AD630089

INFORMATION STORAGE AND RETRIEVAL
A STATE-OF-THE-ART REPORT

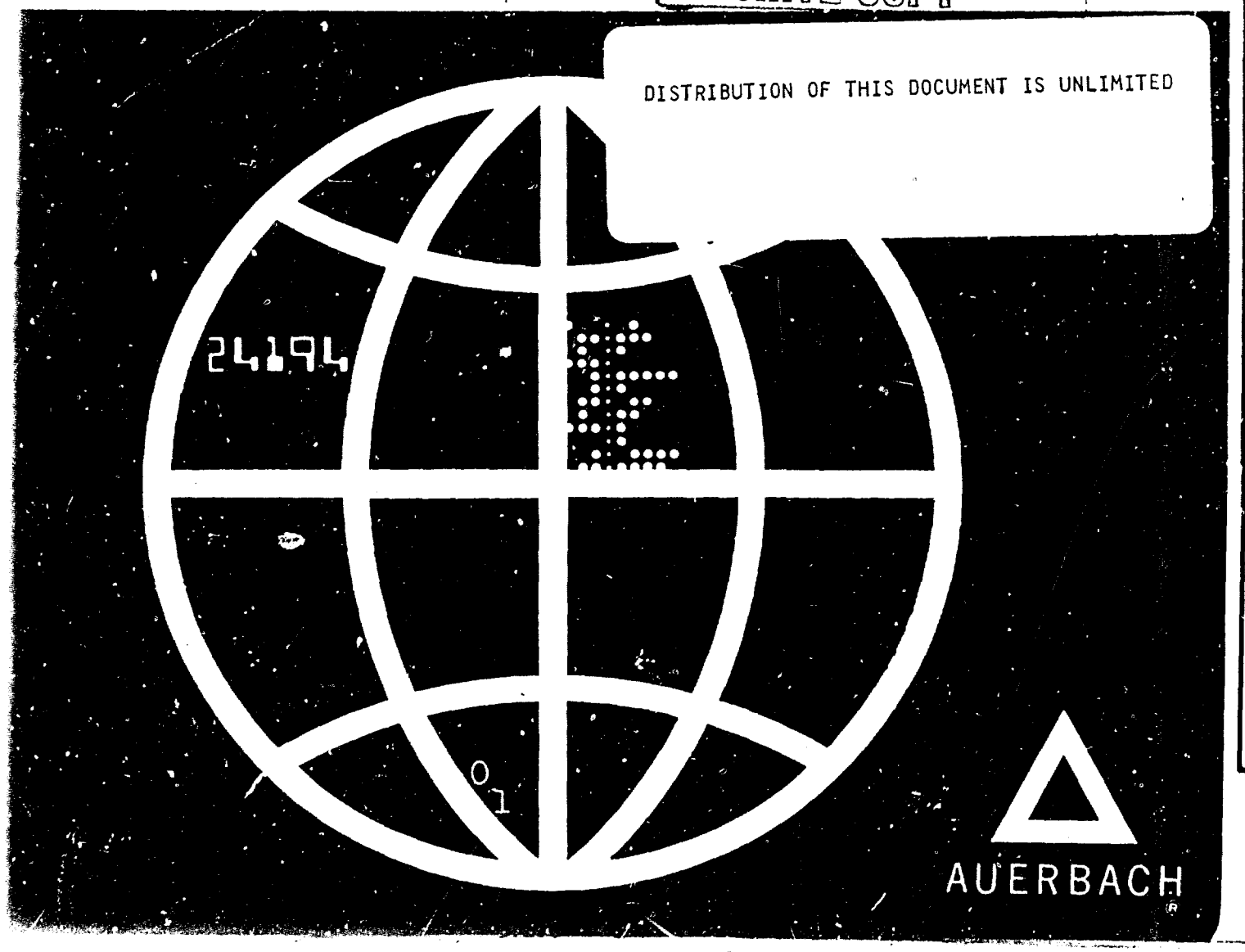
Lawrence Berul
Principal Investigator

Code 1

CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND
TECHNICAL INFORMATION
$6.00 1.50 249 pgs
ARCHIVE COPY COPY

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

24194



INFORMATION STORAGE AND RETRIEVAL
A STATE-OF-THE-ART REPORT

Lawrence Berul
Principal Investigator

September 14, 1964

PR 7500-145

AUERBACH Corporation • Philadelphia 3, Pennsylvania

FOREWORD

The problems of information storage and retrieval have received extensive discussion in the past decade, but few "solutions" to the problems have survived the cruel test of real world application. Accordingly, the field continues to engender strong controversies on matters ranging from problem definition and user need to the best techniques for using computers. In such a climate, it is rare that one finds a dispassionate, factual, and comprehensive state-of-the-art report.

This volume provides what I consider to be one of the more useful state-of-the-art reports now available, even though it brings the field up to mid-1964.

The report was prepared by Lawrence Berul and his associates at the AUERBACH Corporation for a commercial customer. Because the report contains extensive data on government activities, my office became aware of its preparation.

Several months ago I asked the AUERBACH Corporation and its client if there was a possibility that it might be made available to selected government personnel on a restricted basis while the information and data it contains are still current.

In a most generous action, the report has been made freely available to the government, and it seems desirable that it also be made promptly available through the channels normally used for public dissemination of technical reports prepared for the Department of Defense.

I wish to express a deep appreciation to the AUERBACH Corporation and its client for waiving their copyright to this report in order that it may receive prompt and widespread distribution. While my office and the Department of Defense take no responsibility for the data it presents and the conclusions it draws, I feel that this state-of-the art report does deserve to be read by anyone who is seriously interested in gaining a better grasp of the field of information storage and retrieval.

Walter M. Carlson
Director of Technical Information
Department of Defense

AUERBACH®

ACKNOWLEDGMENTS

This state-of-the-art report was compiled by synthesizing the background, experience, and insight of a number of AUERBACH staff members. The primary AUERBACH contributors, besides the principal investigator, included Mr. Eugene Wall, Mr. Murray Dodge, and Dr. Jack Minker, who each prepared portions of this report and, in addition, provided a critical review of the entire text. Acknowledgment is also made of the advice and critical suggestions made by Mr. Paul Kerstetter, Mr. Roger Sisson, and Dr. Harold Wooster.

The cooperation and assistance of the various organizations described in this report are also gratefully acknowledged. Among those who provided particularly helpful assistance are Mr. Lawrence I. Chasin, of the General Electric Company, Missile and Space Division Library; Messrs. Howard G. Alloway, Van Wente, and Irving Lebow of the National Aeronautics and Space Administration, Scientific and Technical Information Division; Mr. Parker Dagget of the Naval Air Technical Services Facility; Mr. Gregory Abdian of the Defense Documentation Center; and Mr. Walter Carlson of the Office of the Director, Defense Research and Engineering, Department of Defense.

A warm acknowledgment is also due the Publications department of the AUERBACH Corporation, which provided outstanding assistance in technical editing, artwork, composition, and printing. Mr. George Neborak was the chief technical writer, responsible for the final editing and coordination of the composition and assembly of this report.

AUERBACH®

TABLE OF CONTENTS

Paragraph	Title	Page
	SECTION I. INTRODUCTION	
1.1	BACKGROUND	1-1
1.2	OBJECTIVES	1-2
1.3	OUTLINE OF REPORT	1-2
	SECTION II. INFORMATION STORAGE AND RETRIEVAL — A COMMUNICATIONS PROBLEM	
2.1	METHODS OF PERSON-TO-PERSON COMMUNICATION	2-1
2.2	COMMUNICATIONS CONTINUUM	2-2
2.2.1	Feedback	2-2
2.2.2	Abstractness	2-2
2.2.3	Zone of Retrievability	2-4
2.2.4	Information Retrieval - Data Retrieval	2-4
2.2.4.1	Document Retrieval	2-4
2.2.4.2	Fact Retrieval	2-4
2.2.5	Graphic vs. Digital Information	2-5
2.3	CLASSIFICATION OF INFORMATION SYSTEMS	2-5
2.3.1	Current-Awareness and Retrospective Services (Mode of Use)	2-7
2.3.2	Use of Information	2-7
2.3.3	Categories of Use	2-8
2.3.4	Performance Characteristics of Information Systems	2-8
2.3.5	Forms of Information	2-10
2.3.5.1	Individual Items	2-11
2.3.5.2	Reference Tools	2-12
2.3.5.3	Correlations	2-12
	SECTION III. INFORMATION STORAGE AND RETRIEVAL CONCEPTS AND TECHNIQUES	
3.1	TRADITIONAL CLASSIFICATION AND INDEXING	3-1
3.2	COORDINATE INDEXING	3-3
3.2.1	History of Coordinate Indexing	3-4
3.2.2	The Thesaurus Approach	3-9
3.2.2.1	Listing of Vocabulary Terms	3-9
3.2.2.2	Exhibiting Relationships Among Terms	3-10
3.2.2.3	Defining of Terms	3-11
3.2.2.4	Practicability of Thesauri	3-11

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE
3.2.3	Syntactical Problems	3-12
3.2.3.1	Linking	3-12
3.2.3.2	Role Indicators	3-13
3.3	AUTOMATIC INDEXING AND ABSTRACTING	3-14
3.3.1	Automatic Indexing	3-14
3.3.1.1	KWIC Indexes	3-14
3.3.1.2	Automatic Indexing "In Depth"	3-15
3.3.2	Automatic Abstracting	3-17
	SECTION IV. INFORMATION PRODUCTS AND SERVICES	
4.1	CURRENT-AWARENESS SERVICES	4-1
4.1.1	Journals	4-1
4.1.2	Abstract Journals	4-6
4.1.3	Contents Journals	4-7
4.1.4	Key-Word-in-Context Indexes	4-7
4.1.5	Initial Dissemination Schemes	4-9
4.1.5.1	Preprint and Report Dissemination	4-9
4.1.5.2	Microform Dissemination	4-9
4.1.5.3	Selective Dissemination of Information (SDI)	4-12
4.2	RETROSPECTIVE SEARCH SERVICES	4-14
4.2.1	Book Form Indexes	4-15
4.2.1.1	Subject Heading Index	4-16
4.2.1.2	Uniterm Index	4-16
4.2.1.3	Tabledex Index	4-16
4.2.1.4	Citation Index	4-16
4.2.2	Card Form Indexes	4-20
4.2.2.1	Edge-Notched Cards	4-20
4.2.2.2	Interior-Punched Cards	4-21
4.2.3	Machine Search Services	4-21
4.2.3.1	Search to Retrieve Document Number	4-22
4.2.3.2	Search Producing Citations or Abstracts	4-22
4.2.3.3	Search Producing Document Display	4-23
4.2.3.4	Search to Provide Microform or Hard Copy	4-23
4.2.3.5	Search to Produce Data Correlations	4-23

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE
	SECTION V. INFORMATION STORAGE AND RETRIEVAL SYSTEM FUNCTIONS	
5.1	GENERAL	5-1
5.2	BASIC FUNCTIONS	5-1
5.3	BASIC SYSTEMS	5-2
5.4	INTERACTIONS AMONG FUNCTIONS IN SYSTEMS	5-3
5.5	DESCRIPTION OF BASIC INFORMATION STORAGE AND RETRIEVAL SYSTEMS	5-5
5.5.1	Origination	5-5
5.5.2	Acquisition	5-6
5.5.3	Surrogation	5-7
5.5.3.1	Cataloging	5-7
5.5.3.2	Abstracting	5-7
5.5.3.3	Indexing	5-8
5.5.4	Announcement	5-9
5.5.5	Index Operation	5-10
5.5.6	Document Management	5-10
5.5.6.1	Document Retrieval	5-11
5.5.6.2	Document Dissemination	5-12
5.5.7	End-Use	5-12
	SECTION VI. TYPICAL APPLICATIONS	
6.1	GENERAL	6-1
6.2	MISSION-ORIENTED INFORMATION CENTER (NASA)	6-3
6.2.1	NASA — Scientific and Technical Information Facility (Hereinafter Called the Facility)	6-3
6.2.2	Inputs and Outputs	6-4
6.2.3	Announcement and Dissemination Process	6-5
6.2.3.1	Document Dissemination	6-5
6.2.3.2	Announcement	6-7
6.2.3.3	Index Generation, Storage, and Dissemination	6-9
6.2.4	Request Processing	6-9
6.2.4.1	Requests for Copies	6-9
6.2.4.2	Requests for Bibliographies	6-10
6.2.5	Selective Dissemination	6-10

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE
6.3	SATELLITE INFORMATION CENTER	6-11
6.3.1	Inputs and Outputs	6-11
6.3.2	Storage and Announcement Process	6-14
6.3.3	Request Processing	6-17
6.3.4	Recapitulation and Analysis	6-18
6.3.4.1	Duplication of Descriptive Cataloging, Indexing, and Abstracting	6-18
6.3.4.2	Equipment and Compatibility Problems	6-19
6.3.4.3	Duplication of Camera Microfilming	6-19
6.4	TRAFFIC ROUTING CENTER (Army Chemical Information and Data System)	6-19
6.4.1	General	6-19
6.4.2	Army Chemical Information and Data System	6-20
6.4.2.1	CIDS Traffic Routing Center	6-23
6.4.2.2	User-CIDS Interface	6-24
6.4.3	Recapitulation and Analysis	6-24
6.5	ENGINEERING DATA CENTER	6-26
6.5.1	User Community	6-27
6.5.2	Inputs and Outputs	6-27
6.5.3	Files	6-29
6.5.4	Input Processing	6-30
6.5.5	Request Processing	6-30
6.5.5.1	Request for a Line Item	6-30
6.5.5.2	Request for a Set of Drawings	6-31
6.5.5.3	Request for Bid Set	6-31
6.5.6	Recapitulation	6-31
6.6	REAL ESTATE TITLE SEARCHING	6-33
6.6.1	General	6-33
6.6.2	Typical Search in a Recorder of Deeds Office	6-33
6.6.3	Typical Title Company System	6-35
6.6.3.1	Input Processing	6-35
6.6.3.2	Output Processing	6-37
6.6.4	A Punched Card System for Improved Grantor-Grantee Indexes	6-38
6.6.5	Computer System for Real Estate Tax Searching	6-40
6.6.6	Recapitulation and Analysis	6-40

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE
	SECTION VII. ECONOMIC ANALYSIS	
7.1	INFORMATION CENTER OPERATING COSTS	7-1
7.1.1	Input Processing and Announcement	7-4
7.1.2	Output-Request Processing Costs	7-5
7.2	RELATIVE OVERALL COSTS OF VARIOUS SYSTEM FUNCTIONS	7-5
7.3	FACTORS AFFECTING UNIT COST	7-5
7.4	ANALYSIS OF COSTS OF RETROSPECTIVE SEARCH	7-7
7.4.1	Comparison of Several Retrieval Methods	7-7
7.4.1.1	Sample Problem Definition	7-9
7.4.1.2	Roll Microfilm	7-11
7.4.1.3	Magnetic Tape	7-12
7.4.1.4	Magnetic Discs	7-13
7.4.1.5	Magnetic Cards	7-13
7.4.1.6	Conclusions	7-14
7.5	COPY FULFILLMENT COSTS	7-15
7.5.1	Effect of Number of Pages Per Document	7-15
7.5.2	Effect of Number of Copies Disseminated on Total Cost	7-17
7.5.3	Unitized Microforms	7-17
7.5.4	Request Copy Fulfillment	7-20
7.5.4.1	Pre-Stocking By an Overrun on Initial Dissemination	7-21
7.5.4.2	Reprinting Policy as A Function of Document Age or of Demand History	7-23
7.5.5	Lower Cost Printing	7-24
7.5.6	Lower Costs for On-Demand Copying	7-24
	SECTION VIII. HARDWARE CONSIDERATIONS	
8.1	IS&R SYSTEM FUNCTIONS AND ASSOCIATED EQUIPMENT	8-1
8.1.1	Origination	8-1
8.1.2	Acquisition	8-3
8.1.3	Surrogation	8-5
8.1.4	Announcement	8-5
8.1.5	Index Operation	8-9
8.1.5.1	Card Catalog Files	8-10
8.1.5.2	Edge-Notched and Interior-Notched Cards	8-10
8.1.5.3	Miscellaneous Devices	8-11
8.1.5.4	Computerized Indexes	8-11

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE
8.1.6	Document Management	8-11
8.1.6.1	Document Dissemination	8-12
8.1.6.2	Document Storage and Retrieval	8-12
8.1.6.3	On-Demand Copying	8-13
8.1.7	Correlations	8-13
8.2	USER FUNCTIONS AND ASSOCIATED EQUIPMENT	8-13
8.2.1	User Function	8-15
8.2.2	User Requirements	8-15
8.2.3	Query-Response Equipment	8-17
8.2.3.1	Look-up and Look-at	8-17
8.2.3.2	Take-Away	8-17
8.3	RELATIONSHIP BETWEEN INDEX AND DOCUMENT FILES	8-18
8.3.1	General	8-18
8.3.2	Early Search Systems	8-19
8.3.3	Computer Search Systems	8-20
8.4	FILE STRUCTURE	8-21
8.4.1	Early File Structures	8-21
8.4.2	File Structures on Tape-Oriented Computers	8-22
8.4.3	Random Access File Structure	8-23

SECTION IX. INFORMATION STORAGE AND RETRIEVAL SOFTWARE

9.1	SCOPE	9-1
9.2	PERMUTED TITLE INDEX PROGRAMS	9-2
9.2.1	IBM 1620 KWIC Index Program	9-2
9.2.2	IBM 1401 - 1410 KWIC Index Program	9-4
9.2.3	GE-225 KWIC Index Program	9-4
9.3	SEARCH PROGRAMS	9-4
9.3.1	General	9-4
9.3.2	Electronic Data Processing Search Programs	9-7
9.3.3	IS&R Programs Which Search Data	9-7
9.3.4	IS&R Programs Which Search Linear Reference Files	9-8
9.3.5	IS&R Programs Which Search Inverted Document Surrogate Files on Magnetic Tape	9-9
9.3.6	RCA RECOL (REtrieval COmmand Language)	9-9
9.3.7	Information Processing System (IBM IPS)	9-11

PARAGRAPH	TITLE	PAGE
9.3.8	Documentation Inc. Linear File Search System	9-12
9.3.9	GE-225 "Text Searching" System	9-13
9.3.10	IBM 1401 Inverted Card File Retrieval System	9-15
9.3.11	University of Pittsburgh Legal Retrieval System	9-16
9.4	SELECTIVE DISSEMINATION PROGRAMS	9-18
9.4.1	General	9-18
9.4.2	IBM 1401 Selective Dissemination of Information (SDI-3)	9-18
9.5	AUTOMATIC INDEXING AND ABSTRACTING	9-20
9.6	FILE MAINTENANCE PROGRAMS	9-20
9.7	EXECUTIVE PROGRAMS	9-20
	SECTION X. EVALUATION OF THE STATE OF THE ART AND PREDICTION OF TRENDS	
10.1	SCOPE	10-1
10.2	ORIGINATION	10-1
10.2.1	Evaluation	10-1
10.2.2	Predicted Trends	10-2
10.2.2.1	Publication of Separates	10-2
10.2.2.2	Cooperative Typesetting of Serials	10-2
10.2.2.3	Secondary Distribution of Separate Literature	10-3
10.2.2.4	Information "Packages"	10-3
10.3	ACQUISITION	10-3
10.3.1	Evaluation	10-3
10.3.2	Predicted Trends	10-4
10.4	SURROGATION	10-4
10.4.1	Evaluation	10-4
10.4.1.1	Cataloging	10-4
10.4.1.2	Abstracting	10-5
10.4.1.3	Indexing	10-5
10.4.2	Predicted Trends	10-7
10.4.2.1	Cataloging	10-7
10.4.2.2	Abstracting	10-7
10.4.2.3	Indexing	10-8
10.5	ANNOUNCEMENT	10-8
10.5.1	Evaluation	10-9

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE
10.5.1.1	Announcement Journals	10-9
10.5.1.2	Selective Dissemination of Information	10-9
10.5.2	Predicted Trends	10-9
10.5.2.1	Contents Journals	10-9
10.5.2.2	KWIC Indexes	10-10
10.5.2.3	Selective Dissemination Information	10-10
10.5.2.4	Automatic Composing	10-10
10.6	INDEX OPERATION	10-10
10.6.1	Evaluation	10-10
10.6.1.1	Separation of Index and Document Files	10-10
10.6.1.2	Large Index Files	10-11
10.6.1.3	Small Index Files	10-11
10.6.1.4	The Intellectual Aspect of Searching	10-12
10.6.1.5	Performance Measurement	10-12
10.6.2	Predicted Trends	10-12
10.6.2.1	Automatic Document Retrieval	10-12
10.6.2.2	Computer Index Searching Systems	10-12
10.7	DOCUMENT MANAGEMENT	10-13
10.7.1	Evaluation	10-13
10.7.1.1	Document Dissemination	10-13
10.7.1.2	Document Retrieval	10-13
10.7.1.3	Pre-Stock vs. On-Demand	10-14
10.7.2	Predicted Trends	10-15
10.7.2.1	Microfiche	10-15
10.7.2.2	Pre-Stocking	10-15
10.7.2.3	Satellite Collections	10-15
10.7.2.4	On-Demand Copying	10-15
10.7.2.5	Microfilm Standards	10-15
10.7.2.6	Microfilm Acceptance	10-15
10.8	CORRELATIONS	10-16
10.8.1	Evaluation	10-16
10.8.1.1	Small Fact Retrieval Systems	10-16
10.8.1.2	Large Fact Retrieval Systems	10-17
10.8.2	Predicted Trends	10-18
10.8.2.1	Fact Retrieval Systems	10-18
10.8.2.2	Real-Time Systems	10-18

TABLE OF CONTENTS (CONTD.)

PARAGRAPH	TITLE	PAGE

BIBLIOGRAPHY

APPENDIX A. SUPPLEMENTAL BIBLIOGRAPHY

APPENDIX B. RANDOM ACCESS FILE STRUCTURES

APPENDIX C. GLOSSARY

LIST OF ILLUSTRATIONS

FIGURE	TITLE	PAGE
2-1	The Communicative Continuum	2-3
3-1	Venn Diagrams	3-5
3-2	Termatrex Cards	3-6
3-3	Printed Uniterm Cards	3-8
4-1	Growth of Journals and Abstract Journals	4-5
4-2	Montage of Samples Showing Range of Typographic Quality	4-8
4-3	Key Word In Context Indexes	4-10
4-4	Typical Microforms	4-11
4-5	Selective Dissemination of Information	4-13
4-6	Tabledex Method	4-17
4-7	Citation Index	4-18
5-1	Possible Combinations of IS&R System Functions	5-4
6-1	Mission-Oriented Information Center (NASA), Input-Output Diagram	6-6
6-2	Mission-Oriented Information Center (NASA), Announcement and Dissemination	6-8
6-3	Mission-Oriented Information Center (NASA), Request Processing	6-12
6-4	Satellite Information Center	6-13
6-5	Satellite Information Center Processing Cycle (GE-MSD Library)	6-16
6-6	Chemical Information and Data System, Input-Output Diagram	6-21
6-7	A Network of Technical Information Centers	6-22
6-8	User - TIC - Traffic Routing Center Interfaces	6-25
6-9	Input-Output Diagram Engineering Data Center	6-28
6-10	Engineering Data Center, Input and Request Processing	6-32
6-11	Title Search In Typical County Recorders Office	6-34
6-12	Typical Title Co. System	6-36
6-13	Punched Card System for Grantee-Grantor Indexes, Registry of Deeds, Norfolk County, Mass.	6-39
6-14	Computer System for Real Estate Tax Searching	6-41
7-1	Effect of Number of Copies on Total Cost of Initial Dissemination for 10,000 - 50 Page Documents	7-19
7-2	On-Demand Copying Costs Vs. Pre-Stocking for Various Numbers of Copies	7-22
8-1	Hypothesized Computer Based Publishing Production Facility	8-4
8-2	Hypothesized Non-Computer Based Publishing Production Facility	8-6
8-3	Sequential Card Composition	8-8
9-1	IBM 1401 KWIC Index System	9-5
9-2	Classification of IS&R Search Programs	9-6
9-3	Flow Chart of University of Pittsburgh Legal Search System	9-17

SECTION I. INTRODUCTION

1.1 BACKGROUND

Information storage and retrieval (IS&R) is a relatively new field which is developing to improve the communication of information from the originators to the end users. It is providing new tools and techniques for communicating information which supplement the more traditional communication methods represented by the professional journal and the abstract journal. Since the author of a technical paper generally does not know all of the people who may be interested in it, the paper or message is in effect broadcast by printing thousands of copies in the hope of reaching the right users. Two factors are presently at work to hinder this traditional method of non-directed communication. The first is the increased volume of technical information generated; the second is the increased number of potential readers. As a result, the channels of communication are becoming clogged with information relevant to only a small percentage of the audience.

