Mag tape	
Gunderson, E.K.E.; Kolb, Douglas	NHRC	To investigate the variables predictive of alcohol and drug abuse and outcome (e.g., biographic data, Comrey scores, Cornell Medical Index, test scores, service records, health records, ARC-ARU patient record forms, narrative summaries, performance data, and discharge data).		Studies of Alcohol and Drug Abuse (several data bases)	1. N>5,000	1. Data from BuPers on patients seen at the Navy's Alcohol Centers and Units from 1972 to the present.	Indiv.		Mag tape	
					2. N=1,200	2. Naval and Marine Corps personnel undergoing rehabilitation at the Naval Drug Rehabilitation Centers at Miramar, Cal. and Jacksonville, Fla. from 1971 to 1973.	Indiv.		Mag tape	
					3. →→→→→	3. DARTS mag tapes prepared by System Development Corp. are now available at the NELC Computer Center.	Indiv.		Mag tape	
					4. N=500 currently; eventually will be a vast file	4. All alcohol and drug data being collected currently.	Indiv.		Mag tape	

(Continued)

DATA BASES AT THE NAVY HEALTH RESEARCH CENTER OF USE IN PERSONNEL RESEARCH (CONT.)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Gunderson, E.K.E.; Miller, Milan	NHRC	To maintain a medical inpatient data base on all Navy and Marine Corps personnel for longitudinal health research. Data include individual data cards at time of discharge from Navy medical facilities, Medical Board and Physical Evaluation Board actions, and death data.		Medical Inpatient System	470,000 patients	All records for all Navy and Marine Corps personnel from all hospitals. (July 1965 to present)	Indiv.		Mag tape	YES
Gunderson, E.K.E.; Miller, Milan	NHRC	To maintain a data base of summary service history for all naval enlisted personnel to be available to NHRC and NPRDC researchers.		Naval Enlisted Personnel Change File	N>30 million records; being reduced to 7 million individual cases	Each record consists of a 132-character abstract from losses and gains to the Naval enlisted population. This data base is being compiled into a single record for an individual summarizing his service history. (1965 to present)	Indiv.		Mag tape	YES
Gunderson, E.K.E.; Ryman, Dave	NHRC	To investigate the relationship between psychological and biochemical factors and performance during the rigorous training that underwater demolition teams undergo.		Underwater Demolition Teams Test File	N=500	Personnel undergoing training in underwater demolition.	Indiv.		Mag tape	

(Continued)

DATA BASES AT THE NAVY HEALTH RESEARCH CENTER OF USE IN PERSONNEL RESEARCH (CONT.)

Person Interviewed	Code	Mission/Project Objectives	Problems Encountered	Data Base Name	Sample Size	Sample Description	Unit of Data Collection	If Questionnaire, Made Available?	How Stored Now?	File Format Made Available?
Gunderson, E.K.E.; Ryman, Dave	NHRC	All military and civilian personnel wintering over in the Antarctica were tested & examined as applicants. Follow-up data after wintering over consisted of peer nominations, illness records, and supervisor ratings.		Antarctica Studies	4,000 applicants; 1,500 wintered over	All military and civilian personnel applying to winter over in the Antarctic since 1964.	Indiv.		Mag tape	
Gunderson, E.K.E.	NHRC	Family and social histories and drinking patterns of Navy career men were examined to determine the usefulness of these data in predicting alcoholism. Follow up by means of personnel and medical records will be conducted.		Re-enlistment Study	1,500	Navy enlisted men re-enlisting at two naval air stations and one naval base.	Indiv.		Mag tape	