The antithesis of the broadcast method of communication is the direct method wherein the user is in direct communication with the originator and a dialogue or conversation can take place between them. Direct communication, however, requires that the originator know the end user or vice versa in order that a direct connection or channel can be established.

IS&R is effectively attempting to create a bridge between originators and users of information. If user needs or interests can be defined and the information content of messages described, information can be automatically switched or routed to the end user, even though the originators and users do not know each other. The role of IS&R, therefore, is to get the right information to the right person in the right form at the right time. This will thereby reduce the degree of redundancy resulting from the reliance on the broadcast of messages as the primary method of communication. It will also decrease the "noise level" or degree of intake of non-relevant information resulting from this method of communications. The net effect should be an improvement in the productivity of professional or technical people, who are heavily dependent on external sources of information in their work.

1.2 OBJECTIVES

This report was prepared for a major manufacturer of office equipment, interested in penetrating the market for IS&R equipment and systems. The objective of the study was to compile relevant background and interpretive material using the information resources of the AUERBACH Corporation and the experience of its technical staff, and prepare a state-of-the-art report which would put the developments in the IS&R field into perspective and serve to orient a newcomer into this field. The study specifically avoided the cataloging of special purpose IS&R equipment. The approach rather was to define the functions of an IS&R system and indicate how and what general type of equipment could be used in each particular function. The study demonstrates that a typical IS&R system is a seemingly unrelated set of equipment and procedures. The key to successful implementation is a well defined set of objectives and an effective system design which ties the varied equipment and procedures into a unified system.

1.3 OUTLINE OF REPORT

This section presents a brief outline or "road map" to the report. Many sections contain information which may appear redundant. To an extent, this was unavoidable as each section of the report is intended to be relatively self-contained with respect to its purpose, but also each section views the field from a different vantage.

Section II describes IS&R as a problem in person-to-person communications. It presents the overall scope of the problem, basic definitions, and some various methods for categorizing information systems.

Section III presents basic IS&R concepts and techniques. It traces the history of traditional librarianship through early IS&R systems and up to the present state-of-the-art. It discusses the major problems in the field which include semantics, syntactics, viewpoint, indexing, classification, file organization, vocabulary control, and automatic indexing and abstracting.

Section IV describes the various information products and services which are being employed to aid the process of communication between the originators and users of information. Both current-awareness and retrospective-type services are discussed. These information products and services represent the output of information systems; each product or service is described in terms of its purpose, i.e., the particular user requirement it is intended to fulfill and its possible forms.

Section V categorizes the various information system operations required to produce the products and services, described in Section IV, into eight basic system functions from which all IS&R systems can be assembled. These basic functions are: origination, acquisition, surrogation, announcement, index operation, document management, correlation, and end-use. A description of these eight basic functions and their role in an IS&R system are also presented.

Section VI describes five examples of IS&R applications or systems, for the purpose of showing how the various functions described in Section V are combined to provide a wide variety of products and services. The particular applications described are the NASA Scientific and Technical Information Facility, the General Electric Company Missile and Space Division technical library, which is a "satellite" of the NASA Facility, the Army Chemical Information and Data System concept (Traffic Routing Center), an Engineering Data Center (The Engineering Data Management Department of the Naval Air Technical Services Facility) and several examples of real-estate title searching systems.

Section VII describes the economics of information center operations for two major government information centers and presents some comparative economic analyses of alternate approaches to (1) the index operations function and (2) the document management function.

Section VIII describes the hardware implications of the various IS&R system functions. It is not intended to be a catalog or evaluation of specific hardware but rather presents a description of the functional requirements and the general type of hardware which has been or can be used to meet these functional requirements.

In addition, a discussion of user functions and their equipment implications is given. This is followed by an analysis of the relationship between index and document files, which helps to explain the reason for the lack of success of those equipment systems which have attempted to combine index and document files. A description of file structures is given for inverted files and linear files as employed on serial searching devices and on random access storage devices. This description is supplemented by an appendix on random access file structures (Appendix B).

Section IX presents a description of computer software (programs) which are available for IS&R functions. Special-purpose devices for searching index files and document files have not been particularly successful. Special-purpose computer programs used with general-purpose computers have been relatively more successful than the special-purpose equipment. Consequently, this section describes the type of functions which can be performed by computer and describes the programs which are generally available for these purposes.

Section X presents a summary technical evaluation of the state of the art and a prediction of trends within each of the eight basic system functions which make up all IS&R systems.

Reference numerals throughout the text refer to a bibliography which immediately follows Section X. A supplemental bibliography on the subject of Automatic Indexing and Automatic Abstracting is included as Appendix A. Appendix B presents a detailed discussion of random access file structures which was considered important enough for inclusion in this report. A glossary of IS&R terminology is included in Appendix C. This glossary has been compiled from numerous existing glossaries in the field. However, it has been modified considerably to eliminate much of the ambiguity which is present in the IS&R terminology.

SECTION II. INFORMATION STORAGE AND RETRIEVAL – A COMMUNICATIONS PROBLEM

The relatively new field of information storage and retrieval is concerned with improving the communication of recorded information among three types of individuals or groups:

(1) The originator of information -- one who develops or first records information.

(2) The indexer or librarian -- one who is responsible for the proper storage of information so that it can later be retrieved and used.

(3) The user of information -- one who requires information as a resource in solving his problems.

2.1 METHODS OF PERSON-TO-PERSON COMMUNICATION

Most of the information received by individuals in their daily lives, and necessary to the conduct thereof, is probably obtained via the sense of hearing, largely because of the omnidirectional nature of this sense. As a result, most of us seem automatically to prefer to receive information "by ear." This is especially true when we can enter into a conversation with the person who is providing us with information. We can provide immediate feedback to clear up misunderstandings and to ensure generally that the message is being correctly received. Unfortunately, messages received "by ear" are ephemeral and cannot be reproduced exactly or referred to later (except in the case of sound recordings, in which the conversational advantage of feedback is lost). Further, it is not possible to talk to, or even to know of, all persons who might provide one with information most useful in solving a particular problem at hand. Hence, we have made increasing use of written, non-ephemeral messages in order to communicate information from the mind of one person to that of others. In fact, abstract or complex information must be received via the sense of sight, preferably in permanently recorded form which permits repeated reference thereto, mulling of the information in the mind, graphical representation, and ultimate comprehension. Upon this fact hinges most considerations of information dissemination, storage and retrieval.

2.2 COMMUNICATIONS CONTINUUM

There are two facets of nearly all communications problems that are significant to consider in the field of information storage and retrieval. These facets are what may be called "feedback of information" and the "degree of abstractness" of the information being communicated. Figure 2-1 illustrates the Communicative Continuum depicting the "degree of abstractness" in the ordinate dimension and the "degree of feedback" in the abscissa dimension.

2.2.1 Feedback

Perhaps the best example of feedback in a communication process is a conversation between two persons, since a conversation provides a two-way communication link over which messages are sent. There is a continual stimulus-response: remarks call up other remarks, and the behavior of the two individuals becomes concerted, cooperative, and directed toward some objective.[(45)*]

Newspapers, magazines, and journals provide greater opportunity for communication between the originator and recipient of information as compared to history, archaeology, and cosmology. However, the feedback derived from such a communications link as the letters to the editor of a newspaper or magazine is still several orders of magnitude lower than the feedback provided by person-to-person conversation. The presence or absence of this type of feedback capability is an important consideration in the design of information systems, which are aimed at improving the process of communication. For example, one design consideration is whether the user should be able to conduct a dialogue with a retrieval system either directly with the machine or through an intermediary.

2.2.2 Abstractness

The ordinate dimension of Figure 2-1 portrays the degree of abstractness of the information being communicated, wherein abstractness refers to the amount of abstract thought required to utilize the information involved. Notice at the low end of the continuum that such things as multiplication and logarithmic tables do not require much abstract thought inasmuch as there is little ambiguity in the interpretation of such information. At the other extreme, however, music, art, humor, and poetry require a considerable amount of abstract thought, thereby creating considerable difficulties in communication.

---

* Reference numerals in parentheses refer to the Bibliography.

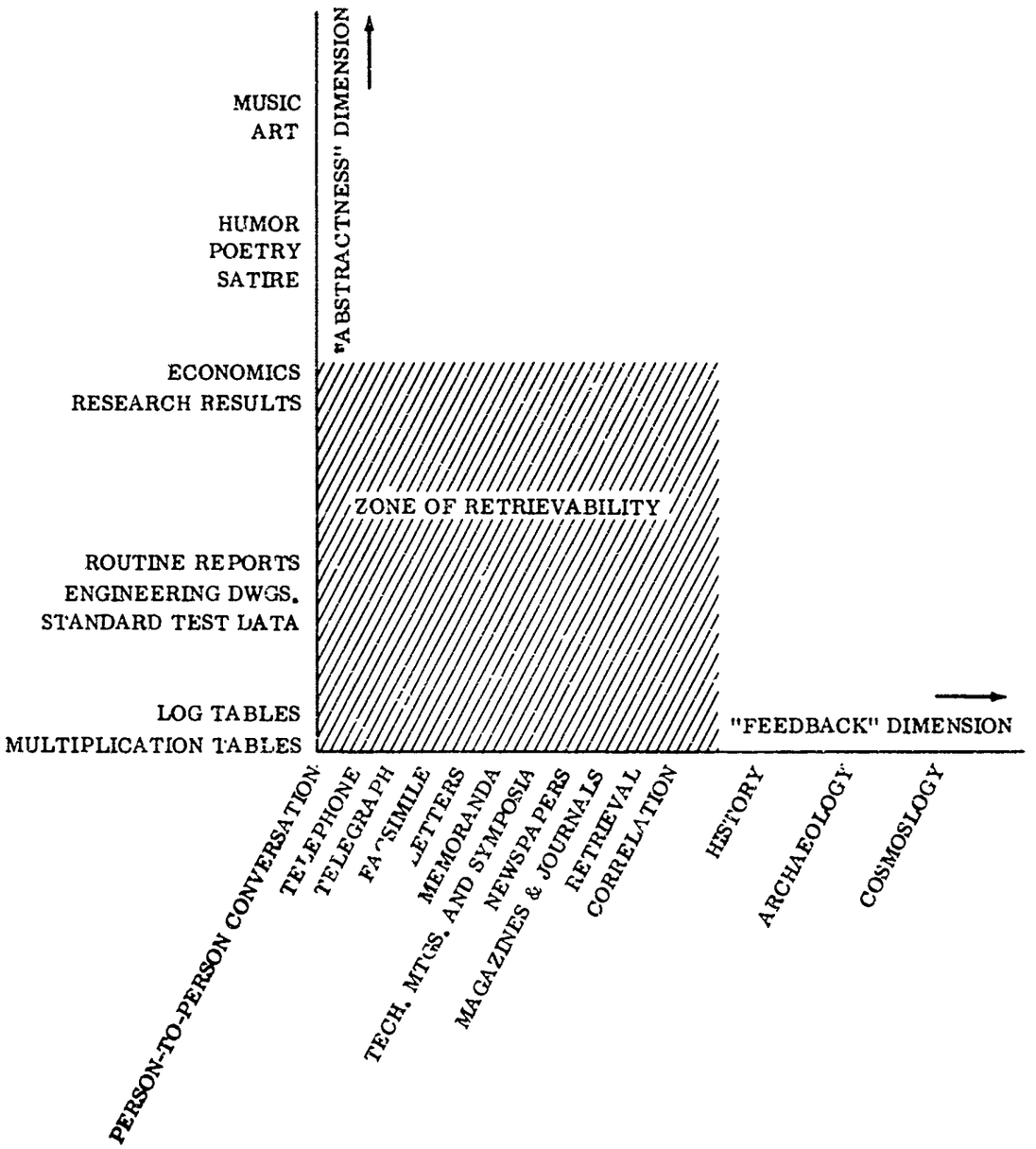


Figure 2-1. The Communicative Continuum

2.2.3 Zone of Retrievability

There is less interest in retrieving highly abstract information such as music, art, poetry, and the like, and in any case, the practical problems involved are quite significant. Also, those forms of information such as history, archaeology, and cosmology, which are at the extreme end of the feedback dimension, are probably outside of the zone of retrievability from a practical standpoint. The shaded area in Figure 2-1 represents the practical limits of the zone of retrievability and it is this area that is treated in this report.

2.2.4 Information Retrieval-Data Retrieval

Within the context of this report, no distinction will be made between information and data or information retrieval and data retrieval. Where these distinctions are made outside this report, data is generally used to characterize material which is easily quantifiable, non-abstract, and which can be formatted; and conversely, information is material which is conceptual, descriptive, usually narrative, and may be judgmental, and is not easily quantifiable or formattable.

Information storage and retrieval (IS&R) as used in this report is generic to all variations of the problem of storing, locating, and selecting information of any kind, whether it is in graphic or digital form and whether the desired output is a document or a specific fact.

2.2.4.1 Document Retrieval. The most common type of IS&R system provides as an output one or more documents which may be relevant to a request. In many cases, the first output of a system is a series of surrogates representing the documents, such as accession numbers, citations, or abstracts. All of these variations are also considered document retrieval systems.

2.2.4.2 Fact Retrieval. Systems are being developed which provide specific answers to inquiries, such as the name of a part having certain characteristics, rather than a document surrogate. These systems, which are also examples of IS&R, are sometimes called "fact retrieval" systems.

The major difference between document retrieval and fact retrieval systems is in the degree of specificity which can be achieved by the system. Fact retrieval systems require a greater depth of indexing and permit much more specific queries to be made of the system.

2 2 5 Graphic vs. Digital Information

Distinctions are sometimes made between graphic and digital information. These distinctions lie essentially in the form in which the information is stored. Graphic information may be stored in the form of books, catalog cards, loose-leaf notebooks, graphs, microfilm, and the like. Digital information includes any coded representation which can be processed by machines without first requiring a transformation into machine language.

2 3 CLASSIFICATION OF INFORMATION SYSTEMS

There are a number of ways of classifying information systems. Table 2-1 presents 10 different methods of classification with a few examples of the individual classes within each method. These 10 methods are:

(1) Class of Information

(2) Subject

(3) Type of Organization

(4) User Job Description

(5) Categories of Use

(6) Mode of Use

(7) Performance Characteristics

(8) Form of Information

(9) Output Media

(10) Type of Information Stored

TABLE 2-1. METHODS OF CL

Class of Information	Subject	Type of Organization	User Job Description	Cat	of Org
Concepts	Law	Stock Brokers	Stock Broker	Refe	okers
Cost and Funding	Chemistry	Pharmaceutical Companies	Lawyer	Surv	eutica
Design Techniques	Medicine	Law Publishers	Doctor	Selec	lisher
Experimental Processes	Biology	Hospitals	Purchasing Agent	Moni	s
Math Aids and Formulae	Electronics	R&D Organizations	Salesman	Veri	anizat
Performance Characteristics	Supply	Insurance Companies	Personnel Director	Colle	e Com
Production Processes and Procedures	Pharmaceuticals	Transportation Companies	Contract Admin.	Diss	rtation
Raw Data	Mathematics	Rate Bureaus	Comptroller	Diss	reaus
Specifications	Transportation	Police Departments	Intelligence Agent	Diss	epartr
Technical Status	Production	Credit Bureaus	Scientist or Engineer	Ref	ureau
Test Processes and Procedures	Personnel and Training	Airlines	Tech. Evaluation		
Competitive Intelligence	Management	Banks	Tech. Admin.		ment A
Market Intelligence	Metallurgy	Employment Agencies	Research		nce Ag
	Mathematics	Intelligence Agencies	Expl. Devel.		tions
	Aircraft and Flight Equip.	Corporations	Adv. Devel.		Record
	Aircraft Instruments	County Recorders	Eng. Devel.		el Dep
	Aircraft Design	Personnel Departments	Oper. Syst. Dev.		aw Fi
	Aircraft Structures	Large Law Firms	Reliability, etc.		onsult
	Flight Operating Problems	Large Consulting Companies			
	Flight Safety				
	Gliders				

TABLE 2-1. METHODS OF CLASSIFYING INFORMATION SYSTEMS

Organization	User Job Description	Categories of Use	Mode of Use	Performance Characteristics	Form of Information
ers	Stock Broker	Reference	Current Awareness	Completeness	Items
itical Companies	Lawyer	Survey	Retrospective	Relevance	Reports
hers	Doctor	Selection		Specificity	Journal Articles
	Purchasing Agent	Monitor and Control		Timeliness	Memoranda
izations	Salesman	Verification of Exist.			Letter
Companies	Personnel Director	Collection			Test Results
tion Companies	Contract Admin.	Dissem. for Info.			Drawings
us	Comptroller	Dissem. for Action			Reference Tools
artments	Intelligence Agent	Dissem. for Future Reference			Abstract Journals
eaus	Scientist or Engineer				Indexes
	Tech. Evaluation				Encyclopedia
	Tech. Admin.				Handbooks
nt Agencies	Research				Information Centers
e Agencies	Expl. Devel.				Bibliographies
ns	Adv. Devel.				Correlations
corders	Eng. Devel.				State-of-the-Art Reports
Departments	Oper. Syst. Dev.				Fact Retrieval Service
Firms	Reliability, etc.				Handbooks
sulting Companies					

B

FYING INFORMATION SYSTEMS

s of Use	Mode of Use	Performance Characteristics	Form of Information	Output Media	Type of Information Stored
d Control n of Exist. or Info. or Action or Future e	Current Awareness Retrospective	Completeness Relevance Specificity Timeliness	Items Reports Journal Articles Memoranda Letter Test Results Drawings Reference Tools Abstract Journals Indexes Encyclopedia Handbooks Information Centers Bibliographies Correlations State-of-the-Art Reports Fact Retrieval Service Handbooks	Teletype TV Display Microforms Opaque Transparent Roll 35 mm 16 mm Chip Microfiche Printed Page Facsimile Computer Printer Book Form Card Form	Inventory Data Usage Data Demand Data Engineering Drawings Specifications Standards Test Results Failure Data Maintenance Data Equipment Population Data Parts Characteristics Interchangeability Data Weather Data Maps Cloud Photography Crop Forecasts Medical Records Criminal Records Title Records Patents

C

Each of these methods of classification may be useful for some purpose and from various points of view. From an applications point of view, for example, classification of type of organization by subject field, by user job description, or by type of information store may be useful for identifying distinct markets. On the other hand, from a systems design viewpoint, classification by categories of use, mode of use, performance characteristics, and form of information are of more direct use and are discussed in more detail in the following paragraphs.

2.3.1 Current-Awareness and Retrospective Services (Mode of Use)

All information services may be categorized as being either a current-awareness service or a retrospective service. Current-awareness services generally aid a user in keeping up with what is happening. Retrospective services are designed to aid a user in determining what has transpired in the past. Newspapers, newsletters, memoranda, journals and magazines, and even abstract journals are examples of current forms of recorded information which are useful for current-awareness purposes. When such items are stored (and when the store is provided with a finding mechanism — such as an index), they become useful for retrospective purposes as well. The name Information Retrieval implies a retrospective purpose, i.e., a search of stored information. Most information systems, however, provide both current-awareness and retrospective products and services.

2.3.2 Use of Information

In general, little is known concerning the needs of users with respect to information content, characteristics, and form, although this is the most important consideration in IS&R. Because of the lack of reliable information concerning users' needs,[6] the art of information dissemination, storage, and retrieval is notably confused. A variety of information products and services with many different characteristics is being offered to the potential users — and sometimes without a clear understanding of their actual needs.

An important aspect concerning the use of information systems is the common reluctance of people to use information systems, however well devised. Mr. Calvin N. Mooers has hypothesized what has come to be known as "MOOERS' LAW" which states:

> "...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."[32]

The rationale behind "MOOERS' LAW" is that it is frequently painful and troublesome to have information since you must obtain it, read it, and then understand it, which is not always easy. It may also require you to make a decision or prove that your work was wrong or needless.

2.3.3 Categories of Use

Information systems may be used by different people for various reasons. One type of user may want a general survey of the literature in a field he is only slightly acquainted with. Another category of use might be called verification of existence, e g., a prior art search for the purpose of ascertaining the patentability of a new device or a search to verify the validity of a real estate title. A manager of a construction or maintenance organization may wish to utilize an information system to pull together all of the information which is relevant to a particular task so that it might be disseminated for action to the foreman along with the wo order.

Each of the above categories of use as well as many others will require different characteristics of information output, which characteristics are described in the next paragraph. Table 2-2 lists a number of categories of use and shows the probable characteristics of th information required, indicates whether the information service is of the current-awareness or retrospective variety, or both, and the probable form of the information required. More reliable predictions of performance characteristics would require a determination of user needs and detailed system requirements.

2.3.4 Performance Characteristics of Information Systems

From the viewpoint of the recipient, information systems have three major performance characteristics. completeness (recall), relevance, and specificity. Completeness refers to how many of the stored documents containing relevant information are supplied to the user. Relevance refers to how much of the information supplied to the user is pertinent to his needs. Specificity is a comparative term which refers to the degree of generality of the relevant information supplied. For example, information on race horses is more specific to an inquiry on race horses than information on horses generally. although the latter information may be relevant.

TABLE 2-2. CATEGORIES OF USE AND TYPICAL PERFORMANCE CHARACTERISTICS OF INFORMATION SYSTEMS

Categories of Use (Examples)	Probable Performance Characteristics of System							
	Mode of Use	Completeness	Relevance	Specificity	Response Time	Form of Output		
						Item	Ref. Tool	Correlation
1. Reference (Handbook) (Dictionary) (Tables etc.)	R. S.*	low	high	high	minutes			X
2. Survey (Bibliography) (Reports) (State-of-art reports)	R. S.	high	low	low	days		X	X
3. Selection (Part Selection) (Personnel Selection)	R. S.	med	high	high	hours	X		X
4. Verification (Patent Search) (Title Search)	R. S.	high	low	low	days	X		
5. Dissemination for Information (Memorandum)	C. A.**	low	med	low	days	X	X	X
6. Dissemination for Action								
(Work Order)	R. S.	low	high	high	hours	X	X	X
(Design Engineering)	R. S.	low	high	high	hours	X	X	X
(Research & Development)	R. S.	high	low	low	days	X	X	X
7. Dissemination for Reference (Directive) (Memorandum) (Report)	C. A.	low	med	low	days	X	X	X
8. Unification (Market Intelligence) (Sales Files) (Central Files)	R. S.	med	med	med	minutes	X		

* = retrospective search
** = current awareness

Different users require different degrees of completeness, relevance, and specificity. A man engaged in basic research in one of the sciences might require rather complete information on the subject in question, particularly in the early stages of his work, but would not require that the information received be completely relevant; he might even be distressed with information that is too specific (e.g., too quantified). A design engineer, on the other hand, would not require complete information (say, on all electric motors) but rather would need highly relevant information coupled with specific performance data.

In 1961, investigations in England[11] found that in all information systems there is a "trade-off" between completeness and relevance. For example, in a retrieval situation, an inquiry might be answered by responding with the entire file; completeness would be high but relevance would be low. Alternatively, the response might consist of only one relevant item; relevance would be high but completeness may be very low. Between these extremes lies a spectrum of possibilities. Invariably, the characteristics of completeness and relevance vary inversely with each other, but not linearly.

There is no optimum completeness/relevance trade-off point for any given information system. Consider the inquiry: "How many tons of coal were produced in West Virginia in 1962?" Although there may be on file 100 items containing this information, the inquirer wants only one of them. Highly specific inquiries require responses with a high degree of relevance but a low degree of completeness. At the other extreme, very general inquiries require responses containing generalized information. The emphasis is on the completeness rather than the relevance of responses. Specificity of information, therefore, is a very important characteristic of information.

2.3.5 Forms of Information

To each user, the form or packaging of the recorded information is important. There are three major degrees of form: individual items, reference tools, and correlations. "Form," as used in this report, has a broader meaning than format. Format has to do with the arrangement of information elements and is important in fact retrieval systems as well as document retrieval systems. Form is more pertinent to document retrieval systems. Individual items include letters, drawings, memoranda, standards, reports or

articles appearing in newspapers, journals, etc. The user usually has to "piece together" from many such individual items the exact information he really needs. Reference tools, appropriately indexed, include abstract journals, catalogs, indexes, libraries, annotated or unannotated bibliographies, and document retrieval systems, all of which enable the user to locate items which may be relevant to his needs and tentatively to evaluate them for relevance prior to examining the full text of the individual items themselves. Correlations include such items as state-of-art reports, handbooks, fact retrieval systems, and the results of manipulations of quantitative information with the objective of providing direct answers to the questions at hand. Within each of these basic forms of information, a large number of variations is possible.

2.3.5.1 Individual Items. The first form of information to be considered is that of the individual item, which is part of either the separate or serial literature. For almost two centuries, the separate literature (books) was reserved for relatively complete information of relatively permanent value, whereas the serial literature (journals) was used to report information quickly of a partial, ephemeral, or interim nature. Libraries tended almost to ignore the serial literature, generally limiting their treatment thereof to the collection and binding of the journals with no cataloging or indexing of individual journal articles. The void in coverage of serial literature was eventually filled, at least in part, by the abstracting and indexing services. Other types of individual items such as drawings, memoranda, letters, etc., were rarely organized by library methods.

Around the beginning of the 20th century the report, a different form of separate literature, became important, but for four decades thereafter it was confined largely to internal operations of individual organizations. In the 1940's, however, the increasing role of Government sponsorship in the area of research and development and other circumstances combined to promote the report to the status of a major medium of mass rather than private communication.

The report is characteristically not subject to critical review prior to publication (as are most journal articles and books), yet it tends to contain information less ephemeral than journal articles. The insubstantial physical construction of reports (compared to books) and their burgeoning numbers also caused problems. But, most important, there was no established mechanism for disseminating reports regularly to all potential users (vs. the ability of any potential user to subscribe to a journal), and the vast quantities of

reports generated precluded the announcement of their existence in the media normally used for the announcement of newly published books.

Centers were created wherein reports could be collected — e.g., ASTIA (now Defense Documentation Center), and the Office of Technical Services. Such centers collected, abstracted, and indexed the accessioned reports and often published abstract journals as current-awareness tools. When a user requested a report, a photocopy was created for him or supplied from stock.

2.3.5.2 Reference Tools. The creation of central collections of reports made essential the creation of effective reference tools (e.g., abstract journals, catalogs, indexes) for both current-awareness and retrospective purposes. Improved announcement techniques had to be developed. Indexes to the collections — indexes better than those previously used for books — had to be created. These reference tools are the second major form of information.

Reference tools direct the user to the individual items which seem to be of interest and permit him to evaluate individual items without actually obtaining them by providing a meaningful surrogate such as an abstract or a title. Whereas the individual items unassisted are generally useful only for current-awareness purposes, the addition of reference tools makes them useful also for retrospective purposes.

2.3.5.3. Correlations. So far, we have discussed two information forms: individual items and reference tools. The third form of information to be considered is that of correlations. Correlations which include state-of-the-art syntheses, handbooks, and fact retrieval provide the user with the exact information required at the moment. Thus they are distinguished from individual items. An item may provide information possibly needed in the future (current awareness) or part (but usually not all) of the information needed at the moment. Users usually have to perform their own correlations of individual items. Correlations are also distinguished from reference tools, which permit the user to evaluate individual items (before actually obtaining them) by directing the user to those individual items which seem to be of interest.

The typical current-awareness correlation is a state-of-the-art report. Such a report appears regularly in some journals. Because of the lack of knowledge concerning user's needs,[6] it is not certain whether many of the state-of-the-art reports now created are truly useful, nor whether other and different types of state-of-the-art reports are required. This form of information is likely to remain in stasis until better means are developed to determine the true and current needs of potential users.

Given a request by a user, it is theoretically possible to "tailor" a state-of-the-art report to satisfy the request In practice, this is so time-consuming and expensive that such procedures are practicable only when wealthy or powerful users make requests of an appropriate information center. There appears to be little likelihood that this situation will change markedly.

On the other hand, retrospective correlation of quantitative (or near-quantitative) information has long been practiced, and even more generally since the advent of computers. Quantitative data, being precisely definable, tends to develop a small vocabulary even for large systems. Further, with some effort, the vocabulary may be established almost completely prior to the initial input of data into the system. Synonymous situations may be eliminated, and hierarchical relationships may be made relatively invariant, i.e., there are no unspecifiable relationships. Finally, the syntactical problem can be solved by formatting, either by specifying certain fields for particular elements of information (closed format) or by tagging various elements of information with a unique code.

SECTION III. INFORMATION STORAGE AND RETRIEVAL CONCEPTS AND TECHNIQUES

The heart of any information storage and retrieval system is that portion which makes it possible for the user (or the user's agent) to locate the information which is needed or of interest. Without a satisfactory locating function, an IS&R system is useless. Thus, over the years, a great amount of time and effort have been expended in developing satisfactory methods of locating information contained in a collection.

The first approaches to the problem of developing a method of locating information had their roots largely in philosophical and (to at least some extent) metaphysical concepts. For centuries, there had been a consensus that there is inherent order in the universe. Hence, there could be order in all our information about the universe if only we could know enough to perceive that order. This idea was encouraged during the heyday of the "causative" or "mechanistic" concept, when it was held that all events are caused by preceding events and can thus be predicted, if the interactions of current events can be detailed sufficiently.

3.1 TRADITIONAL CLASSIFICATION AND INDEXING

The feeling that there was an inherent order in knowledge brought about attempts to organize knowledge accordingly. The result was the development of hierarchical classification, which attempts to (1) group together those subjects which are alike, (2) separate those subjects which are not alike, and (3) provide an essentially continuous gradation of separation with respect to degree of similarity.

The Dewey Decimal Classification (DDC),[16] first published in 1876, was the first widely-used example of hierarchical classification. Even Dewey, however, recognized that "one-place" assignment of a concept in a framework of all knowledge (i.e., assumption of an inherent order) did not reflect the real world, and he provided for "parallel divisions" (i.e., common sub-categories were provided under many of the main categories).

Cutter followed with his Expansive Classification[12] in 1882, which formalized "common sub-divisions," separated from the table of general divisions.

The Universal Decimal Classification (UDC),[26] based upon DDC and developed around the turn of the century, provided in its notation an ability to express two types of relationships between subjects -- the hierarchical (or class inclusion) relationship and a "general" relationship covering all others. In 1901, the Library of Congress Classification (LC) was developed, based largely upon Cutter's work.

The first real break with the concept of inherent order came in 1933, with the publication of the first edition of Raganathan's Colon Classification (CC).[35] Later and more generalized elaborations of CC have become known as "faceted classifications." In such systems, hierarchical relationships are still fundamental. Individual documents, however, can be assigned to several different places in the classification and the complete notation of such assignment will then include the notations of all the classes employed, each individual notation separated from the others with a colon. Elaborate rules concerning the order in which the individual notations are listed, or "rotation" of the complete entry so that each individual notation is used as a filing entry, provide multiple-access capability for CC. Nevertheless, the inflexibility of the fundamental hierarchical scheme penalizes CC just as it does earlier classification. If every new concept must be integrated into an established, rigid framework, either the new concepts will have to be structured "to fit the system" or else the system will have to undergo major revisions. New concepts almost by definition change the classifications, the hierarchical relationships or the structure.

Because of this inherent rigidity of classifications, there developed a trend toward using mere "subject headings," which could be arranged alphabetically. Subject heading systems thus abandoned entirely the concept of inherent overall order. In some instances, however, a certain amount of hierarchy was introduced by providing sub-headings for major subject headings. For example:

Brick

Firebrick
Silica Brick

Bridges

Arch Bridges
Floating Bridges
Lift Bridges
Portable Bridges
Suspension Bridges

Brighteners

Brines

Briquets

etc.

Under these circumstances, it was often difficult to find the term "Arch Bridges" (for example) in an index, in that it appeared under "Bridges" (in the B's) rather than between "Archaeology" and "Arches" (in the A's).

Thus, for the alphabetical systems (whether or not hierarchy was also employed), cross-referencing between terms (e.g., "see," "see also") became essential. The probability of finding relevant terms was, to a large extent, a function of the ingenuity of the user.

3.2 COORDINATE INDEXING

It is not difficult to see how the last vestiges of the concept of inherent order (i.e., hierarchy) could be discarded and the other aspects of faceted classification combined with those of alphabetical term lists to form the foundation for coordinate indexing, a new and more flexible approach. In fact, coordinate indexing did not first arise in this manner, but rather via the utilization of mechanical devices, and the relationship of coordinate indexes to earlier types of indexes and classifications has been recognized, generally, only in retrospect.

Coordinate indexing utilizes the concepts of Boolean algebra in searching and retrieval. For example, it assumes that if a document is indexed by the terms "Reproduction" and "Books" that the document discusses reproduction of books. That is, the logical, Boolean intersection of the set of documents indexed by "Reproduction" with the set of documents indexed by "Books" will be the set of documents discussing reproduction of books.

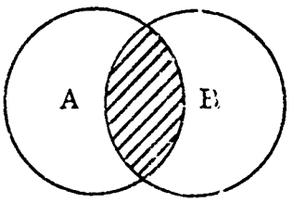
Similar reasoning is followed with respect to logical union. The set of documents about documents is assumed to be the union of the sets of documents indexed by the terms "Books," "Reports," "Monographs," "Letters," "Brochures," etc. Logical negation (A but not B) can be similarly utilized. Thus, coordinate indexing permits the use of less complex terms during indexing and allows the more complex concepts to be formulated during retrieval. Figure 3-1 is a Venn diagram illustrating logical intersections, unions, and negations.

3.2.1 History of Coordinate Indexing

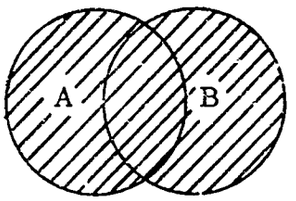
The first well-known technical application of the technique was made by Batten in England during the 1940's, when he indexed documents about chemistry by coordinate methods.[5] He used the dedicated-space, internally-punched card (e.g., peek-a-boo, Batten or "Termatrex"card) as a physical device. (See Figure 3-2.) Each card stood for a given term. Such a card dedicates one of a large number of positions on each card (the same position on each card) to one document. When a particular term is used to index a document, a hole is punched on that term card in the proper document position. During retrieval, the appropriate term cards are superimposed and documents indexed by all the inquiry terms are signified by a coincidence of punched holes on all the term cards (i.e., logical intersection). Logical unions or logical negations are considerably more difficult to perform with such a device.

Mooers[33] then developed the "Zatocoding" system, which provided one edge-notched card per document. A certain combination of notch positions was allocated for each term, and the entire deck of cards (one per document) was examined (via the usual "knitting needle" technique) to find those with the notched positions of all inquiry terms (i.e., logical intersection). Logical unions could be performed by searching for each term in sequence and combining the cards found for each term. The primary advantage of this random superimposed coding method is that it reduces storage requirements by making effective use of hole positions. It does produce some "false drops" or wrong answers, however, because the code combinations are not all unique.

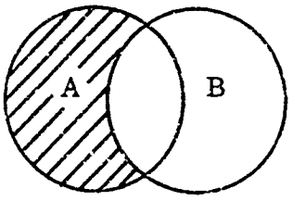
Taube[42] then promoted the "uniterm" system which was similar to the Batten system (one card per term), except that space for each document was not dedicated. Rather, document numbers were entered (posted) on the cards and retrieval was performed, for intersections, by finding matched postings on all the cards of the inquiry terms. Unions



Intersection
A∩B
(A and B)



Union
A∪B
(A or B)



Negation
$A \cap \overline{B}$
(A but not B)

Figure 3-1. Venn Diagrams

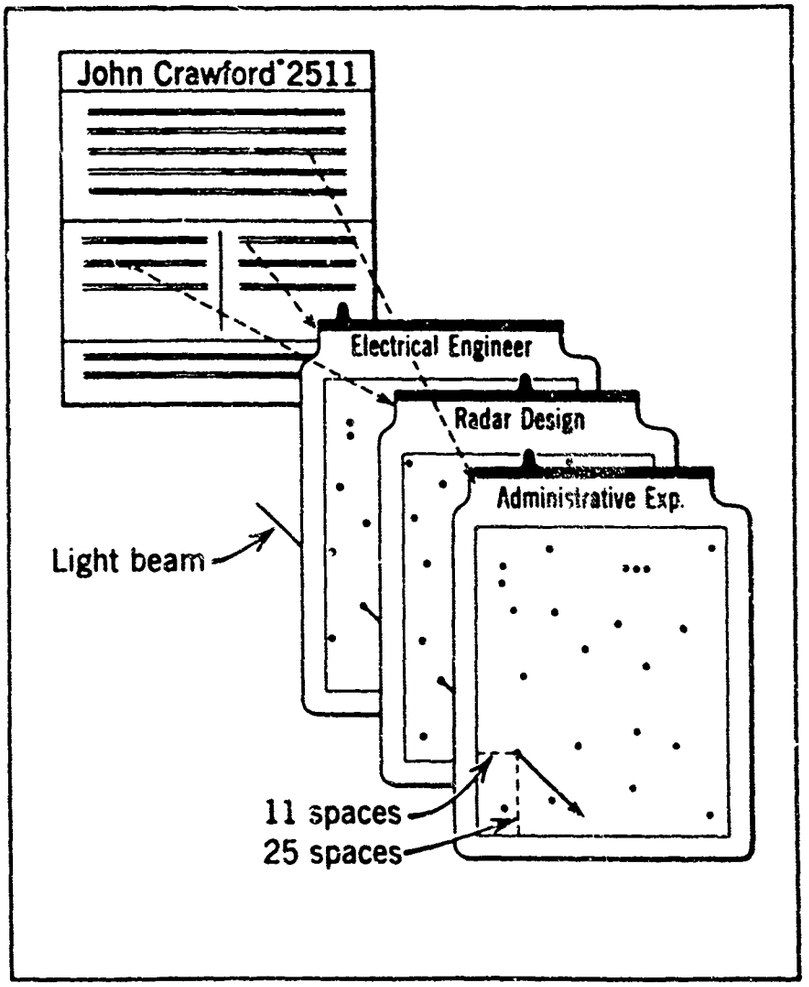


Figure 3-2. Termatrex Cards

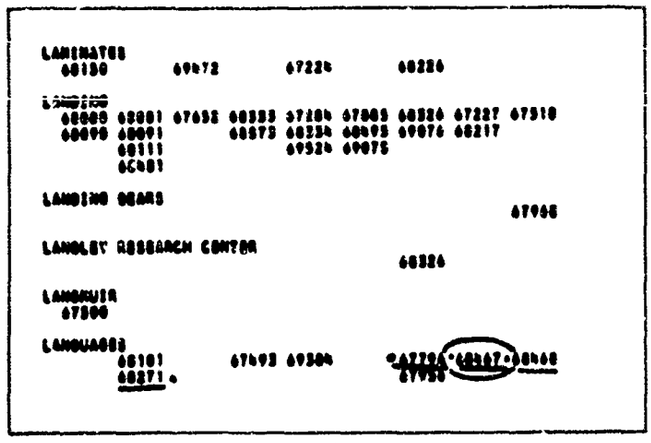
could be developed by merging the postings on term cards. Figure 3-3 illustrates a logical intersection on printed uniterm cards.

Both Mooers and Taube promulgated intellectual techniques to accompany their physical index devices. Mooers suggested the use of index terms which he called "descriptors." In this original sense, "descriptors" are very broad terms, each with a scope so well delineated that the "descriptors" become merely "tags" for the delineated concept. Taubes "uniterms" were originally conceived as single-word index terms chosen from the text being indexed (i.e., "free indexing"). Subsequent developments, however, have proved that short phrases rather than single words are often most useful, and that index terms brought to mind by the text (or by some cross-referenced authority list of approved index terms) are also useful.

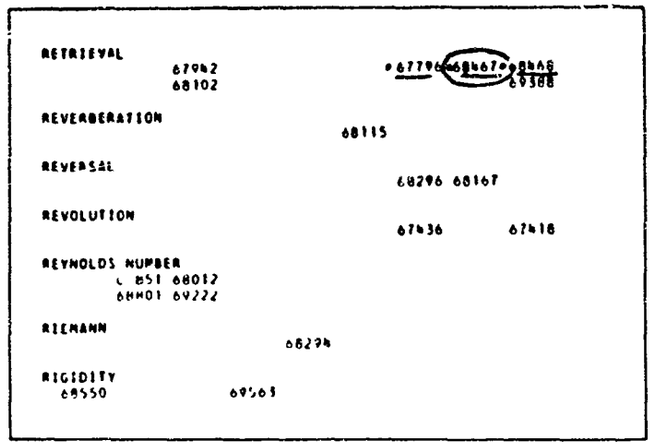
Multi-word terms (or phrases) tend to be more specific than do single-word terms; hence, the evolution of the "uniterm" system has resulted in its diverging considerably from the "descriptor" system. In general, the larger the collection indexed, or the more complex the information contained therein, the more appropriate is some modification of the "uniterm" system. This is particularly true when a current-awareness index is being provided. In such indexes, the terms should "stand alone," instead of being useful only in conjunction with other terms.

Taube's basic idea of "free indexing" using the full range of terms found in the text has proved better in practice than Mooers' small vocabulary of broad "descriptors." When vocabulary control is considered, however, Mooers' approach is more practical. Mooers felt that not all possible words would be useful "descriptors" and that the language of ordinary discourse should be translated into "descriptor" language for purposes of indexing and retrieval. Taube felt that the language of ordinary discourse should be used for indexing and retrieval. Taube's approach, however, is fraught with difficulties, as described below.

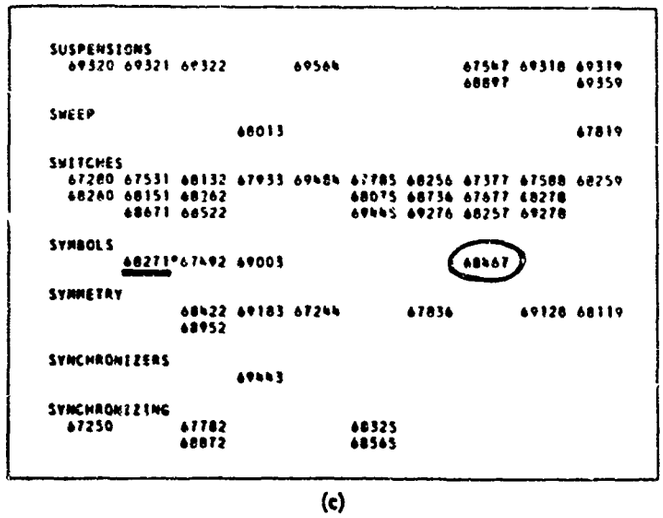
Indexes use words to constitute their index terms, and words are imperfect communication devices. The same word will mean two different things to two different people, or the same concept may be described by two different people using two different words. Further, the viewpoints of different persons affect their use of a common vocabulary. One



(a)



(b)



(c)

Documents on Retrieval ∩ Languages ∩ Symbols

Figure 3-3. Printed Uniterm Cards

person may view a real world referent or concept from a broader point of view than another (e.g., horses vs. (race) horses). Finally, the order or arrangement of words (i.e., syntax) may cause difficulty in communication. It is thus necessary to introduce specific syntax and vocabulary rules into vocabulary development efforts before significantly useful retrieval services can be provided, particularly for large files.

How may such rules and utility be introduced into a retrieval vocabulary? Mooers' approach was to classify the language of ordinary discourse into broad (not necessarily hierarchical) categories, each of which he called a "descriptor." Such an approach is not applicable when a full range of term specificity is to be provided in a vocabulary.

3 2 2 The Thesaurus Approach

The approach to vocabulary control which has become most widely used during the past few years is that of providing a thesaurus[48] for the retrieval vocabulary. An information retrieval thesaurus generally has several characteristics: (1) it lists vocabulary terms authorized for use in the system, (2) it exhibits relationships among these terms — relationships such as synonymy or hierarchy, and also relationships which may indicate synonymy or hierarchy from some points-of-view but not generally, and (3) it defines the vocabulary terms to the extent required.

The functions of an information retrieval thesaurus are (1) to permit indexers of documents containing valuable technical information to index (i.e., describe) more fully, and at different levels of generality and from many technical points-of-view, the information contained in documents.and (2) to permit searchers for information to phrase inquiries appropriate to the scope and degree of their immediate interests — inquiries employing all terms of the retrieval vocabulary which have appropriate meaning and specificity.

3.2.2.1 Listing of Vocabulary Terms. An information retrieval thesaurus may serve as an authority list in the conventional manner. Any term which an indexer wishes to employ to describe a document but which is not in the thesaurus must have its use justified.

It may be that the term being considered for use is sufficiently near in meaning to an already-accepted term that the accepted term can be employed instead or, if the candidate term is more desirable, it can be substituted throughout the index for the previously-accepted term.

It may be that the candidate term is not a near-synonym of an existing term but describes a member of a class of objects or events described by the existing term. If the candidate term is not expected to become important in the system, the broader existing term may be used in lieu of it to index the document and the candidate term rejected accordingly.

3.2.2.2 Exhibiting Relationships Among Terms. There are two basic relationships among terms important in information retrieval; these are the relationships of synonymy and hierarchy. As noted above, however, two terms may be related synonymously from one viewpoint and not from another. For example, "Salvage" and "Reclamation" may be synonyms from many viewpoints, but from at least one viewpoint they are near-antonyms; from this viewpoint "Salvage" is more nearly synonymous with "Scrapping". Similar comments can be made about hierarchical relationships. In such instances, when synonymy or hierarchy cannot consistently be specified, the occasional or possible existence of such relationships must nevertheless be exhibited in the thesaurus.

There are, of course, very few true synonyms other than spelling variations ("Sulfur" vs. "Sulphur") or abbreviations ("mph" vs. "Miles per Hour"). However, in the context of any given information retrieval system (even that of a large system) many terms are sufficiently close in meaning so that they can be treated as synonyms. This is especially true with very specific terms (e.g., "Imperfections" vs. "Defects"). The specification of synonymy, however, must be cautiously undertaken.

Just as there are few true synonyms, so also there are few (if any) fundamental hierarchical relationships. Again, however, within the context of a given system, hierarchical relationships among terms can often be specified (e.g., "Electric Motors" narrower term: "D-C Motors" and conversely "D-C Motors" broader term: "Electric Motors"). In addition, where there are more than two "levels" of hierarchy, all levels should be exhibited.

The "see-also" reference has long been employed by most library systems to indicate the unspecifiable relationship (and in addition, usually, to indicate the "down" hierarchical relationship noted above). The notation "related terms" is becoming increasingly preferred to indicate this relationship, thus distinguishing between hierarchical and unspecifiable relationships.

3.2.2.3 Defining of Terms. By exhibiting synonymous and hierarchical relationships among terms, term definitions are provided as well, partly by extension and partly by intension. Occasionally, however, a word will occur which has several different meanings. These words are homographs. Usually, brief parenthetical modifiers will suffice to distinguish between the terms, e.g., contact (meeting) vs. contact (electrical). Occasionally, particularly when a term is very broad in meaning, there is need for another type of scope note — an instructional scope note; such a scope note limits the meaning of the term and specifies how it should be used so that it will not be overused and thus lose its retrieval utility. Finally, there occasionally appear terms which are esoteric or unfamiliar or used in an unusual manner; such terms require definitive scope notes — i.e., true definitions. Our experience has been that only a small percentage of the terms in a vocabulary require definitions.

3.2.2.4 Practicability of Thesauri. Because any one profession has such a large operational vocabulary, it may reasonably be asked: "How can all terms of the vocabulary be included in the thesaurus and all cross-references be provided without excessive cost and bulk?" The answer, proved in every vocabulary investigation to date, is that the retrieval vocabulary need be only a fraction of the size of the operational vocabulary. For example, the architect may use the term "Gracefulness" as an operational term, but for retrieval purposes a more general term, such as "Esthetic Quality," should be more appropriate.

In fact, most terms in any vocabulary are seldom used, and for indexing and retrieval purposes a few slightly more general terms may be substituted for the numerous very specific terms, with the result that the retrieval vocabulary is relatively small. It has previously been found[20] that the vocabulary size (V) of a scientific or engineering information system is related to depth of indexing (d) and number of items indexed (D) according to the approximation:

$$V = 3330 \log_{10} (Dd + 10^4) - 12600$$

provided that proper names (or their equivalents, such as names of specific chemical compounds) are excluded from the vocabulary. The logarithmic factor represents the fact that the growth of the vocabulary is considerably slower than the growth in the number of items. There would be required, of course, a list of single-entry cross-references from overly-specific operational terms to slightly-broader retrieval terms, and this list would grow larger with the size of the collection. Such a list would require very little space, however, in either a printed or a mechanized system, in that each operational, specific term would appear only once, with a brief reference to the appropriate retrieval term or terms.

3.2.3 Syntactical Problems

The traditional pragmatic and semantic problems of communication are thus shown to be subject to reasonable solution via appropriate techniques of vocabulary control. The traditional problem of syntax has required the application of different techniques.

A symptom of the syntactical problem is the retrieval of non-pertinent information. For example, a document indexed by the terms "Cleaning," "Coal," "Grinding," and "Boiler," might be concerned with "grinding of coal" and "cleaning of boilers," but not "cleaning of coal." Yet items covering the latter would be retrieved as a result of a request for all documents indexed by "cleaning" and "coal."

Retrieval of non-pertinent information as a result of syntactical ambiguity becomes more likely as the depth of indexing increases, because there are more possible term combinations, many of which are spurious. Thus, if deep indexing is employed, syntactical ambiguity must be minimized. Two basic means of doing this have been employed, either singly or together.

3.2.3.1 Linking. The first technique used to combat syntactical ambiguity is that of item subdivision. In effect, each item is indexed as if it were a number of smaller or less comprehensive items, thus reducing the depth of indexing on each individual part of each item. The technique is known as linking and has also been called punctuating, or interfixing.