RELATED INTERVIEWS

Person Interviewed	Navy Component	Description of Objectives	Description of Computer	Charge Structure	Miscellaneous
Bigbee, L. R.	MARDAC representative at the NPGS, Monterey, California.	MARDAC is a staff service to the Assistant Secretary of Defense for Manpower. MARDAC monitors outside researchers for ASDM. An ancillary purpose is to make data available to researchers.	The computing capability at NPGS is twin IBM 360/67's that work together as one, providing multiprocessing in a time-sharing environment.		MARDAC maintains an historical collection of 1800 reels of magnetic tape (binary packed data). They plan to bring one-third of this collection to Monterey (i.e., the cleaner data).
Harding, Frank; Goral, John; Lipowitz, Andrea	MARDAC, Survey Research Division, Alexandria, Virginia.	MARDAC is a staff service to the Assistant Secretary of Defense for Manpower. The Survey Research Division creates and maintains computer files of responses to military and civilian surveys, analyzes the results, and publishes research reports. They also respond to special inquiries.			The Survey Research Division of MARDAC maintains a Survey Data Bank. As of January 1975, 24 surveys were on file including 14 DoD-wide surveys, 2 surveys of civilian personnel, 2 Air Force surveys, and 7 Army surveys. When completed, the data bank will include all major military surveys (DoD-wide, civilian, and individual surveys) dating back to July 1970.
Small, Dana	Naval Electronics Laboratory Center, San Diego, California.		The NELC computer is an IBM 360/65 running under HASP OS Release 21.7. It has 7 tape drives, and uses 3330 disk packs. The system is running almost at capacity now.	The NELC charge structure for computer services makes a low charge for core, but penalizes users with high I/O utilization.	

(Continued)

RELATED INTERVIEWS

Person Interviewed	Navy Component	Description of Objectives	Description of Computer	Charge Structure	Miscellaneous
Messinger, Charles	Naval Undersea Center, San Diego, California.		The NUC computer is a UNIVAC 1110 running under the EXEC 8 operating system. It has seven 9-track tape drives (1600 bpi capacity) and two 7-track tape drives. It is disk limited for on-line operation. Currently the system is running at less than half capacity. They request 6-9 months lead time to add significant usage to the system, which could involve adding more disk.	The NUC Computer Center employs a cost recovery charge structure. A copy of the accounting algorithm has been obtained.	

DISTRIBUTION LIST

Chief of Naval Operations (OP-103B)
Chief of Naval Operations (OP-987P10)
Chief of Naval Personnel (Pers-10c)
Chief of Naval Personnel (Pers-6)
Chief of Naval Personnel (Pers-65)
Chief of Naval Material (MAT 0344)
Chief of Naval Material (MAT 035)
Chief of Naval Research (Code 450) (4)
Chief of Naval Research (Code 452) (3)
Chief of Naval Research (Code 458)
Chief of Naval Education and Training (N-5)
Chief of Naval Technical Training (Code 016)
Commander Training Command, U.S. Atlantic Fleet (Code N3A)
Naval Electronics Laboratory Center
Fleet Combat Direction Systems Training Center, Pacific (Code 03A)
Naval Education and Training Program Development Center
Naval Development and Training Center (Code 0120)
Naval Training Equipment Center (Technical Library)
Navy Recruiting Command (Code 20)
Naval Aerospace Medical Research Laboratory (Code L5)
Naval Research Laboratory (Technical Information Division) (6)
Naval Research Laboratory (Code 2029) (6)
Office of Naval Research Branch Office, Boston
Office of Naval Research Branch Office, Chicago
Navy Health Research Center
Human Resource Management Center, London
Human Resource Management Center, Washington
Human Resource Management Center, Norfolk
Human Resource Management Center, San Diego
Human Resource Management Center, Pearl Harbor
Human Resource Management Center Attachment
Human Resource Management School, Memphis
Navy Occupational Task Analysis Program (NOTAP)
Naval Postgraduate School (Code 2124)
Commandant of the Marine Corps (Code MPI-20)
Office of the Deputy Chief of Staff for Personnel (DAPE-PBR), U.S. Army
Army Research Institute (Library), Rosslyn, Va.
Headquarters, AFSC, Andrews AFB (Environmental and Life Sciences Division)
AFOSR (NL), Arlington
Personnel Research Division, Air Force Human Resources Laboratory (AFSC), Lackland AFB
Occupational and Manpower Research Division, Air Force Human Resources Laboratory (AFSC), Lackland AFB
Technical Library, Air Force Human Resources Laboratory (AFSC), Lackland AFB
Technical Training Division, Air Force Human Resources Laboratory, Lowry AFB
Advanced Systems Division, Air Force Human Resources Laboratory, Wright-Patterson AFB
Flying Training Division, Air Force Human Resources Laboratory, Williams AFB
Manpower Research and Data Analysis Center (MARDAC)
Science and Technology Division, Library of Congress
Defense Documentation Center (12)