For the example given above, one link (A) of the item would be indexed by the terms "Grinding" and "Coal" while another link (B) would be indexed by "Cleaning" and

"Boilers" thus avoiding false retrieval. Presumably, a third link (C) would be indexed by "Coal" and "Boilers." The index for this example (DOC #100) could be symbolized as follows:

TERM	DOC. #	LINK(S)
GRINDING	100	A
COAL	100	A, C
CLEANING	100	B
BOILERS	100	B, C

The use of "links" or their equivalent increases markedly the cost of indexing. Care must be taken to make all appropriate "connections" between term pairs. Otherwise, pertinent information may not be retrieved during searching. The resulting redundancy also increases the size of the index.

3.2.3.2 Role Indicators. The second technique used to combat syntactical ambiguity is that of using role indicators (also known as modulants). In effect, each term is inflected by a small number of quite general role indicators to create a larger vocabulary of more specific terms. The technique avoids syntactical ambiguity which cannot be avoided by the use of "links." For example, a document discussing the use of Mylar* for packaging might be indexed by the terms "Packaging" and "Mylar." However, it would be retrieved by a request for documents discussing the packaging of Mylar. The use of links would not avoid this type of false retrieval. The application of a role indicator, such as "Uses of," to the term "Mylar" would solve the problem.

A set of role indicators ideally should be mutually exclusive (and thus relatively small in number) and collectively exhaustive (and thus relatively large in number). The ideal is obviously never achieved. A "good" set of role indicators is not easy to develop. In general, the tendency is to emphasize the use of a small number of roles, thus minimizing the complexity of rules for their use. The use of a small set of roles increases indexing costs by perhaps only 10 percent, but increases the effective size of the index vocabulary at least several-fold. The best-known set of role indicators today is that developed by the Engineers' Joint Council.[49]

---

* Registered Trademark.

Several studies have been made to evaluate the effectiveness of the use of links and roles in information retrieval systems. One rather detailed investigation made by Sinnett, and reported on as a master's thesis,[39] concludes that roles are not effective in improving recall and relevance, and increase indexing time and question formulation time by a significant amount. Links, on the other hand, were found to reduce irrelevant information retrieved by over 56 percent, with less than a five percent loss of relevant information.

Barbara Montague of duPont in another study,[31] however, concludes that roles appreciably increase relevance but reduce recall through errors in application, and represent 11 percent of total indexing cost. Links were found to represent four percent of total indexing cost, but did not appreciably affect the recall or relevance of patent retrieval. This latter point is inconclusive, however, as patents normally a 'imited to one composition or process, which can usually be indexed in one link.

3.3 AUTOMATIC INDEXING AND ABSTRACTING

3.3.1 Automatic Indexing

Because indexing by human beings is at best costly and at worst inconsistent, efforts are being made to develop automatic indexing techniques which employ computers to assign index terms to documents. Input to such an operation may be citations, abstracts, or full text. The present method of input is to keypunch the material to be indexed. Keypunching is excessively expensive for full-text but fairly cheap for citations or abstracts. Optical character readers are being developed such that text can be converted automatically to machine-readable media. Indications are, however, that optical character readers will be either too specialized (i.e., for only one-type font and size) or so costly that only large-scale operations can afford them. Machine-readable text may also be obtained as a by-product of document preparation where automatic typesetting is employed, but this is limited probably to those sources which the "system" can control.

3.3.1.1 KWIC Indexes. At any rate, once the text of the item surrogate is in machine-readable form, a number of things can be done with it. Titles can be used to create key-word-in-context (KWIC) indexes.[30] The computer uses each word in the title as an index term, and lists therewith the title. Thus, the title appears in the index once for each word in the title. "Non-informing" words (e.g., "and," "or," etc.) are excluded, based on a "common-word" list against which the computer compares each word in the

title. Unfortunately, KWIC indexes are inherently lightly indexed and no vocabulary control is exercised, (other than eliminating non-significant words), thus making them only partially effective for retrospective purposes. They are, however, quick and inexpensive and are immeasurably better than no index at all.

3.3.1.2 Automatic Indexing "In Depth". The use of abstracts or full-text as input is fraught with more fundamental difficulties. Vocabulary control (as described above) becomes increasingly important. Not all "non-common words" should be selected as index terms. Further, it is often true that a phrase is meaningful whereas the individual words therein are not (e.g., "Last Clear Chance"). Hence the problem is to develop an algorithm which will select from each text the "right" words or phrases as index terms. Even then, there exist the problems of hierarchy and near-synonymy, as discussed above. The "right" words or phrases may not even appear in the text.

The first attempts[29] to automatically index text depended upon selection, as index terms, of the words which appeared most frequently in a given text, followed by deletion of the "common words." It was next attempted to assign all selected words to one or more "notional families" (i.e., broad concepts) and to index the text by the "notional families' occurring most frequently. Rationally derived "notional families" were not easily developed, however. Assuming that a thesaurus could be made available to the computer, the words selected from the text could presumably be augmented automatically with synonymously and hierarchically related terms from the thesaurus. Even this is of dubious value in that context is important as with the term LOGIC when indexing different documents about philosophy and computer design.

The most recent efforts[17] attempt to determine the distribution of each word in all texts of a collection. It is then assumed that "non-informing" words will be evenly distributed over all texts and can thus be detected and deleted. The remaining words would be "informing" words, and each such word would be used as an index term for those texts containing it most frequently. There is a certain rationale to approaches of this type, but they still leave unsolved the problem of automatic index augmentation and the detection of meaningful phrases.

It is slowly being realized that automatic indexing will be just as inconsistent as "human indexing," because automatic indexing depends primarily upon the words used in a

text written by a human being. The realization of the unattainability of a high degree of consistency in "human indexing" brought about action to develop thesauri for use both in creating and in searching "human indexes" (as described above). So also is the realization of the unattainability of a high degree of consistency in automatic indexing bringing about efforts to develop and use thesauri via computers. The realization has not yet generally dawned, however, that automatic use of thesauri (whether they be created by man or computer) to augment search terms is probably self-defeating again because of the problems of context.

First is the fact that term augmentation (either during indexing or search) ideally should select quite different thesaurus terms in response to slightly different connotations of the same text or search term. (Witness the example cited above with respect to the single term LOGIC when used by a philosopher and a computer designer.) No automatic method is in view for making this distinction among terms listed in a thesaurus as being related to LOGIC. It would appear that human intervention in the augmentation process is unavoidable. Second, there is only a modest effort being devoted to the automatic detection of meaningful phrases rather than single words when using a computer to generate a thesaurus based on text. Much of the work pertinent to analysis of text is being done by linguists and computer program (format) language developers as well as by IS&R specialists. (See Appendix A.)

Automatic indexing might succeed via use first of a syntactical analysis program(A8)* on full text, thus to detect meaningful phrases rather than single words. For example, one of the following sentences should be "described" by the term "information retrieval" and the other should not.

(1) Because the capsule contains secret information retrieval is essential.

(2) This study dealt with the retrieval of data and information.

Phrases detected via this syntactical approach could then be processed to ensure extraction of all significant candidate index terms (e.g., "retrieval" from "information retrieval") and their addition to the candidate term list. The number of uses of each candidate term thus extracted from each document could then be tallied and the distribution

* References(A8) etc. refer to the supplemental bibliography in Appendix A.

of candidate term usage over a large number of documents could be determined. Candidate terms with highly skewed distributions would then be accepted as index terms for those documents in which they occurred with more than a threshold frequency. The automatic indexing process might stop at this point, without augmentation of the indexing either automatically or "cerebrally."(A21, A22)

Retrieval should consist of a human being referring to a thesaurus to phrase inquiries in all manners (and using all terms) by which the computer might have indexed the desired documents. (The same philosophy applies to human-indexed systems.) The usual Boolean expressions could then be utilized to retrieve documents either manually or mechanically.

Such a scheme of automatic indexing (a combination of several independent approaches(29) (17) (A8) now being investigated) would seem to have a good chance of being technically sound.

3.3.2 Automatic Abstracting (Extracting)

Efforts to abstract documents automatically have consisted of an extension of Luhn's technique of indexing via selection of the words (excluding "common words") appearing most frequently in the text.(28) Each text sentence is then ranked according to how many index terms appear therein, and how important (based on occurrence) the index words are in that given document. Highest ranked sentences are then selected as an "auto-abstract" or, more correctly, an "auto-extract."

A number of refinements have been suggested which would tend to increase the validity of Luhn's extracting technique. Baxendale(A7) claimed that a larger list of "uncommon" words would eliminate more noise and increase the probability of selecting the most meaningful words and sentences. She favored using only the "prime sentences" of a document consisting of the title, the summary, and the first and last sentences of each paragraph. She also attempted to isolate meaningful prepositional phrases statistically, since she believed such phrases to be an active part of speech which have a great influence on meaning.

It has been suggested that a thesaurus be built which would fit in the computer and which would be used to select only those words which bear strong meaning in a particular field of endeavor. But thesauri are difficult to build, consume much memory space, and increase the operating cost by slowing down the processing.

The statistical approach has been advanced a further step by Doyle[A10, A11] who advocates the introduction of large volumes of full text into the computer for statistical analysis. The analysis would permit the calculation of a weight for each word in a thesaurus to show its relative significance. A true evaluation of this approach must await the more widespread use of an inexpensive optical character reading device to convert into machine language the large amount of text required for the analysis.

Up to this point, the approaches described have been statistical. Climenson[A8] has developed a syntactical technique for parsing sentences in a computer in order to eliminate the modifying phrases and to retain the skeleton of each sentence. If the parsing is valid, the syntactical approach has an advantage over the statistical approach, which presents whole sentences and large gaps between them, causing a lack of continuity. The syntactical approach, on the other hand, avoids the problem of discontinuity by presenting at least a condensed form of each sentence in the original text. However, the automatic parsing of sentences has not yet reached an operational, much less economical, stage for commercial use.

Automatic extracting techniques have been generally unsuccessful, although occasionally a document is found which results in an extract which should be sufficiently informative, concise, and coherent. Most extracts produced by these various techniques lack one or more of these three important characteristics.

SECTION IV. INFORMATION PRODUCTS AND SERVICES

This section describes the various information products and services which are being employed to aid the process of communication between the originators and end-users of information. These products and services may be broadly categorized into two groups:

(1) current awareness services and

(2) retrospective services.

Table 4-1 classifies information products and services by their form, i.e., individual items, reference tools and correlations; and by their mode of use, i.e., current awareness or retrospective.

Table 4-2 summarizes some of the key aspects, including advantages and disadvantages, of the various information products and services described in this section.

.1 CURRENT-AWARENESS SERVICES

Until 1830 the scientific journal was the principal medium for disseminating scientific information. By the year 1830, there were about 300 journals being published. To cope with the problem of keeping abreast of so many journals, the first abstract journal was introduced in the year 1830. By 1950, there were some 30,000 primary journals and 300 abstract journals.[15] It has been suggested by Dr. Derek de Solla Price that the information problem is such today that what is needed is a significant advance, i.e., something which is to the abstract journal as the abstract journal was to the primary journal in the year 1830.[15]

4.1.1 Journals

There are a number of aspects which are common to most scientific and technical journals, which are worth describing here. First a journal is generally a collection of several articles. In most cases, there is a "critical review" by an editorial staff and others in the field before an article is accepted for publication. The majority of scientific

TABLE 4-1. CLASSIFICATION OF INFORMATION PRODUCTS AND SERVICES BY FORM AND MODE OF USE

Form of Information			
Mode of Use	Items	Reference Tools	Correlations
Current-Awareness Services	Journals	Abstract Journals	State-of-the-Art Report
	Reports	Contents Journals	
	Memoranda	KWIC Indexes	
	Correspondence	Current Bibliographies	
	Newsletters	Citation Indexes	
	Preprint Dissemination	Loose Leaf Services	
	Microform Dissemination	Selective Dissemination (abstract cards)	
	Test Results		
Retrospective Services	Journal Articles	Document Retrieval Systems	Fact Retrieval Systems
	Abstracts	Citation Indexes	State-of-the-Art Reports
	Drawings	KWIC Indexes	Handbooks
	Reports	Book-Form Indexes	
	Memoranda	Catalogs	
		Coordinate Indexes	
		Maps	
		Demand Bibliographies	

TABLE 4-2. SUMMARY OF KEY ASPECTS OF VARIOUS INFORMATION SERVICES

Current Awareness Services	Advantages	Disadvantages
Journal	Collection of articles Critical review Graphic arts quality Wide circulation	Too many journals High cost
Abstract Journal	Broad coverage Reduces bulk Provides bibliographic control	Narrowly indexed
Contents Journal	Inexpensive Convenient for browsing	
KWIC Index	Computer produced Quickly produced	Poor typographic quality Difficult to browse Poor indexing quality
Pre-Print and Report Dissemination	Direct Quick	Narrow circulation Relatively expensive Too much information
Microform Dissemination	Decentralizes copying Inexpensive Direct	Narrow circulation Too much information Inconvenient to read
Selective Dissemination	Tailored distributions based on profiles Automatic distribution Possibility of feedback	Expensive Requires deep indexing Feedback is difficult to obtain
Retrospective Services	**Advantages**	**Disadvantages**
Subject Heading Index	Easy to use Inexpensive to produce	Low depth of indexing
Uniterm Index	Ease of publication Coordinate searching Deep indexing	Difficult to use
Tabledex Index	Coordinate searching Can get item count directly Deep indexing	Difficult to use

TABLE 4-2. SUMMARY OF KEY ASPECTS OF VARIOUS INFORMATION SERVICES (Continued)

Retrospective Services	Advantages	Disadvantages
Citation Index	Provides searching ideas History of use of an article	Expensive to produce
Library Card Catalog	Ease of updating Provides a useful surrogate Catalog to index combined	Difficult to publish Low depth of indexing
Edge-Notched Cards	Needle sorting Can combine image and coding	Limited fields
EAM Cards	Machine sorting and reproduction Can combine image and coding	Limited fields
Peek-a-Boo Cards	Ease of searching Simplicity	Dedicated space Expensive to input
Machine Search - Document No.	Fast Efficient	Additional step required to find surrogate and then to find document
Machine Search - Citations or Abstracts	Provide surrogate to "Lock at" Relatively efficient	Separate step to find document
Machine Search - Document Display	Provides "Look at" function	Slow Expensive Inefficient
Machine Search - Microform	Provides "Look at" function Provides "Take away" function	Slow Expensive Inefficient
Machine Search - Data Correlation	Provides "facts" Provides "answers" Provides analyses	Difficult to set up

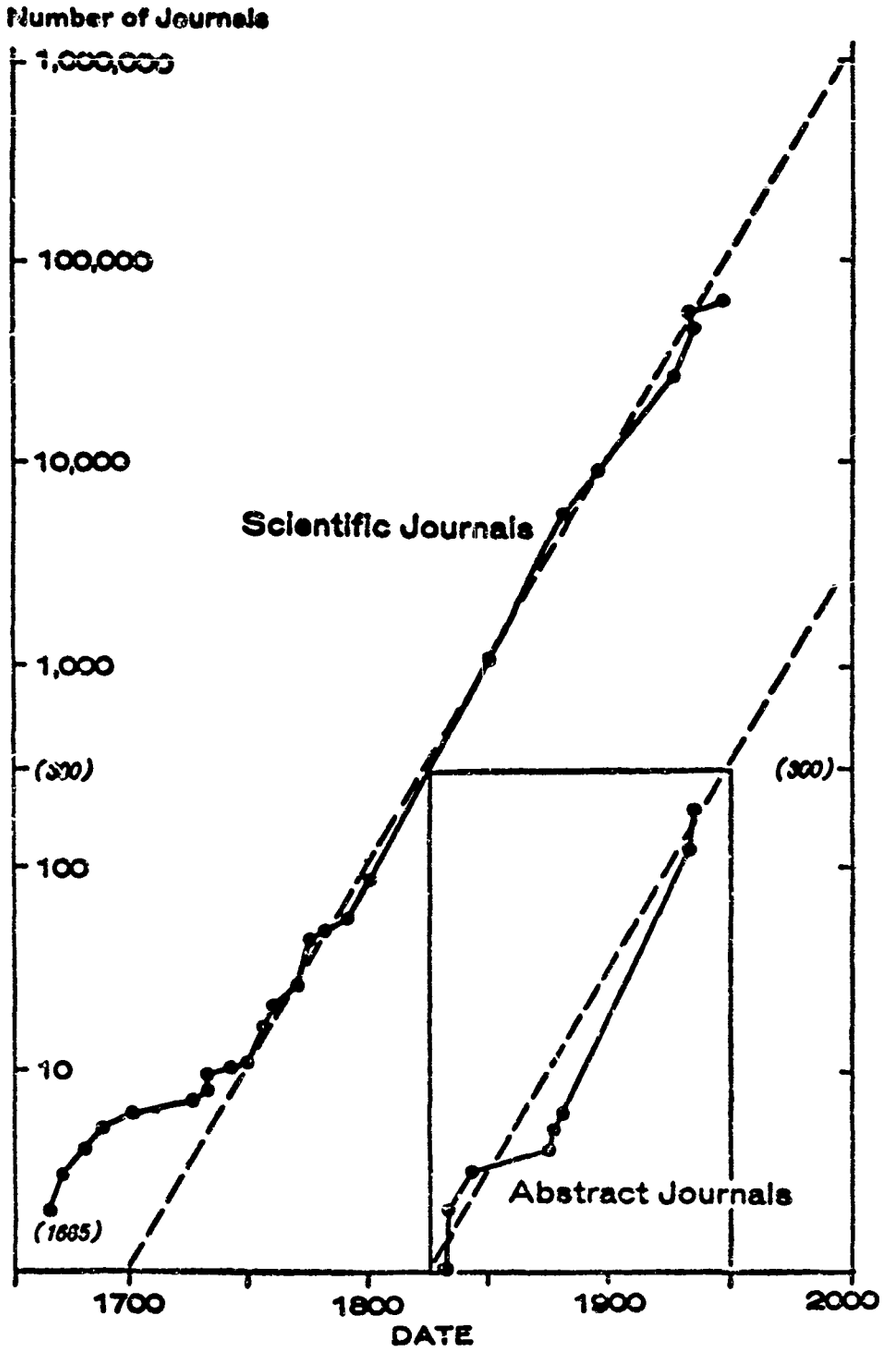


Figure 4-1. Growth of Journals and Abstract Journals
(Taken from de Solla Price, D. J. Little Science, Big Science)[15]

and technical journals still utilize graphic arts, quality typesetting and printing, although many journals, especially those produced by the smaller professional societies, are in financial difficulty. This is due primarily to the high cost of typesetting and printing and to the lack of growth in the number of subscribers. The latter cause is the result of a trend toward subscription by libraries and information centers rather than by individuals, wherein each copy receives more use (and frequent photo-copying) than does the copy of an individual subscriber. As a result, many scientific publications are being subsidized by Government agencies and non-profit foundations. This situation suggests the possibility that the primary journal may not continue to enjoy the major role it has played thus far in the communication of scientific and technical information. The journal article, however, is a social as well as technical phenomena, inasmuch as it contributes to the author's prestige as well as to the reader's knowledge.

4.1.2 Abstract Journals

Abstract journals originally were intended to perform a current-awareness function by compressing or digesting journal articles and thereby reduce the bulk of material which the subscriber had to read. Abstracts may be merely indicative (indicating content), in which case they are used to help the reader decide whether he should read the full text of the article, or informative (condensations or summaries), in which case the abstract may be a sufficient substitute for the article itself. With the great proliferation of information we have experienced in recent years, the so-called profession-oriented abstract and index services, e.g., Chemical Abstracts, Biological Abstracts, and Index Medicus are less frequently used for current-awareness purposes. Rather, they are purchased by libraries and utilized primarily in the retrospective search mode. On the other hand, the more specialized abstract and index services such as Cancer Chemotherapy Abstracts are more frequently utilized in the current-awareness mode as well as in the retrospective mode.

As a result of the trend toward the centralization of acquisitions of reports, journals, and other documented information, the component within the organization responsible for these acquisitions frequently publishes what has come to be known as an announcement journal. The announcement journal may be merely an acquisition list or may take the form

of a high-quality abstract journal with associated subject, author, report number, source, and other indexes, which are cumulated and republished periodically.

The arrangement of the abstracts within an abstract journal may vary from a completely random arrangement to a highly classified arrangement which tends to resemble an index. Under the purely random (accession number) arrangement, there is always only one entry of the abstract. Under the classified arrangement, there may be one or more entries, depending upon cost and bulk considerations and upon whether there is a separate index to the abstracts themselves.

In contrast to the primary journal, the majority of abstract journals are not of graphic arts quality, provided by monotype, linotype, or photocomposition. In fact, a wide range of qualities is prevalent, ranging from a low of all upper-case computer output "camera-ready" copy to the highest graphic arts quality copy produced by photo-composition. (See Figure 4-2.)

4.1.3 Contents Journals

An interesting current-awareness product entitled "Current Contents" has been successfully marketed by the Institute for Scientific Information in Philadelphia. Current Contents is a collection of the tables of contents from a wide number of primary journals in a specified field. It allows the user to "browse" through the tables of contents of a larger number of journals than he could ever conceive of subscribing to. If he is interested in reading an article, he can either obtain it from his library or request a copy from the publisher. Current Contents has proved to be an effective current-awareness service. Perhaps one of the reasons for its success is that the user is not forced to think about what he is looking for while browsing through each issue. This is the feature which distinguishes Current Contents from book form indexes, which, while sometimes considered to be current-awareness tools, are more frequently used for retrospective teaching.

4.1.4 Key-Word-in-Context Indexes

The key-word-in-context (KWIC) index utilizes key words selected from the text or title of a document along with the surrounding words of context, as index entries. The most common type of key-word-in-context index is the so-called permuted title index.

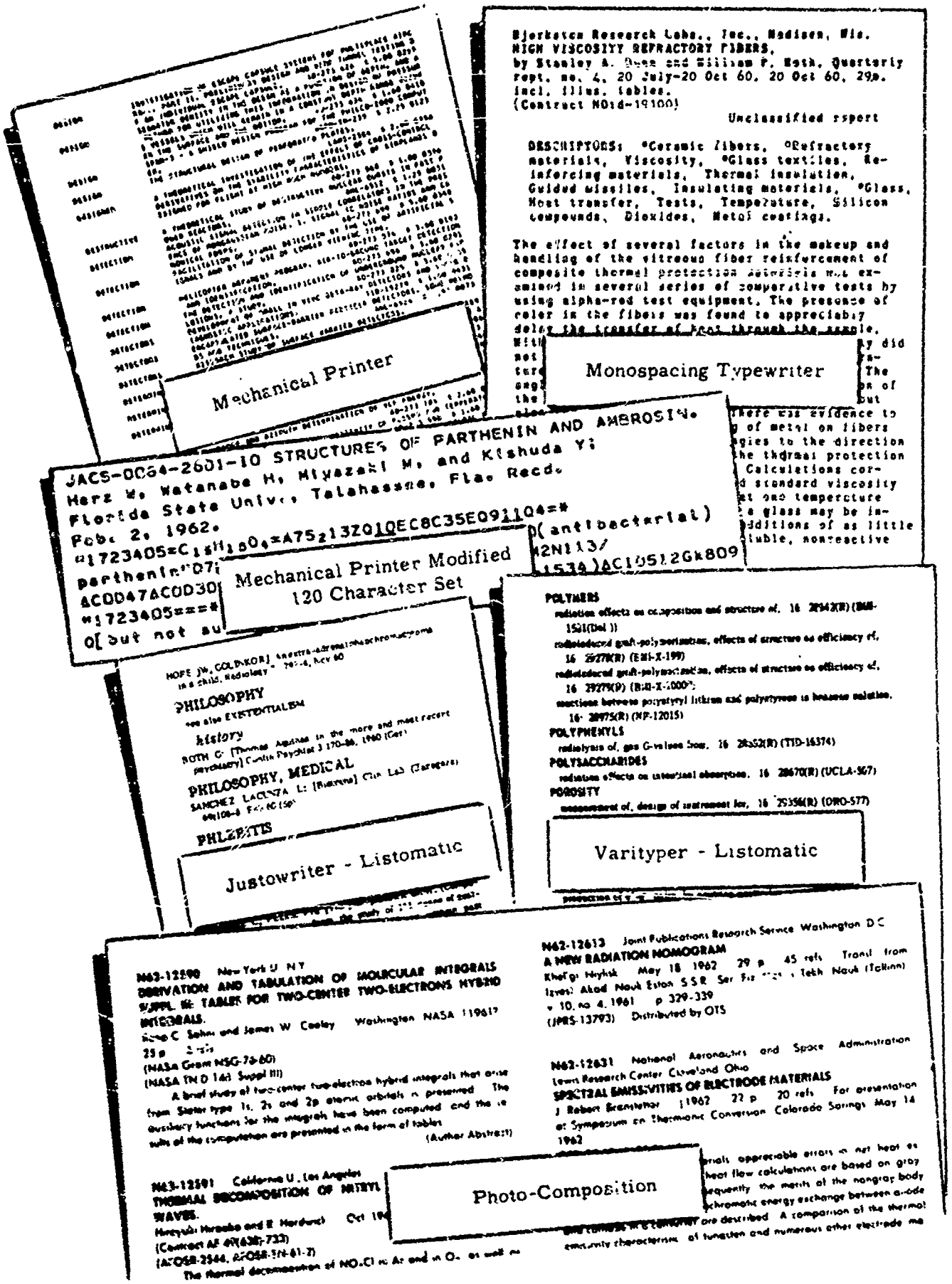


Figure 4-2. Montage of Samples Showing Range of Typographic Quality
(Taken from Libraries and Automation, Library of Congress, Washington, D. C., 1964)

which is generally produced by computer. The form of the index entry may either be the wraparound variety where those words not in immediate proximity to the key word are chopped off as illustrated in Figure 4-3 (a) or the variety illustrated in Figure 4-3 (b), wherein the key word is displayed separately.

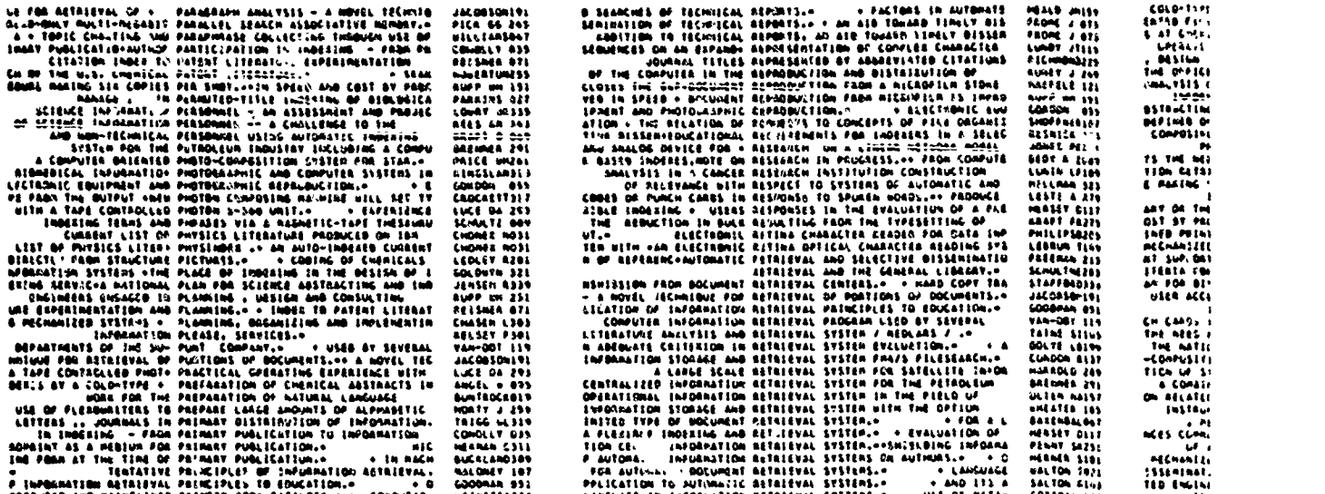
The number of KWIC indexes being produced is increasing; however, this may be an indication not of their acceptance or utility, but rather of the ease with which they can be produced. The KWIC index is useful, however, as a retrospective tool where a more refined index is unavailable or impractical to produce. One manifestation of the advent of KWIC indexing is the recognition by authors of the need for informative titles.

4.1.5 Initial Dissemination Schemes

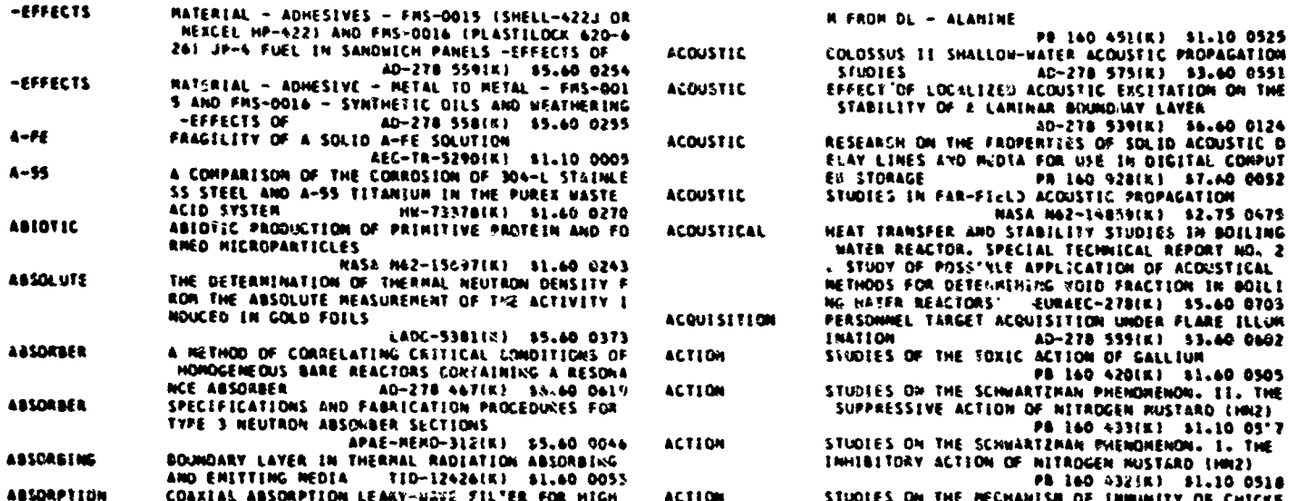
Another common form of information communication is by the automatic dissemination of printed material. These schemes may range from the formal distribution of preprints to a handful of interested recipients by the author to a highly organized system for selective dissemination of information.

4.1.5.1 Preprint and Report Dissemination. The sponsors of research and development generally will see to it that the results of research and development are disseminated. They will distribute anywhere from ten to several thousand copies to those people or organizations on a particular mailing or distribution list. This distribution may be made from the initial printing by the generator of the report or from a special reprinting made for this purpose. Generally, depository libraries and information centers are on most of the distribution lists for Government sponsored research.

4.1.5.2 Microform Dissemination. A number of Government document information centers are making initial distributions in microform. The microfilm aperture card (Figure 4-4 (a)) is the standard microform for engineering drawings. For technical reports, there will probably be a standardization on the microfiche, i.e., transparent microfilm sheet (Figure 4-4 (b)). Standards are presently being developed covering size, format and reduction ratios by the National Microfilm Association. One of the purposes of the microform dissemination program is to decentralize the production of hard-copy. Reproducibles are made available to those groups which would otherwise be likely to request hard-copy from the central document store. The distribution of these microforms is usually on a "broadcast" basis by broad categories and not based on a highly selective

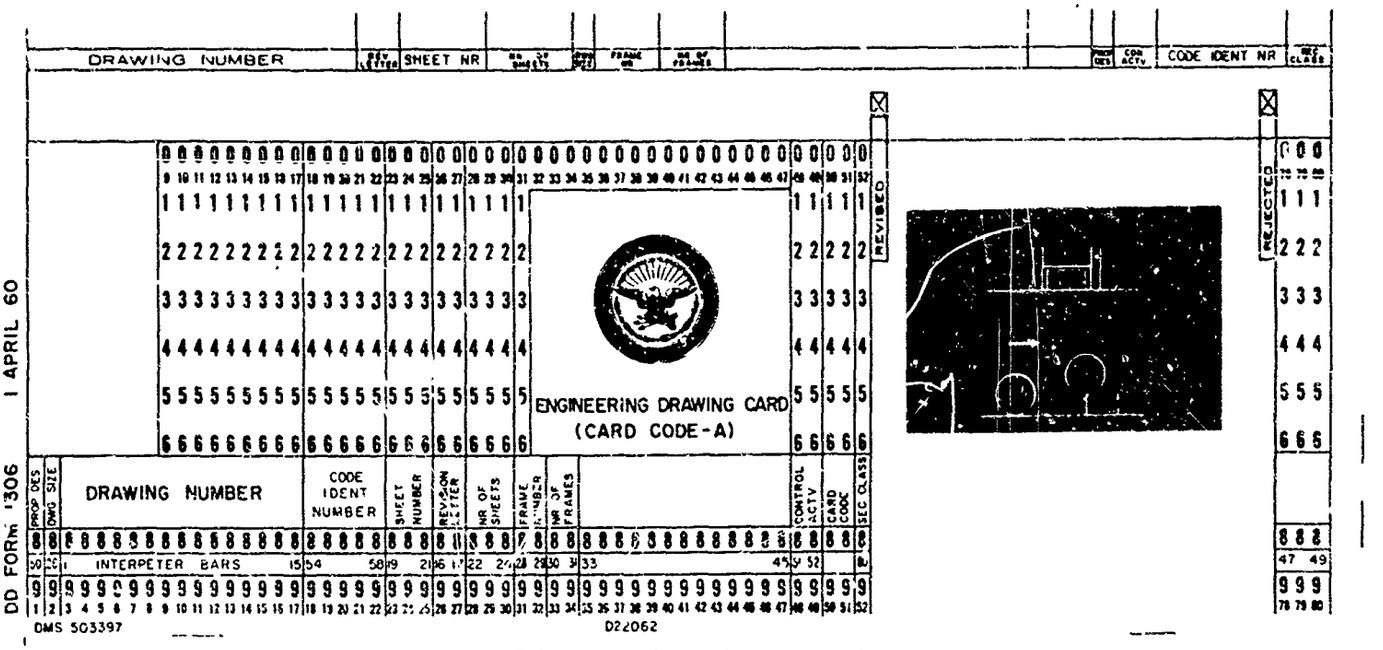


(a)

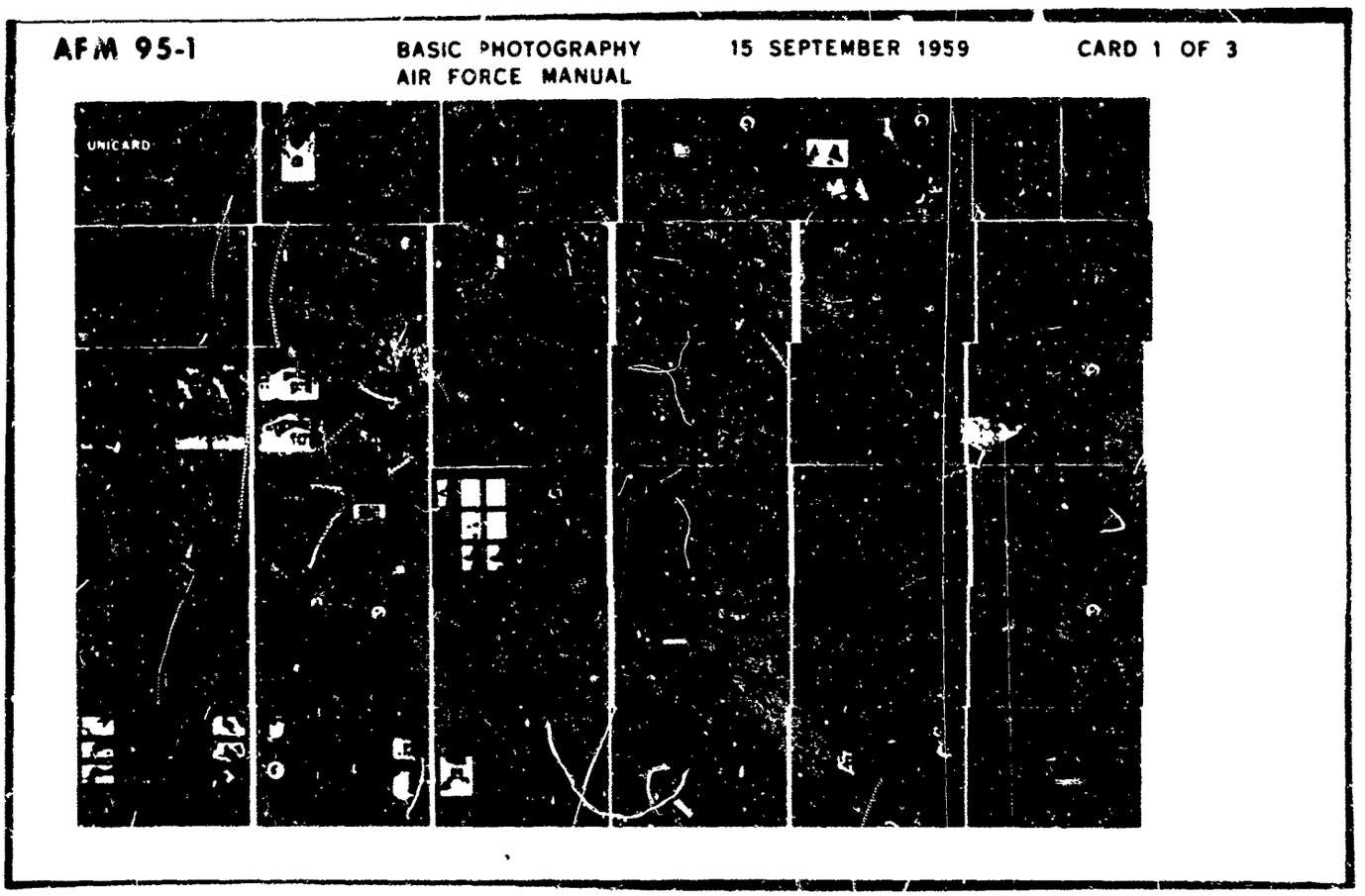


(b)

Figure 4-3. Key Word in Context Indexes



(a) Microfilm Aperture Card



(b) Microfiche

Figure 4-4. Typical Microforms

profile of the recipient's interest. A discussion of the economics of various types of microform dissemination is included in Section VIII.

4.1.5.3 Selective Dissemination of Information (SDI). A relatively sophisticated current-awareness service has come to be known as SDI for Selective Dissemination of Information. Under SDI, an individual (or an organizational component) automatically receives from an information center the announcement of all material which is relevant to his or its work. Selective Dissemination as illustrated by Figure 4-5 is accomplished by matching the index terms assigned to the current accessions of the information center against all of the index terms assigned to the interest profiles of the individuals and organizational components serviced by the center. An interest profile is prepared initially by a combination of interview and questionnaire techniques. The profile is then translated into the language of the system and may be considered as standing questions stated in Boolean terms, e.g., logical intersections, logical unions, and even logical negations of index terms. However, most SDI systems only employ logical unions. The interest profiles are generally updated to add new projects as well as to delete completed projects.

An essential ingredient to the success of any SDI system is feedback, i.e., the ability to automatically adjust the system in response to the changing experience of and evaluation by the users of the system. For example, if the system is flooding the recipient with information, a means must be provided for correcting this situation.

A common form of output from an SDI system is a tabulating card on which an abstract, a citation, or merely a title is printed. Alternatively, this output might be a card containing several pages of images, or perhaps it might be an extremely inexpensive microform containing the entire document with a few elements legible enough for the reader to evaluate its relevance. Provision is made on the SDI cards for requesting a hard-copy of the document. There is also a place for indicating whether or not the document was relevant, which closes the feedback loop to the system. Unfortunately, where this feedback aspect is not effectively policed, the users of the system may become inundated with non-relevant information. This suggests that SDI systems covering individuals may only be practicable within the confines of a single organization wherein all users are under the management of that organization. On the other hand, SDI services directed at organizational components or so-called "interest centers" might be effectively provided by information centers.

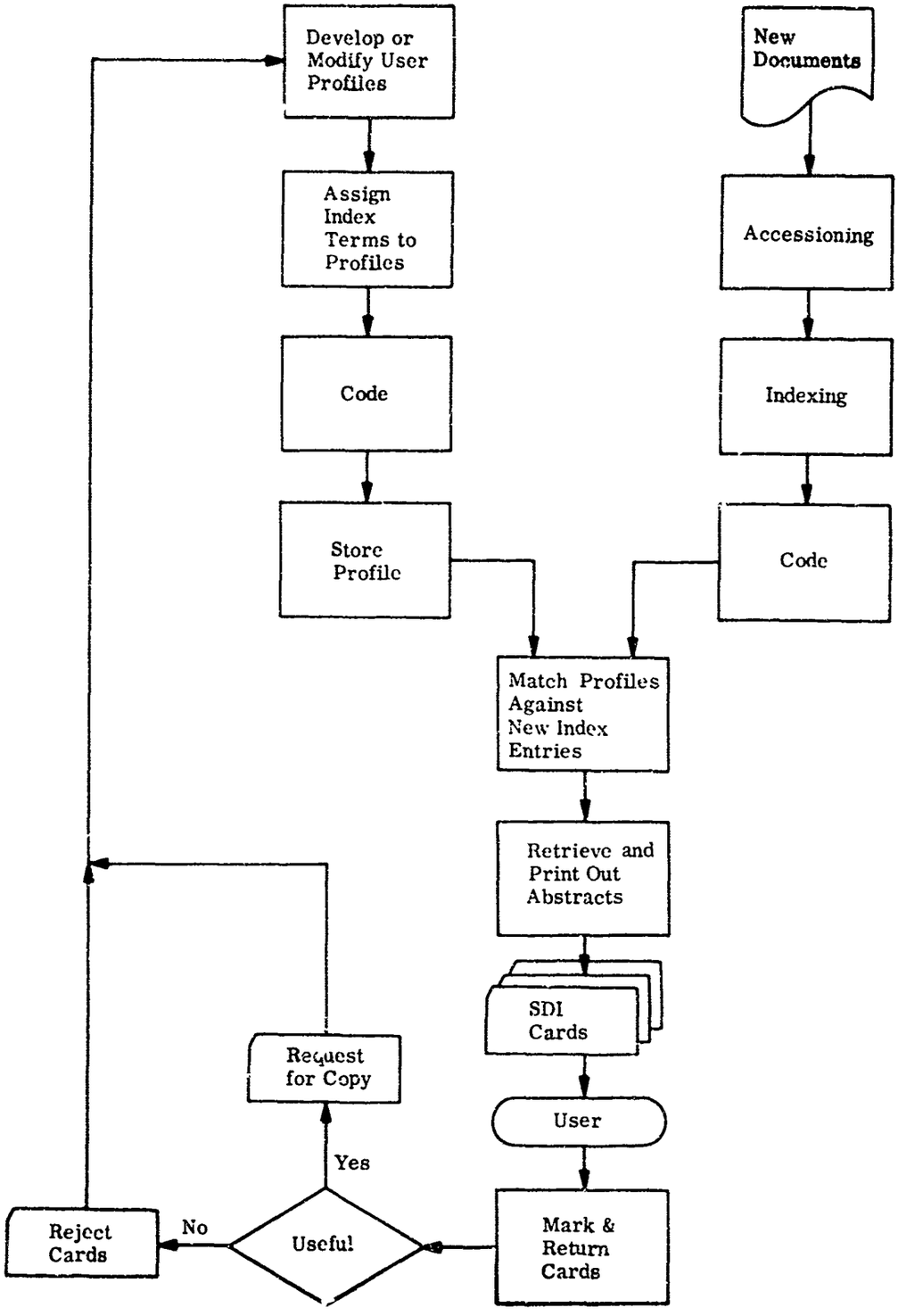


Figure 4-5. Selective Dissemination of Information

4.2 RETROSPECTIVE SEARCH SERVICES

All retrospective search services whether manual or machine-assisted must provide for three basic functions from the user's standpoint. These are:

(1) look-up (locate)

(2) look-at (examine)

(3) take-away (use, retain).

In the look-up function, the user will utilize either a manual index or an index in a machine to locate documents or, more often, surrogates of documents which may be of interest. Manual indexes may be in book form or card form. He will then have to look at these surrogates, which may be citations, titles, abstracts or even the full text, to determine whether the documents are probably relevant to his problem. These two steps, i.e., look-up and look-at, may be repeated several times. Lastly, he will want to take away a copy of the relevant documents, either on a loan basis or for retention.

With a user-controlled index, such as a book form index, the user may in effect conduct a dialogue with the index. He may readily modify his search as his deductive ability and serendipity leads him to more fruitful avenues of approach. By scanning the surrogates the user is able to screen out those items which do not appear to be relevant. Machine search systems do not generally permit a dialogue, but they can handle a much greater depth of indexing and thereby respond to more highly specific questions, and (hopefully) eliminate the need for looking at a host of non-relevant surrogates or documents.

Machine searching is intended to replace a process which now involves human intellectual effort. Its success, therefore, will depend upon the ability to separate the intellectual functions from the clerical functions. This has been accomplished in most existing machine systems by (1) having the requester (or an intermediary) formulate the question in the accepted language of the system, (2) look at and evaluate the output of the machine search, and (3) modify or reformulate the query as required. The machine system performs the routine, but time-consuming, function of matching or comparing the criteria of the question with the index entries in the file.

The take-away process is, to an extent, independent of the method of search employed. The usual first output of a machine search system is a list of citations or abstracts which the user must first look at before he determines that it is economically desirable to request a full-size copy. A major disadvantage, therefore, of the centralized machine search service is the additional delay necessitated by having to make a second request for hard-copy. Even where the user searches in a book form index, he will not be able to take away the full-size document unless his library happens to have a loan copy on hand. There are two approaches in use which tend to avoid this delay involved in the second request. The response to the first request may include a low-cost reproducible microform of the entire text which the user can throw out if not relevant. Another approach is to provide complete microform collections at little or no cost to the libraries of the user organizations. Such collections can be "complete" relative to a specific field of interest. These collections decentralize the physical access or take away problem and reduce the turn around time required to supply copies to the user.

4.2.1 Book Form Indexes

One means of decentralizing the search or look-up function is by publishing the retrieval service in book form. Published retrospective search services face the serious problem of bulk because of the relatively large volume of accumulated information (rather than merely current information) which must be accommodated. Published retrospective services must either be of great bulk and cost, or else reduce index depth, or provide only non-informative index entries, such as accession numbers. Usually, a compromise of some sort is reached, but the basis for making each such compromise is neither apparent nor objectively justifiable, in light of the general lack of knowledge of true users' needs. Book form indexes frequently include subject indexes, author indexes, report number indexes, and source (issuing organization, journal, etc.) indexes.

Book form subject indexes may be arranged in a variety of fashions. They provide varying depth of indexing, and can provide informative or non-informative surrogates. There is generally a trade-off between the depth of indexing and the size of the surrogate. If an item is indexed by 20 descriptors, it will be listed 20 times. The space required for this multiple listing is made available by making each entry much shorter. In a book form index, if the depth of indexing is increased, the size of the surrogate is generally decreased. Otherwise, the cost and bulk of the index would increase in proportion to the depth of indexing.

4.2.1.1 Subject Heading Index. Subject heading indexes generally have a depth of indexing of from one to three index entries per item. The subject headings are generally selected from an authority list, such as Subject Headings Used In the Dictionary Catalogs of the Library of Congress. The subject headings are generally multiple word terms and frequently have modifiers. A multiple term subject heading may be considered to be a precoordination of index terms, as distinguished from coordinate indexes which allow the searcher to coordinate terms in framing his query.

4.2.1.2 Uniterm Index. The Uniterm index is one embodiment of the concept of coordinate indexing, which was discussed in the previous section. While a Uniterm index is more difficult to use than other coordinate indexes, its prime advantage is its relative ease of publication. (See Figure 3-3). A searcher merely pulls the card for each of the terms in his inquiry and visually matches the accession numbers posted under each term card. Where the same accession number is posted under two terms, there is a logical intersection. The search process is simplified by arranging the postings from left to right on the page by their terminal digit so that all sevens will be in the same column, all sixes, etc. Another popular form of printed Uniterm index is the double dictionary which contains two identical book form indexes in a loose leaf binding. In this fashion, it is possible to view two different pages of the book at the same time and make a visual coordination.

4.2.1.3 Tabledex Index. Another book form coordinate index is the so-called Tabledex index illustrated in Figure 4-6. Under the Tabledex method, you first determine the number of items (frequency) posted under each term by referring to the list in Figure 4-6 (a). You then select the lowest frequency number and search the table in Figure 4-6 (b) for those items that contain all the codes in the question. While this can be more efficient a method of searching than the Uniterm book form index, its use has not become widespread because its method of use is more difficult to understand.

4.2.1.4 Citation Index. Another type of retrospective index which has existed in the legal profession for nearly a century and is just beginning to appear in the scientific and technical fields is the citation index (See Figure 4-7.) For each article a citation index gives a record of all the later articles which have cited it. In law, this is applied to court decisions and is particularly important because of the importance of precedent.