ONR INFORMAL DISTRIBUTION LIST

NAVY

1 Dr. Eugene E. Gloye
ONR Branch Office
1030 E. Green Street
Pasadena, CA 91106

1 Dr. H. Wallace Sinaiko
c/o Office of Naval Research
Code 450
800 N. Quincy Street
Arlington, VA 22217

1 Cdr. Anthony C. Cajka, USN
Department of the Navy
Human Resource Management Center
Washington, DC 20370

1 Dr. E. K. E. Gunderson
Code 8030
Navy Medical Neuropsychiatric
Research Unit
Navy Health Research Center
San Diego, CA 92152

1 Cdr. Paul D. Nelson, MSC, USN
Head, Human Performance Division
Code 44
Navy Medical R&D Command
Bethesda, MD 20014

1 Dr. Donald P. Gaver
Naval Postgraduate School-OR/AS
Monterey, CA 93940

1 Dr. Kneale Marshall
Naval Postgraduate School
Monterey, CA 93940

1 Dr. C. Brooklyn Derr
Associate Professor, Code 55
Naval Postgraduate School
Monterey, CA 93940

ARMY

1 Dr. Ralph Canter
Army Research Institute
Commonwealth Bldg.
1300 Wilson Blvd.
Rosslyn, VA 22209

1 Dr. David Segal
Army Research Institute
Commonwealth Bldg.
1300 Wilson Blvd.
Rosslyn, VA 22209

1 Dr. Arthur Gilbert
Army Research Institute
Commonwealth Bldg.
1300 Wilson Blvd.
Rosslyn, VA 22209

1 Major Frederick Trone
DAPC 844C, Hoffman Building
Alexandria, VA 22332

AIR FORCE

1 Mr. Robert Stephens
AFAC
Room 5C162, Pentagon
Washington, DC 20350

1 Dr. Robert A. Zawacki
Assistant Professor of
Behavioral Science
6457B
United States Air Force Academy
USAFA, CO 80840

MARINE CORPS

1 Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, DC 20380

1 Mr. E. A. Dover
Manpower Measurement Unit
Code MPI
Arlington Annex, Room 2413
Arlington, VA 20380

OTHER DOD

1 Dr. Frank Harding
Manpower Research and Data Analysis Center (MARDAC)
300 N. Washington Street
Alexandria, VA 22314

1 Mr. Lawrence Bigbee
Manpower Research and Data Analysis Center (MARDAC)
550 Camino El Estero
Monterey, CA 93940

1 Col. Austin W. Kibler
Human Resources Research Office
ARPA
1400 Wilson Blvd.
Arlington, VA 22209

OTHER

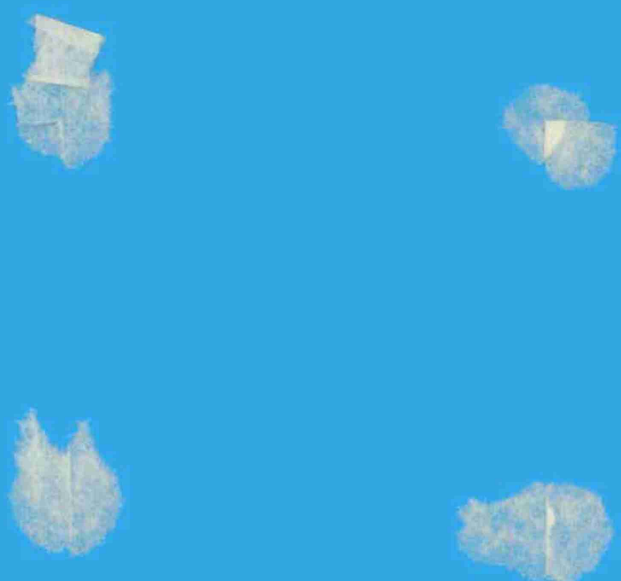
1 Dr. David G. Bowers
Institute for Social Research
University of Michigan
Ann Arbor, MI 48106

1 Dr. John J. Collins
9521 Cable Drive
Kensington, MD 20795

1 Mr. Joel Ellermeier
Bureau of Training, CSC
Room 7626
1900 E St., N.W.
Washington, DC 20415

1 Dr. Gloria L. Grace
System Development Corporation
2500 Colorado Avenue
Santa Monica, CA 90406

1 HumRRO (ATTN: Library)
300 N. Washington Street
Alexandria, VA 22314



NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER
SAN DIEGO, CALIFORNIA 92152

Naval Postgraduate School
Monterey, CA 93940
ATTN: Library (Code 2124)