1.001 ABLATION
1.001 ABSORPTN
1.002 ACAD SCI
1.002 ACOUSTICS
4.001 AERODEX
4.002 AFRICA
1.003 AIR FORCE
24.001 AIRCRFT
12.001 ALASKA
5.001 ALBEDO
2.001 ALERT
2.002 ALPHA RAY
10.001 ALTITUDE
1.004 AMATEUR
2.003 ANDES
1.005 ANGLE
98.001 ANTARCTIC
2.014 ANTENNA
56.001 ARCTIC
2.005 ARGENTINA
2.006 ARGUS
1.006 ARMY
1.007 ARNEB
1.008 ARTIC
2.007 ASCAGRAPH
3.003 ASIA
11.001 ASTRONOMY
10.002 ATLANTIC
1.009 ATLANTIS
119.001 ATMOSPHR
11.002 ATOMIC
70.001 AURORA
12.002 AUSTRALIA
2.008 AUTOMATIC
1.010 AVIATION
3.004 BALLISTIC
17.001 BALLOON
2.009 BATTERIES
1.011 BEAUFORT
3.005 BELGIUM
1.012 BERMUDA
6.001 BIBLIOG
5.002 BIOLOGY
3.006 BLUE GLAC
3.007 BOLIVIA
11.003 BOOK
1.013 BRAVO
1.014 BREID BAY
1.015 BROOKS RG
2.010 BRUNNEAR
1.016 BRUSSELS
4.003 BURST
4.004 BYRD
9.001 CALIBRATN
1.016 CHIZHMATA
6.003 DUST
21.002 DYNAMICS
11.001 EARTH
7.003 ECLIPSE
36.002 EL CENTRO
47.001 ELECTRONS
5.011 ELLSWORTH
11.006 EMISSION
18.001 ENERGY
12.001 EQUATOR
4.004 EQUIPMENT
1.018 EROSION
2.014 EXCAVATN
6.002 EXOSPHERE
29.002 EXPEDITN
4.005 EXPLORER
7.004 EXPLORER
1.019 EXPLORER
4.007 EXPLORER
1.020 EXPLORER
2.015 EXPLORER
10.004 EXPLOSION
11.002 FACILITY
5.007 FADING
3.013 FEDCHENKO
26.001 FIELDS
1.021 FILCHNER
5.014 FINLAND
24.002 FLARES
2.016 FLETCHER
2.017 FLOWMETER
4.006 FORECAST
21.003 FRANCE
2.018 FRITZ PK
21.004 FT CHURCH
5.008 GAMMA RAY
62.001 GENERAL
23.001 GEODESY
11.003 GEOLOGY
102.001 GEOMAGNET
1.042 GEORGIA
14.001 GERMANY
46.001 GLACIERS
2.019 GOETZ
6.005 GRADIENT
1.043 GRAHAM LD
30.001 GRAVITY
10.005 GREENLAND
4.008 GRENADES
21.005 GT BRIT
1.044 GUAM
10.007 HANILE
14.002 HAPS
3.019 HAWSON
3.020 HICALL
2.027 HCMURDO
189.001 HEASUBMAT
13.005 HECHANICS
9.010 HEDICINE
31.002 HEETING
4.013 HESONS
9.011 HESOSPHR
91.001 HETEOROL
27.003 HETEORS
1.064 HEUDON
2.028 HICROCOMO
2.029 HICROLOGY
4.014 HICROWAVE
4.015 HINNEAPLS
1.067 HINNESOTA
6.006 HOON
1.068 HOONWATCH
8.004 HOUNTAINS
1.069 HOUSE
1.070 HT AFTER
2.030 H.AMERICA
2.031 HAT CORN
2.032 HAVIGATN
9.007 HAVY
13.006 HEUTRONS
7.007 HEW ZEAL
1.071 HIGERIA
7.008 HIGHT
5.012 HIGHT SKY
1.072 HIKE
1.073 HITROGEN
5.013 HOCTILCNT
27.004 HOISE
2.033 HORFASLD
2.034 HORWAY
1.074 HUCLEAR
2.035 HUCLEI
1.075 OBLATE
81.001 OBSERVATN
16.002 OBSERVTRY
27.005 OCEAN
34.001 OCEANOG
1.076 OLYMPIC
1.077 OLYMPUS
6.005 OPTICS
21.006 ORBIT
1.078 ORIONIDS
19.001 OSCILLATN
1.079 OSIRIS
9.008 OXYGEN
36.003 SEISMOL
11.008 SFERICS
1.095 SHAPE
15.006 SHIP
1.097 SHOWA
1.098 SILICON
1.099 SIMEONEG
2.064 SKY
5.015 SOFLAMS
8.006 SNOW
1.100 SODIUM
101.001 SOLAR
13.002 SOUND
24.002 SOUNDING
2.066 SOUTH
3.021 SPACE
1.101 SPAIN
23.003 SPECTRA
4.017 SPITZBRGN
2.067 SPUTNIK
8.007 SPUTNIK 1
4.018 SPUTNIK 2
8.008 SPUTNIK 3
1.102 S ANDREW
1.103 STARS
17.001 STATIONS
19.003 STORMS
4.019 STRATIG
8.009 STRAT.SPH
6.010 SUMMARY
4.020 SUNSPOTS
5.015 SWEDEN
5.016 SWITZ
1.104 SYMBOLS
4.021 T-3
2.068 TADZHIK
1.105 TAMANRAS
1.106 TASMANST
14.004 TELEMETRY
2.069 TELESCOPE
38.004 TEMPERATR
25.002 THERMAL
4.022 THERMOSPH
1.106 THICKNESS
1.107 THULE
1.108 TIBET
16.004 TIDES
9.011 TIME
1.021 TIROS
8.010 TIROS 1
1.109 TIROS 2
1.110 TIROS 3
1.111 TOKYO

(a) List of Retrieval Words With Frequency Codes

5.014-7.005

5.014(CONTINUED)

0665	21.005	105.001	119.001	132.001	161.001
0664	91.001	105.001	119.001	132.001	161.001

5.015

0148	5.016	16.001	56.001	62.001	
0640	5.016	28.002	29.002	56.001	62.001
0640	5.016	29.002	56.001	161.001	
0282	15.006	23.005	81.001	185.001	
0634	31.002	56.001	132.001	161.001	

5.016

0153	10.005	14.001	21.003	46.001	
0148	16.001	56.001	62.001		
0682	21.001	21.005	22.001	23.003	66.001
		81.001	161.001	199.001	
0690	28.002	29.002	56.001	82.001	
0640	29.002	56.001	161.001		

5.017

0886	6.002	23.001	29.003	68.001	98.001
0777	6.001	22.001	28.002	91.001	56.001
		63.001	91.001		
0406	18.001	21.002	23.001	101.001	119.001
		199.001			
0725	30.002	91.001	119.001		
0669	63.001	119.001			

5.018

0613	19.001	24.002	37.001	98.001	132.001
		199.001			
0590	31.002	62.001			
0636	31.002	132.001	161.001		
0601	31.002	132.001			
0604	98.001	161.001			

6.001

0432	13.007	22.001	23.002	26.001	70.001
		81.001	119.002	185.001	
0493	16.001	189.001			
0581	56.001	62.001	98.001	161.001	
0280	62.001	84.001			
0285	62.001				
0477	189.001				

6.006(CONTINUED)

0835	23.001	81.001			
0718	50.001	70.001	81.001	98.001	

6.007

0771	7.003	8.001	56.004	67.001	68.001
		101.001	132.001		
0266	7.005	11.006	13.004	23.003	26.001
0756	7.011	17.001	36.002	70.001	81.001
		102.002	119.002	185.001	
0413	9.009	10.001	13.008	189.001	
0549	10.001	10.003	13.004	39.002	70.001
		119.002	185.001		
0783	12.005	13.006	15.001	16.002	81.001

6.008

0587	6.010	10.003	10.009	16.002	27.003
		39.002			
0678	13.007	29.002	161.001		
0033	13.007	62.001			
0138	29.002	56.001			
0451	29.002	161.001			
0440	36.003	91.001			

6.009

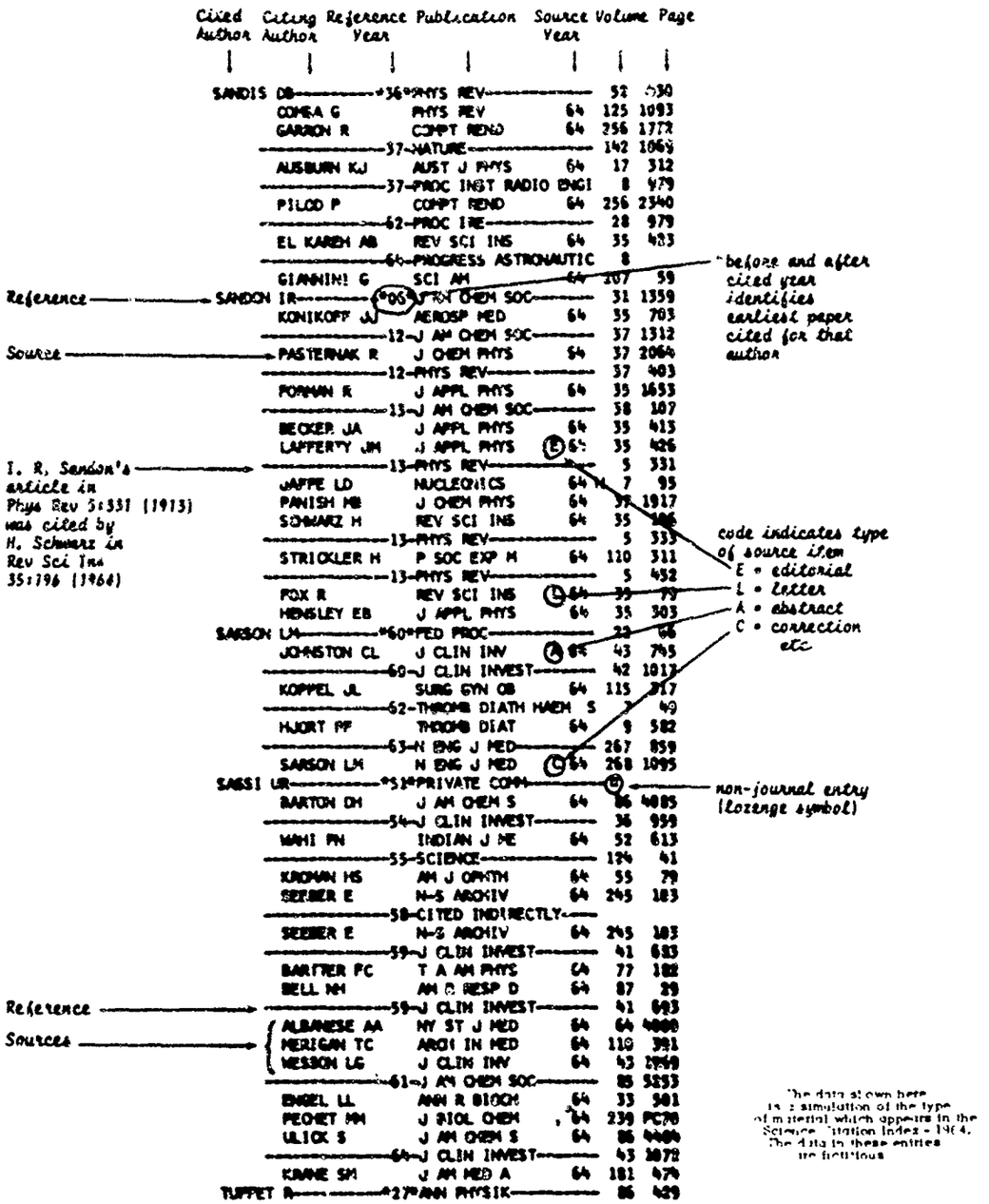
0781	7.002	14.003	35.001	98.001	119.002
		132.001	169.001	199.001	
0717	7.005	9.001	11.005	26.001	29.003
		38.002	70.001	119.002	
0782	9.001	13.009	14.001	67.001	68.001
		105.001	119.002	132.001	185.001
		199.001			
0326	10.004	36.002	50.001	105.001	119.001
0392	21.002	23.003	132.001		
0482	31.002	189.001			

6.010

0663	10.002	10.009	15.004	29.002	38.001
		84.001	161.001		
0587	10.003	10.009	16.002	27.003	39.002
0676	21.005	39.001	56.00'	161.001	
0677	21.005	105.001	132.001	161.001	
0627	66.001	105.001	161.001		
0585	81.001	84.001	119.001	132.001	

(b) Frequency Coded Tabledex Tables

Figure 4-6. Tabledex Index



To locate sources which cite a particular paper, look first for the cited or reference author which is located on the left. For each cited paper by that author there is a dashed line which continues to the column reserved for the year of reference publication, followed by journal, volume, and page. When a given reference has been cited more than once, the sources are arranged alphabetically. Each type of source item is further identified by a code letter.

Figure 4-7. Citation Index

The citation index is based on the assumption that the author of a document is best qualified to define other material relevant to his document, and that the author defines the relevant material by means of citations or references. The citation index is organized in the following fashion:

(1) The names of authors within the document universe are arranged alphabetically.

(2) Each author's name is followed by references to documents he has written (in chronological order).

(3) Each document is followed by references to documents (listed in chronological sequence) which have cited the original document.

The index is used by selecting an author (or an article by an author) which is known to be related to the desired subject area. The index provides a list of documents which cite the author (or article) and bring the searcher forward in time to the most recent documents which bear on the subject area.

The advantages of a citation index are:

(1) A citation index points to the latest published documents in a subject area (descendants) whereas most bibliographies point back in time to older work (antecedents).

(2) Entry into a citation index can lead to a variety of related but different fields, not bounded by a narrow discipline. For instance, an article about a dye indicator might be cited by articles on biochemical studies using dyes, medical research, or clinical diagnosis of patients.

(3) A critical review of a document (or an author) can be obtained by seeing who has cited the document (or author) in the past, and evaluating what was said.

(4) A citation index requires no logic or intellectual involvement. The index uses only the citing authors' sense of relevancy. Thus, the citation index avoids the confusion which surrounds other indexes and classifications.

(5) A scientist, for example, is kept up to date by a citation index on who is using his material or on who is interested in the same subject area.

Some of the disadvantages of a citation index are:

(1) The one-time costs for producing a citation index are extremely high. Aside from programming costs, a vast number of references must be keypunched before a suitable index evolves. For instance the Institute for Scientific Information had to keypunch the references for all the world's scientific literature for 1961 in order to produce a citation index for the field of genetics. Once the references were analyzed, only five percent of the total references were found to be pertinent to the field of genetics. Consequently, citation indexes will tend to be produced centrally by a publisher and will cover broad fields, e.g., Law, Medicine, Science. They will not be a primary tool for a local system i.e., for locating internal company documents.

(2) The relevancy of the interconnection between the articles and the citations is entirely dependent on the author's judgement, which may be bad. The author may cite erroneously, incompletely, or excessively.

4.2.2 Card Form Indexes

Card form indexes have been utilized for many years, primarily because of the ease in sorting, merging, rearranging, and updating unit record files as compared to book form files. One of the primary difficulties with card form index systems is that each set must be separately maintained, and consequently, their use is generally centralized. In contrast the book form index can be published and widely utilized on a decentralized basis, although maintained centrally. The primary disadvantage of the book form index is its inability to be updated except by supplement or re-publication.

The most familiar form of card form index is the library catalog card file which is frequently referred to as the dictionary-catalog or sometimes as the catalog-index. A number of new mechanical techniques have been developed to make the unit record more flexible. Basically, these fall into two categories: edge-notched cards and interior punched cards.

4.2.2.1 Edge-Notched Cards. Edge-notched cards permit the selective sorting of records which fall into a single category or the selection of records which fall into a specified combination of categories (logical intersection). This is accomplished by skewering a deck of cards with a needle which will allow those cards which have been notched to drop out, the

needle retaining all of the unnotched cards. Because of the limited number of holes available on a card, various coding schemes are employed for representing information. The most efficient of these, from the standpoint of effective utilization of available hole positions, is the random superimposed coding. By this method several items of information may be superimposed in the same coding field.[33]

4.2.2.2 Interior-Punched Cards. The most commonly utilized form of punched card is the EAM or Tabulating Card. These cards are usually machine sorted or collated on electromechanical sorting equipment. They may, however, be utilized as an optical light coincidence (peek-a-boo) card. A more common form of peek-a-boo card is the Termatrex Card produced by Jonkers Business Machine Company. Peek-a-boo cards are utilized for coordinate indexes in inverted file arrangement. There is a card for each term in the vocabulary. There is a dedicated space for each document in the collection on each card. If a document has been indexed by a given term, a hole will be punched in the term card in the dedicated space for that document. Logical searches are made by superimposing term cards for each term in the question and shining light through them (see Figure 3-2). Each coordinate where light shines through represents a particular document which has been indexed by each of the terms in the question. Devices are available for conveniently reading these coordinates and even for printing out the results. It should be noted that, since peek-a-boo cards are generally utilized in the inverted file mode, they are never combined with abstracts or other surrogates. On the other hand, edge-notched cards, when not used in the inverted file mode, frequently have some form of surrogate recorded in the interior.

4.2.3 Machine Search Services

Nearly all of the machine search techniques involve some form of "concept coordination" or "coordinate indexing." As described above, coordinate indexing can be accomplished by manual techniques including book form indexes and card form indexes. Machine searchable coordinate indexes are necessary and desirable however, when any or a combination of the following characteristics is required: deep indexing (an average of ten or more index entries per item), relatively large collections (over 25,000 items), logical capability (the ability to perform logical unions, intersections, negations, differences, of descriptors, etc.) of fast retrieval, and high degree of completeness.

Some retrospective search systems can provide an approximate or actual count of the number of responses there would be to a query before the search is actually made. This becomes particularly important where the file is large, and the output may be lengthy. The following paragraphs describe the various outputs which may be produced by a machine search system.

4 2.3.1 Search to Retrieve Document Number. The simplest form of machine search output is a listing of the accession numbers of all documents which meet the criteria of the search question. The file consists of index terms and document accession numbers. The file may be stored on any medium the system requires. The output media may be an EAM card, a printout from a computer printer, or a console typewriter.

A list of document numbers is seldom an adequate product for delivery to the user The document number is not a useful surrogate to look at to determine whether to obtain the full text of the document. This form of output is most commonly utilized as a picking list, i.e., for pulling copies of catalog cards, abstract cards, microfiche copies of reports, or full-size copies of documents. Where decentralized files exist, the list of accession numbers may be utilized by the user to find a microform or full-size copy of the document to view.

4.2.3.2 Search Producing Citations or Abstracts. A more useful output of a mechanized search system, from the user's viewpoint, would be a special bibliography containing complete bibliographic citations and possibly abstracts of all documents meeting the specific terms of the search query. There are many more mechanized systems which produce bibliographic citations only than produce citations with abstracts. This is particularly true in computer systems. The abstract generally requires at least 1,000 characters and cannot be economically stored in a computer system. If an inverted file arrangement is utilized, it is necessary to maintain a separate file of abstracts and citations in accession number order, even if mechanized. Therefore, the search would first produce a list of document numbers which answer the search inquiry and would store the list in memory. Then the citations and/or abstracts would be extracted from a separate file. Even if a linear file arrangement is utilized, it may still be more practical to separate the citations and/or abstract file so that the logical search itself can be made efficiently (see Paragraph 9.3). If a mechanized search is to be supplemented by a manual picking of a

catalog card or abstract card, the value of mechanization is diminished. The value of manual picking can be increased, if the output of the manual picking process is a visible abstract card plus the entire contents of the document in a reproducible microform. This would eliminate the necessity for the user to make a second request of the system for the complete text. In other words, he would have the abstract card to look at for evaluating whether it is worth reading the document. He would also have a microform of the document to take away. If the document is not relevant, he can merely throw the card away or perhaps return it to the system.

4.2.3.3 Search Producing Document Display. With the three elements of all retrospective search systems, i.e., look-up, look-at, and take-away in mind, it is evident that systems which directly provide a display of the document (or a substantial surrogate such as an informative abstract) serve both the look-up and look-at functions. This implies that the user will be in direct communication with the system, either sitting at a remote console device or at the main console of the system itself. From the user viewpoint, a system which provides an immediate response in the form of display would be highly desirable. He could immediately look at the display and make the determination as to whether or not he would like the document. He might even modify the search question as the search proceeds.

4.2.3.4 Search to Provide Microform or Hard Copy. The direct output of some search systems is microfilm (e.g., Rapid Selector or FileSearch). More frequently, microfilm may be the form of response from the information center to the user, while the output of the search sub-system may be merely a list of document numbers. The reasons for the separation of index files from the document files are discussed in Paragraph 9.3.

The ultimate product desired by the user of a document retrieval system is generally a hard copy of all documents relevant to his query. Because of the difficulty in adequately framing a query, the first output of a system is generally a surrogate of a document rather than a hard copy.

4.2.3.5 Search to Produce Data Correlations. The product of a fact retrieval type of information system is the answer to the specific inquiry rather than a document or series of documents containing the answer. When dealing with linguistic and qualitative information, this is frequently too difficult to achieve, especially by mechanized techniques. However,

when dealing with quantitative or near quantitative information, the data correlation is readily achievable, particularly due to the limited vocabulary and relative lack of ambiguity which prevails. The output of such a search might be a man's name, a list of names, a chemical process, the physical or chemical properties of a particular element or compound of matter, or perhaps the credit rating of a particular individual or organization.

SECTION V. INFORMATION STORAGE AND RETRIEVAL SYSTEM FUNCTIONS

5.1 GENERAL

This section discusses important functions common to many information systems. The discussion is not intended to constitute an "operating manual" or a listing of all factors, or even a detailed description of all important operations. Rather, it concentrates on facets of each operation which are critical, or on those facets which are often overlooked in information system design and operation. All information products and services described in the preceding sections are produced by some combination of the following system functions.

5.2 BASIC FUNCTIONS

There are eight basic functions of IS&R systems from which all such systems can be assembled:

(1) Origination (including initial publication) of an information item.

(2) Acquisition and/or selection and evaluation of informational items for use or for input into current-awareness, announcement, and/or retrieval systems into correlation processes.

(3) Surrogation including indexing and/or abstracting.

(4) Announcement of informational items.

(5) Index Operation including recording of index information into a physical medium and the searching of that medium to provide an output of references and/or other item surrogates.

(6) Document Management including storage, retrieval (based upon item addresses only), reproduction, dissemination, inventory control, etc., of documents.

(7) Correlation of many informational items.

(8) End-Use information.

5.3 BASIC SYSTEMS

The basic functions can be combined into 10 basic types of systems as follows:

System 1

Origination
End-Use

Examples:
Preprints
Telephone calls
Memoranda

System 2

Origination
Acquisition
End-Use

Example:
Journal subscriptions

System 3

Origination
Acquisition
Surrogation
Index Operation
Document Management
End-Use

Example:
Science Information Exchange

System 4

Origination
Acquisition
Surrogation
Index Operation
Document Management
Correlation
End-Use

Example:
Advanced State-of-the-Art Technology Group of Institute of Applied Technology, National Bureau of Standards

System 5

Origination
Acquisition
Surrogation
Announcement
End-Use

Examples:
Chemical Abstracts
Applied Science and Technology Index

System 6

Origination
Acquisition
Surrogation
Announcement
Document Management
End-Use

Example:
Engineering Index and Engineering Societies Library

System 7

Origination
Acquisition
Surrogation
Announcement
Index Operation
Document Management
End-Use

Example:
Defense Documentation Center, NASA Scientific and Technical Info. Facility

System 8	System 9	System 10
Origination Acquisition Surrogation Announcement Document Management Correlation End-Use	Origination Acquisition Surrogation Announcement Index Operation Document Management Correlation End-Use	Origination Acquisition Correlation End-Use
Example:	Example:	Example:
TIROS Data System	Specialized information centers with retrospective stores.	Use of complex information in a project from multiple sources.

Figure 5-1 illustrates how the eight basic functions are interconnected to provide the 10 basic systems. The solid lines indicate the flow pattern of information systems generally in that they subsume all 10 systems. The numbers on the arrows correspond to the system types.

Interfaces between information systems can occur, bringing about in effect, loops in the above pattern. Two principal types of interfaces can exist: first type, indicated by a dotted line, may be termed reorigination; second type, indicated by a dashed line, may be termed reacquisition.

Reorigination is said to occur when new information is created as an end in itself, via the correlation function. A state-of-the-art report is an example.

Reacquisition is said to occur when one system acquires documents previously acquired, processed, and announced or available from another system.

5.4 INTERACTIONS AMONG FUNCTIONS IN SYSTEMS

Although there are only eight basic functions and 10 basic systems, the details of the functions are subject to significant variations when they are combined to form different systems. For example, the details of the surrogation function suitable for System 6 (origination, acquisition, surrogation, announcement, document management, and end-use) would be entirely unsuitable for use in System 7 (origination, acquisition, surrogation, announcement, index operation, document management, and end-use). The

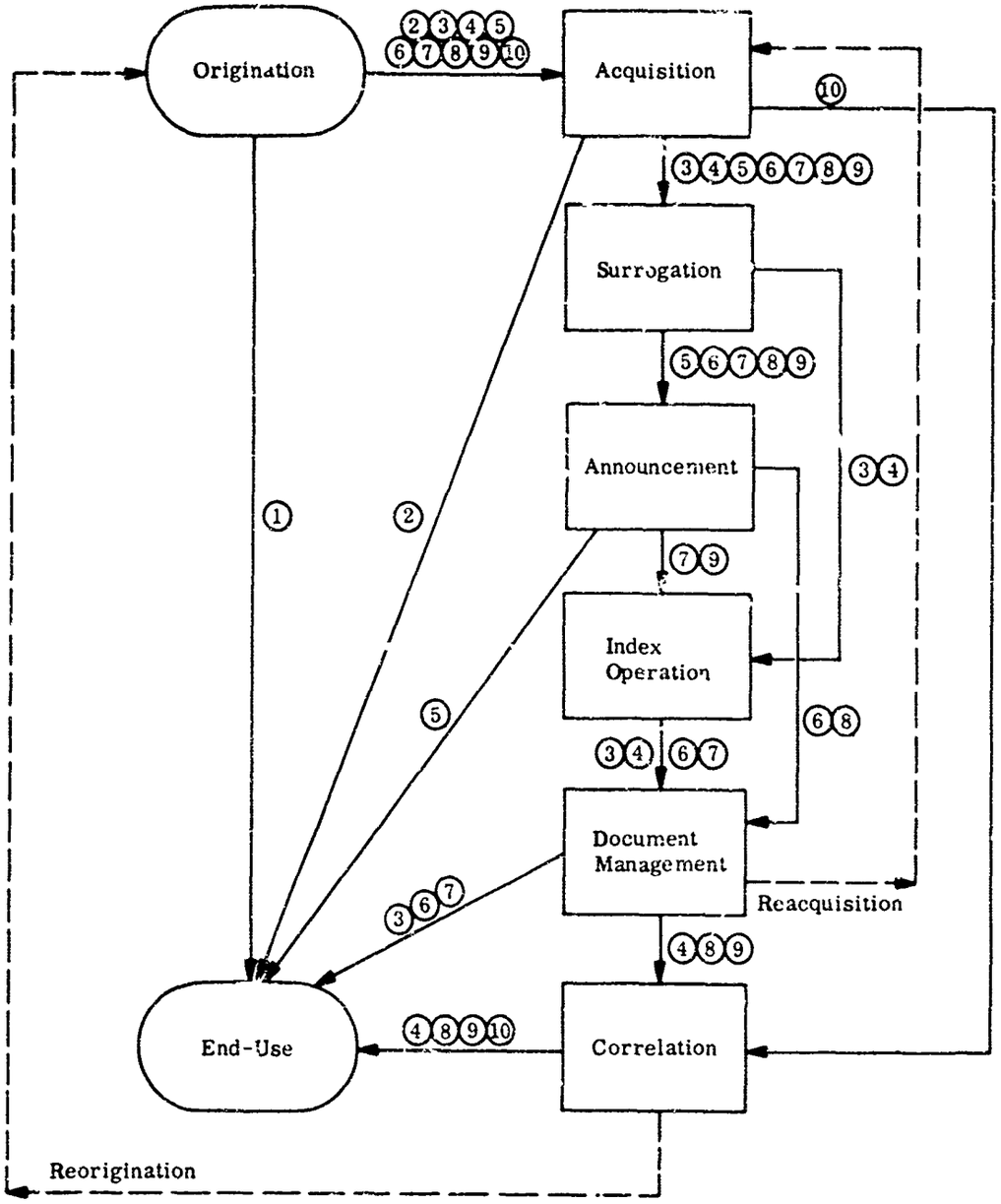


Figure 5-1. Possible Combinations of IS&R System Functions

difference lies in that the surrogation function in System 7 must provide not only for an announcement function (for which "shallow" indexing would be adequate), but also for a retrospective retrieval function (for which "deep" indexing is required).

Hence, each function must be considered in relation to the other basic functions with which it must operate as a system.

5.5 DESCRIPTION OF BASIC INFORMATION STORAGE AND RETRIEVAL SYSTEMS

5.5.1 Origination

Origination involves the publication of information in some recorded form. A publication may be intended for general or private distribution and the information contained therein may be of ephemeral or lasting value.

There are two basic forms of general publication: the serial literature and the separate literature. The separate literature (e.g., books and monographs) was typically used for rather complete and extensive works of lasting value; whereas the serial literature (e.g., journals) was utilized for less complete or extensive communications of a generally ephemeral value and was segmented by specialized fields of interest. About two decades ago, the technical report, a new form of separate literature, became a major medium of mass communication. The technical report has a characteristic of often being more extensive than an article in the journal literature. It has the disadvantage of not being subject to evaluation and critical review prior to publication. It also has the disadvantage of not being effectively distributed since the initial distribution is in the hands of the originator (and what distribution list can possibly encompass all potentially interested readers) rather than in the hands of the potential readers, e.g., the freedom to subscribe to a journal. The effective dissemination of technical reports, therefore, is tending to be left to the secondary distribution function provided by information centers through initial dissemination, announcement, and request copy fulfillment, which functions are discussed on the following pages. The production operations involved in the publication process normally involve editing, re-drafting, typesetting, printing, collating, binding, and distribution. The typesetting function is undergoing radical technological change. The employment of computer-assisted typesetting and electronic photo composition is likely to bring about significant changes in the organization and methodology employed in the production of existing primary journals.

5.5.2 Acquisition

Most acquisition, evaluation, and selection processes assume that the documents are generated external to the organization being served by the information system. Externally-generated documents are important. Much can be done to ensure adequate coverage of externally-generated documents in the technologies of interest. It is sometimes not realized that the capture of valuable internally-generated information may be as difficult as the acquisition of externally-generated information. Modifications in established acquisition procedures are often required to acquire effectively the internal information.

Not all documents acquired are sufficiently valuable to warrant processing into the system. Documents having the requisite value will be selected for input processing. Determination of the requisite value is quite important and is usually an intellectual process. The value depends upon a number of factors, particularly upon the type and completeness of services to be provided to the users and also upon budgetary and staff limitations.

One process often overlooked by those not experienced in operational information systems is the handling of duplicates during the selection process. "Duplicate-checking" can be quite time consuming and expensive, although several efficient and effective techniques are known. These techniques usually involve the checking of author index card files either before or after the cataloging of the inputs (depending upon the percentage of duplicates encountered).

Usually every document is given a unique number for accounting control purposes prior to evaluation and selection. Yet if many documents are not selected, it may be necessary to convert the originally-assigned number to a permanent accession number or to provide cross-indexed files from temporary to permanent numbers. The alternative is to use the first assigned number as an accession number and keep track of those numbers which have been eliminated in the selection process. The choice between these alternates depends upon a number of factors and the wrong choice can result in higher costs, performance delays, and in errors in the future.

5.5.3 Surrogation

Surrogation is the process of substituting for the document or item something which will represent it or stand in its place for various purposes. Surrogation may include one or more of the processes of cataloging, abstracting, and indexing. A surrogate may be an accession number, a title, a citation, an extract, or an abstract.

5.5.3.1 Cataloging. Information systems generally require that a cataloging function:

(1) Assign permanent accession numbers and/or class numbers (often more than one class number, depending on the form of current-awareness service to be provided to the clientele).

(2) Record titles, authors, and sources in a specified fashion.

(3) Control the source name vocabulary (e.g., is it ESSO? Humble? FTC? etc.).

(4) Note certain characteristics of the documents (e.g., it contains graphs, bibliographies, data, etc.).

This information constitutes part (or sometimes all) of the surrogate of the document, said surrogate to be employed for many purposes later in the information system processes. The term cataloging as used by librarians also includes some of the functions labeled as indexing in this report. This is probably due to the fact that the physical form of a library index is a card catalog.

5.5.3.2 Abstracting. If abstracts are required, a choice must be made between accepting "author" abstracts and creating them. With respect to those abstracts which must be created, a spectrum of choices of types is available, ranging from informative condensations or summaries of the document (so that the summaries can often be used in lieu of the documents) to merely indicative single-phrase notations of content. The choice of abstract type will usually represent a compromise between what the clientele would consider ideal and what the information system can afford.

5.5.3.3 Indexing. Just as the type of abstract required is determined by the needs of the clientele, so also are the depth, type, and arrangement of indexes. Index entries may properly consist of classification notations (e.g., Dewey Decimal) for some purposes (particularly when the information being indexed is quantitative in nature), but for most purposes, index entries consist of terms (or subject headings) of an adequate range of specificity -- terms standing for concepts ranging from the general to the very specific.

Many environments require that index entries provide indications of the type of relationship existing among terms (e.g., role indicators, "links," etc.); this type of indication is generally required only when indexing must be provided in depth for most documents.

Indexing processes may range from the controlled (i.e., use of a rigid authority list of some type) to the uncontrolled (so-called "free indexing"). Most information systems require at least some degree of control to ensure that retrieval will be both effective and efficient (e.g., confounding of synonyms, use of proper word-form, etc.). Many systems require even greater degrees of control (e.g., choice of alternate and/or additional terms — broader, narrower, or nearly synonymous — chosen from an authority list supplemented by an extensive cross-reference structure, such as a thesaurus). The maintenance and updating of vocabulary authorities is a complex but vital function which must be performed.

For manual operation, index terms are often left in natural language (English), except when they are classification notations. For mechanized operation, index terms may be coded or stored in natural language. Natural language is easier for the indexer and user as it avoids a code look up. It requires considerably more keypunching, however and has problems of spelling, accepted word forms, and the like. Term codes may be assigned at random or in various sorts of orderly fashions. Random codes are most useful when used with certain types of manual linear files employing marginally-punched, superimposed-coded cards. Ordered code sequences may be based on the alphabetical sequence of terms, the frequency of term usage, or a classification notation. Ordered codes may be dense or sparse.

Properly designed sparse codes permit at any time the insertion of a code for a new term into the vocabulary in its proper location in the code list without recording any previously existing terms. Dense codes usually require that the system maintain a separate sparse code for ordering (sorting) purposes. A code can be developed with the property that between any two codes one can always insert a new code without changing the arrangement or significance of any of the codes involved. For example:

$$1.5121 \longrightarrow \begin{matrix} 1.512 \\ 1.513 \end{matrix}$$

The physical form of an index may be varied widely to best meet local requirements. For some environments, a catalog card index is preferable; for others, a marginally-punched card index is best. For yet others, an inverted internally-punched card index (e.g., Batten card, peek-a-boo, etc.) is suitable, or perhaps a terminal-digit-posted card index (e.g., "Uniterm"). Printed indexes may consist essentially of mounted, photographed cards of the above types, or they may have their components created and assembled especially for index purposes.

All types of printed indexes (and many types of card indexes) can be assembled and printed in camera-ready form by a computer. Whether or not this is done depends principally upon the economics of the particular situation.

5.5.4 Announcement

Current-awareness needs may be served by announcing newly obtained documents as well as by formally circulating or disseminating them. Announcement bulletins are particularly useful when the system's clientele is large. The entries in announcement media may or may not be grouped in broad categories; the entries may be merely citations or they may be abstracts; and the medium may or may not include an index to the entries.

If abstracts are employed for announcement purposes, their production should be in a physical form so that labor to assemble the announcement medium will be minimized (e.g., no re-typing). The same holds true even if less complete surrogates (e.g., citations) are employed instead of abstracts. In fact, a system should be devised to ensure that the initial creation of each surrogate in its final form should serve all future purposes, including announcement, storage (e.g., catalog cards), and bibliography assembly. Various techniques are employed to accomplish this ranging from sequential card composition to computer-driven photocomposing machines.

5.5.5 Index Operation

The physical medium constituting an index, whether it be a manual (e.g., file card) or a mechanized (e.g., magnetic tape) system, must be processed in some manner to permit its updating with new index terms and/or index entries. Particularly for mechanized systems, the structure of the index file itself controls the file maintenance techniques; index file structures are discussed in Section VIII. Similarly, search procedures are governed by the index file structure. A human intermediary or a real-time response (which will permit the inquirer to have a "dialogue" with the system), is essential to ensure that the inquirer and the system are able to communicate without subtle or inadvertent misunderstanding.

Outputs from the index operation are responses to inquiries consisting of document references and/or other surrogates. As the response to the search of a mechanized index, document number references are usually obtained which can be used as input to yet another operation to obtain more informative surrogates, such as abstracts.

Abstracts (or less-complete surrogates of documents, if such are satisfactory) must be stored and retrieved for the creation of bibliographies in response to inquiries received by the retrieval system. Abstracts can be stored on cards, which can be extracted (after using the index) as required and reproduced in card form (or mounted and reproduced in page form) to create a bibliography in response to an inquiry. Alternatively, abstracts may be stored in paper-tape form and used to print page-form bibliographies. It is possible to store the surrogates on magnetic tape and print bibliographies therefrom via a computer; this is usually practicable (to date) only when short surrogates, such as citations, are employed.

5.5.6 Document Management

Documents may be stored either in original (full-size) form or in reduced-size form (microform). The choice between these alternatives rests upon the frequency of reference to the collection and the volume of reproduction traffic. Full-size documents require more storage space, although a proper choice of storage facilities can minimize this factor. They are legible without enlargement or reproduction equipment, although

if they must be reproduced frequently, another sizeable cost factor is introduced. Microform documents require less space for storage although storage and access can still be troublesome if roll or strips of microfilm (rather than microfiche - sheet microfilm) must be stored. Microform documents require relatively expensive equipment for their reading or full-size reproduction. The cost of microfilming and duplication of documents is discussed in Section VII.

Some information systems (particularly those serving a small, specialized clientele or covering a narrow area of technology) find it possible to group some or all of the stored documents, or their microforms, by broad subject classes and to index to individual documents assigned to these classes; that is, the class number is made a part of the document number. More frequently, however, documents are stored in accession number order.

When large amounts of continually-updated quantitative data must be stored, storage on magnetic tape, disc, or random access cards becomes economical.

5.5.6.1 Document Retrieval. Once the documents to be retrieved, have been identified, it may be desirable to obtain the actual document (either in full-size or in microform). No better method usually exists than to send a human being to the file to extract the documents when they are stored in full-size. This is also usually the best technique even when the documents are stored in microform. There are other means, however, which involve more-or-less automatic equipment (see Paragraph 8.1.6.2).

Occasional reproduction of full-size documents is a minor cost factor, but frequent reproduction (or reproduction in quantity) may become quite costly. (See Section VII). Such conditions might justify storage in microform, from which full-size copies may be produced at moderate cost and microform copies at low cost. In many cases where full-size copy is frequently requested, the center will reprint the document and maintain an inventory. Whether to pre-stock and replenish or to produce copies solely on demand depends primarily on the number of requests per unit time for a given document or for a given segment of the file. (See Section VII.)

5.5.6.2 Document Dissemination. It is conventional to serve current-awareness needs by circulating journals (possibly multiple copies) and by providing distribution lists for reports. These techniques have a distinct value but one which cannot be extrapolated or enlarged indefinitely, particularly when classified documents are involved. Alternative methods of document dissemination and their comparative costs are discussed in Section VII.

More recently, the selective dissemination (SDI) techniques have been developed. SDI is still a very expensive operation, and its effectiveness, except in limited environments, is yet to be proved.

Initial dissemination of microforms are frequently made by broad category definition to collective groups of users of a system. The assumption here is that the microforms are inexpensive and reproducible, thereby serving to decentralize future accesses to these documents.

5.5.7 End-Use

Little is known about the end-use of information, i.e., how information is utilized in the performance of a task. Where information is supplied to a user in either a digital form or as a microform, some means of converting it to visual form is required (e.g., mechanical printer, display, microfilm reader - printer).

SECTION VI. TYPICAL APPLICATIONS

6.1 GENERAL

This section describes five "typical IS&R applications" including a description of the information products and services produced within each IS&R system and the detailed processes by which they are produced. Four of the five applications described involve some form of scientific or technical information (STINFO). The reason for the preponderance of STINFO applications in the following discussion is the fact that the state of the art in the field of Information Storage and Retrieval is more advanced in the area of STINFO than in other fields of interest. The IS&R concepts and techniques which have already been applied to STINFO communication problems will, in the next decade, be applied to many other application areas.

The purpose of this section is to present some of the details, complexities and practical considerations that arise in operating IS&R systems. The sample of five is too small to permit drawing any general conclusions about needs or current practices. (A survey-based analysis of the state of the art is not within the scope of the work.) Many aspects of current IS&R practice are brought out in these typical application descriptions.

The five applications which are described are:

(1) Mission-Oriented Information Center. The NASA Scientific and Technical Information Facility is described as an example of a mission-oriented information center. This information system falls within System 7 as defined in Section V. The major system functions are origination, acquisition, surrogation, announcement, index operation, document management, and end-use. The system is heavily oriented toward the dissemination and retrieval of documents rather than toward facts. Its products and services are supplied to other information systems which are often in closer contact with the ultimate user.

(2) "Satellite Information Center." General Electric's Missile and Space Division (MSD) Technical Library is a major user of the services of the NASA Facility. Consequently, such major users are sometimes referred to as "satellite information centers." The GE-MSD Library is a complete information system in its own right, performing the same general functions as the NASA Facility. It is classified as System 7 in the reacquisition mode because it acquires and reuses the output of another information system. Its primary emphasis is on reference services and acquisitions on behalf of the user. It frequently acts as an intermediary between another information system and the ultimate user.

(3) Traffic Routing Center. (Army Chemical Information and Data System) This application is still in the research and development stage. One of several design concepts for the Army Chemical Information and Data System is discussed. The Traffic Routing Center (TRC) serves the function of routing a request to an appropriate Technical Information Center which is qualified to handle it. The TRC files might consist primarily of an index to the qualifications of each of some 60 Technical Information Centers plus basic data on the physical, chemical, and biological properties of some 2.5 million chemical compounds. It can be classified as System 3 since it involves the functions of origination, acquisition, surrogation, index operation, document management, and end-use. The emphasis is on retrospective search and on fact retrieval. The document management function consists of storing and retrieving the data records which will probably be stored in a computer system.

(4) Engineering Data Center. A description of the Engineering Data Management Department of the Naval Air Technical Services Facility (NATSF) follows. NATSF is the central repository of engineering data for the Bureau of Naval Weapons. The system utilized is System 3 but it differs from the Traffic Routing Center in practically every other respect. Its emphasis is on the dissemination and fulfillment of requests for copies of documents, specifically engineering drawings. Its index management function is of minor importance as the great majority of its requests are made by drawing number. Like the NASA system, NATSF acts as a wholesale distributor to other satellite centers which are in closer contact with the user. The users are primarily involved in purchasing or maintenance of production equipment and not in design engineering or research and development.

(5) Real Estate Title Searching System. Several examples of title searching systems are given including those of the Recorder of Deeds = City of Philadelphia, a large title insurance company, the Recorder of Deeds at Deedham, Massachusetts and the Title Insurance Company of Los Angeles. All of these systems are of System 3 as they involve the functions of origination, acquisition, surrogation, index operation, document management and end-use. In the recorders' offices, separate grantee-grantor indexes are maintained because of the need for being able to determine if any property is owned by a particular, named individual. The title companies usually maintain a file of documents by property or lot number wherein once the property location is clearly identified, all the documents pertaining to that property will be found in one place. The surrogation function is that of analyzing the legal description in the document to arrive at the property or lot number, (which is the only surrogate assigned to the document). In the recorders' offices, the names of the grantee and grantor or mortgagee and mortgagor, and the date of the instrument are also part of the total surrogate of the document. The various parts of the surrogate are stored in separate index files — usually bound volumes.

6 2 MISSION-ORIENTED INFORMATION CENTER (NASA)

The mission-oriented information center is characterized by its interdisciplinary nature, i.e., its coverage of a number of disciplines. The most widely known mission oriented information centers are the Defense Documentation Center (formerly ASTIA), the Technical Information Division of the Atomic Energy Commission, the Scientific and Technical Information Facility of the National Aeronautics and Space Administration, the National Agricultural Library and the National Library of Medicine.

6.2.1 NASA — Scientific and Technical Information Facility. (Hereinafter Called the Facility)

The Facility, which is operated for NASA by a contractor (Documentation, Inc.), acquires and selects technical materials, mostly reports, to be added to the NASA collection, abstracts and indexes them as appropriate, prepares camera-ready copy for announcement journals and book form indexes, provides current and continuing dissemination services in both hard copy and microfiche, (sheet microfilm), provides a supporting reference service, and compiles bibliographies in specialized subject areas.

The NASA system interfaces with both government and non-government information systems. Non-classified NASA publications are made available to the public through The Clearinghouse for Federal Scientific and Technical Information (formerly OTS) as well as to the 12 Federal Regional Technical Report Centers. NASA, DOD contractors, and authorized government personnel are serviced through a complex of NASA field centers as well as through the Facility. These service points include the information divisions at NASA centers, major contractor facilities, and ASTIA field centers.

An important interface exists with the American Institute of Aeronautics and Astronautics (AIAA). Through an arrangement with AIAA, all of the journal literature related to the NASA program is processed and fed into the NASA Facility in a compatible form. AIAA produces its own announcement journal entitled International Aerospace Abstracts.

6.2.2 Inputs and Outputs

Figure 6-1 illustrates the annual inputs (raw materials) provided to the Facility and the outputs (services) provided by the Facility. You will note that there are two basic kinds of inputs: Technical Reports and Bibliographical Information on Journal Literature which is processed by AIAA and furnished to the Facility on magnetic tape, along with compatible microfiche copies of all non-copyrighted journal articles. The technical reports are received from DOD, NASA, other Government agencies, NATO, and Government contractors in hard-copy form and sometimes along with a small inventory of printed copies. The Atomic Energy Commission furnishes the Facility with a microfiche copy in compatible form, of all reports processed by its Technical Information Division, which are relevant to the NASA system.

The major outputs of the NASA system are represented along the bottom of Figure 6-1. They include the publication of an announcement journal which is distributed to about 8,000 locations, the initial distribution of microfiche copies of new accessions within specified categories (of which there are 34) to 100 recipients of the NASA microform. There are presently about 3 million microfiche copies disseminated annually. There are also demand request services for hard copies of documents, microfiche copies of documents, and selective bibliographies.

An interesting characteristic of the NASA information system is that its services are generally not directly utilized by the ultimate consumer. Note in Figure 6-1 that the typical categories of users of the Facility are Libraries, DDC field offices, NASA Research Centers, other government information centers, and NASA contractors. In effect, therefore, the NASA facility is the wholesaler and its customers are the retailers of its information products and services. These organizations act as an intermediary between the ultimate consumer, the engineer or scientist, and the NASA information system. To further understand this reacquisition process a satellite center, which is a typical customer of the NASA facility, is described under Paragraph 6.3. The primary products which NASA manufactures, therefore, are:

(1) Announcement (abstract) journals.

(2) Book-form indexes.

(3) Magnetic tapes of index data and citations.

(4) Search programs for decentralized searching.

(5) Microfiche copies of reports which can be reproduced or used for producing hard copy enlargements on a decentralized basis.

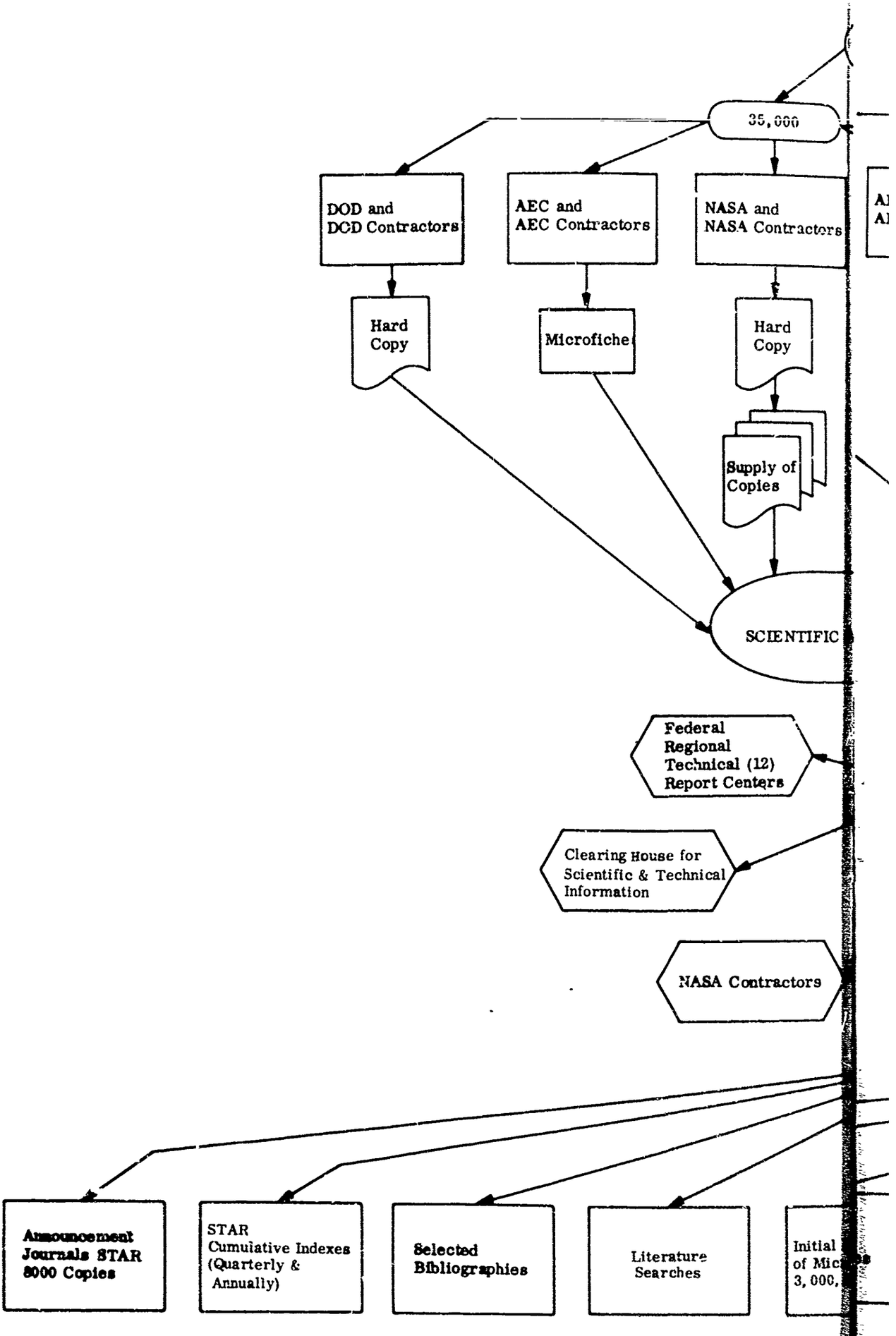
(6) Hard copies of reports on request.

(7) Literature searches.

6.2.3 Announcement and Dissemination Process

The various operations involved in the announcement and initial dissemination process and the order in which they are performed are illustrated in Figure 6-2.

6.2.3.1 Document Dissemination. Acquisitioning is the first system function. It involves checking for duplicates, assigning accession numbers, and storing additional copies. Following this is descriptive catologing, abstracting, and indexing. The next operation is microfilming one copy of each report accession. The bindings of each report are sheared and the report microfilmed at approximately 1:18 on a planetary camera. The microfilm is then processed on a continuous flow processor. A duplicate roll of microfilm is produced on diazo film using a roll-to-roll duplicator. Following this, the silver negative is slit and stripped-up to prepare microfiche masters on a specially designed


35,000
DOD and DOD Contractors
AEC and AEC Contractors
NASA and NASA Contractors
Hard Copy
Microfiche
Hard Copy
Supply of Copies
SCIENTIFIC
Federal Regional Technical (12) Report Centers
Clearing House for Scientific & Technical Information
NASA Contractors
Announcement Journals STAR 8000 Copies
STAR Cumulative Indexes (Quarterly & Annually)
Selected Bibliographies
Literature Searches
Initial
of Mic
3, 000,

A

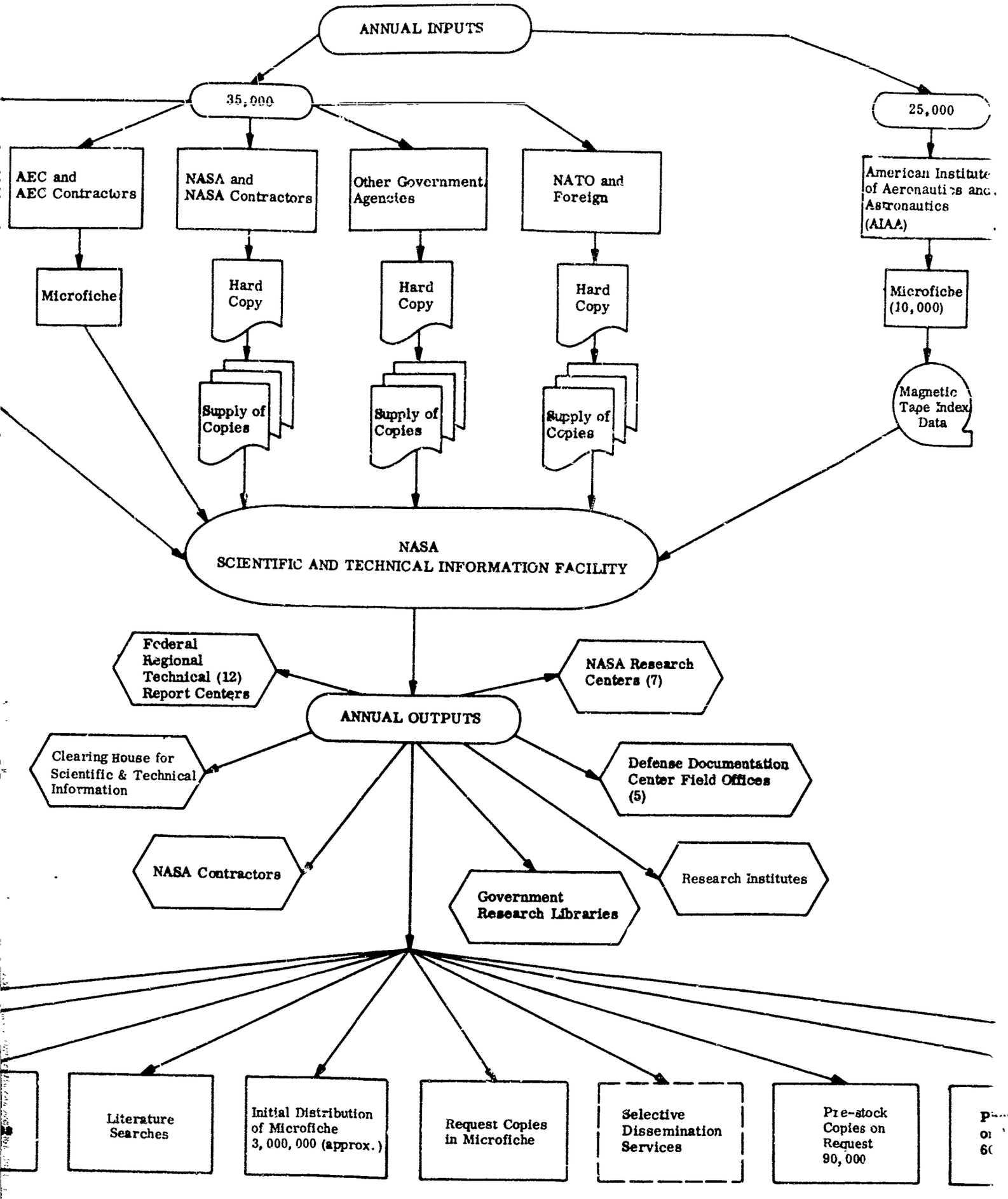


Figure 6-1. Mission-Orie
Input-C

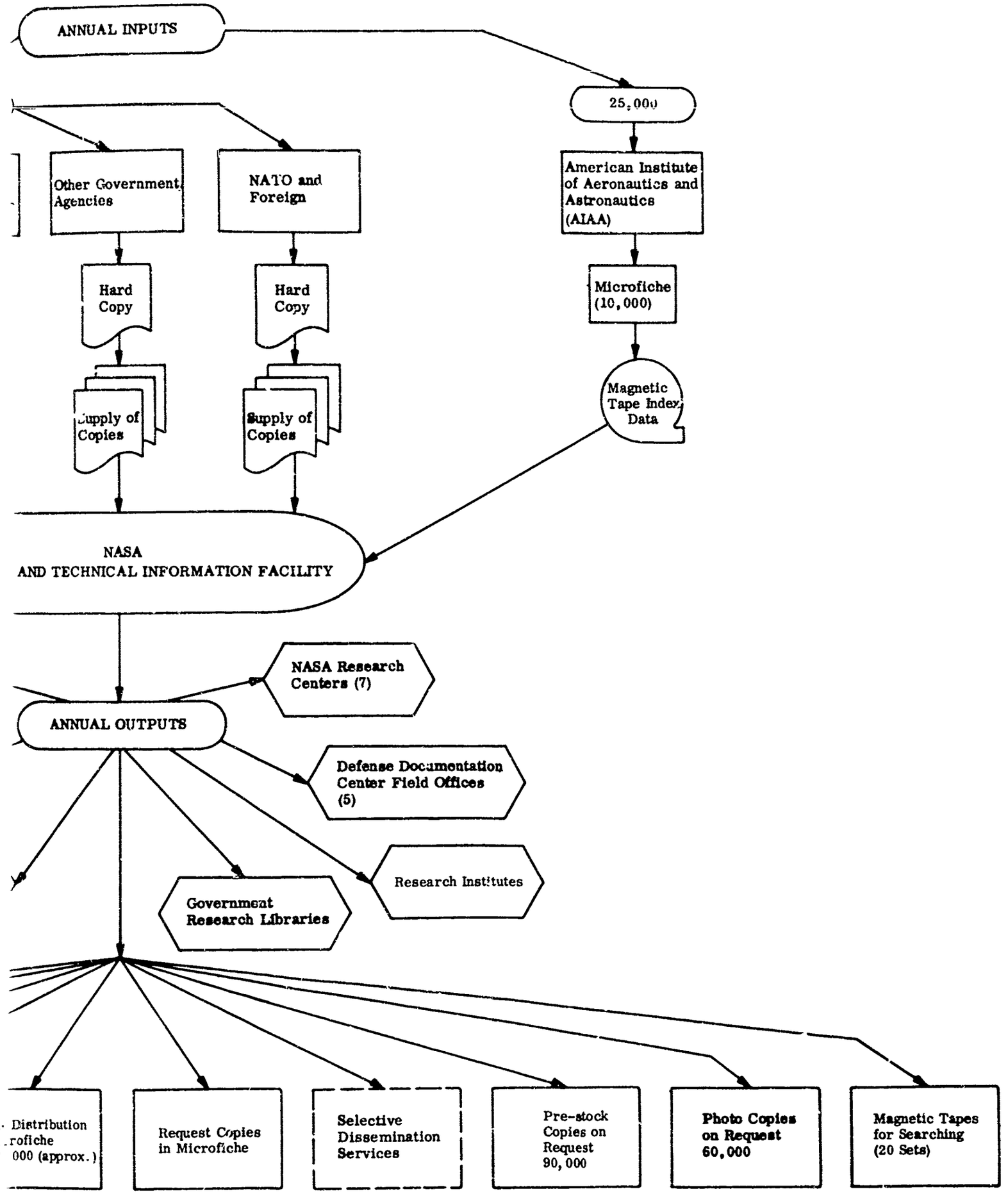
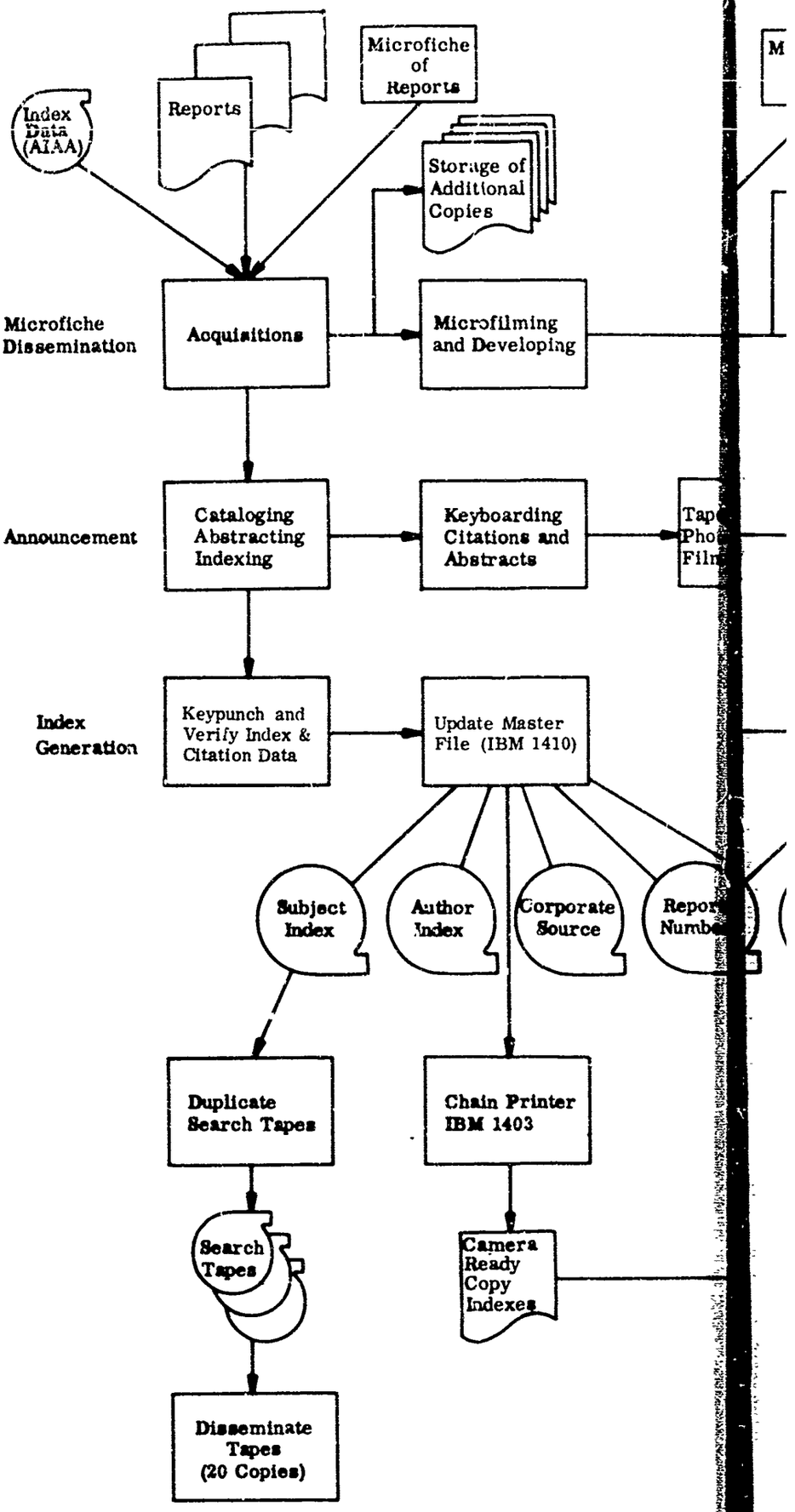


Figure 6-1. Mission-Oriented Information Center (NASA) Input-Output Diagram

work place. The microfiche master is contact printed onto 4 x 6 diazo masters using a point light source, vacuum frame plate making machine. This second generation master is used for making quantity duplicates. Dissemination copies are hand distributed by accession number according to category lists which indicate the accession numbers in each category for the period.

6.2.3.2 Announcement. The accessions to the NASA facility are announced to the public through the medium of the NASA announcement journal, Scientific and Technical Aerospace Review (STAR). The journal literature, which is processed by AIAA, is announced through the medium of its own journal, International Aerospace Abstracts (IAA). The primary intellectual surrogation operations which precede the publication of these journals are: descriptive cataloging, abstracting and indexing. Since these functions have all been described in Section V, they will not be repeated. However, it should be noted that the indexing operation serves two functions. Wherein an average of 15 index terms are assigned for each document, only an average of 3.5 subject index entries will appear in the book-form indexes to STAR or IAA. Deeper indexing is required for retrospective computer searching. All of the indexing, abstracting, and cataloging is recorded on a multiple part form which accompanies each document.

A phototypesetting system was adopted to produce camera-ready copy of graphic arts quality for the citation and abstract portion of STAR, and also to provide a machine-readable input of the citations to the computer-searchable retrospective store. The citations and abstracts are keyboarded on modified Friden ICC-S Justowriters. The paper tape product of this keyboarding operation is utilized directly to drive a Photon photocomposing machine, which produces graphic arts quality copy on photographic paper. The standard direct keyboard operated Photon machine has a 1440 character repertoire and a selection of 12 point sizes. The tape operated Photon system developed by NASA is limited to a 450 character capability plus superscripts and subscripts, super-superscripts, and sub-subscripts which are produced by a series of code actuated off-set lenses.


Index Data (AIAA)
Reports
Microfiche of Reports
Storage of Additional Copies
Microfiche Dissemination
Acquisitions
Microfilming and Developing
Announcement
Cataloging Abstracting Indexing
Keyboarding Citations and Abstracts
Index Generation
Keypunch and Verify Index & Citation Data
Update Master File (IBM 1410)
Subject Index
Author Index
Corporate Source
Duplicate Search Tapes
Chain Printer IBM 1403
Search Tapes
Camera Ready Copy Indexes
Disseminate Tapes (20 Copies)

A

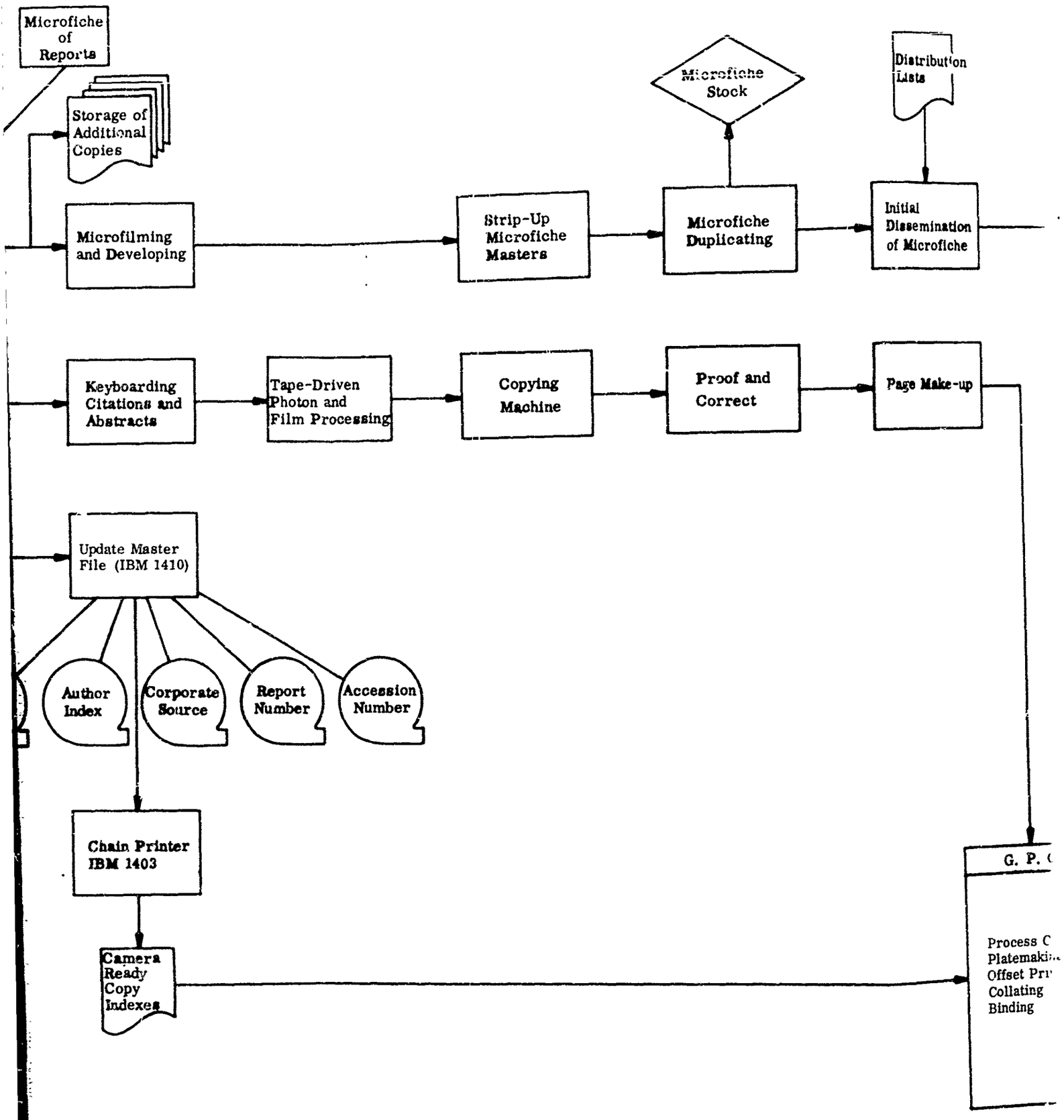


Figure 6-2. Mission-Orien
Announcemen

B

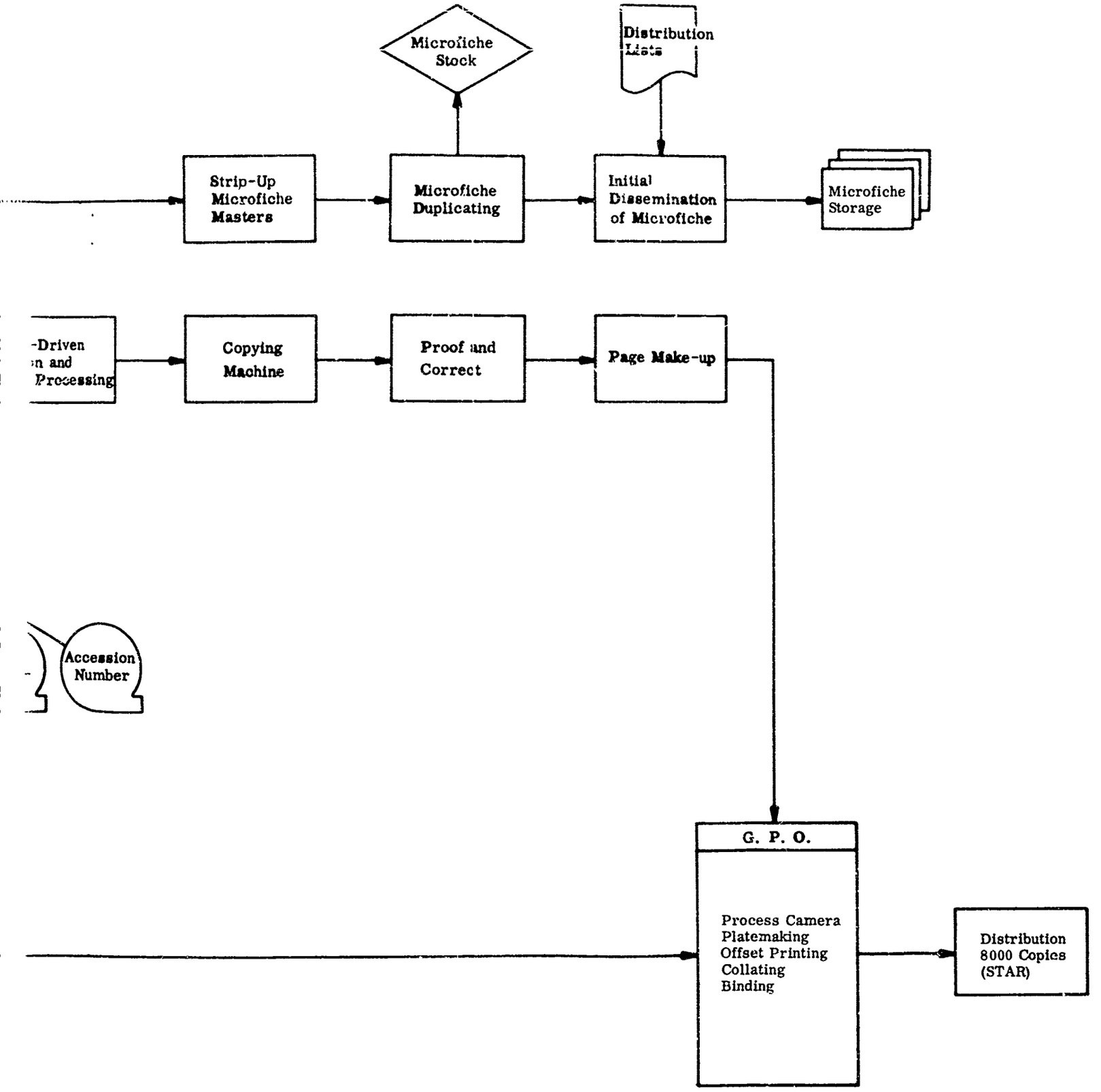


Figure 6-2. Mission-Oriented Information Center (NASA) Announcement and Dissemination

The paper or film output of the Photon is processed and dried and then a diazo proof copy is prepared for proofreading and the marking of corrections. Corrections are keyboarded either directly at the Photon or on the LCC-S keyboards and the output copy is cut and pasted up prior to the page make-up process. The resultant copy along with the computer output of the index section is sent to the Government Printing Office for plate making, printing, collating, and binding.

6.2.3.3 Index Generation, Storage, and Dissemination. The index terms and the descriptive catologing data which have been recorded on a standard work sheet are keypunched and verified and then read into the IBM 1410-1401 computer system. The tape is then stripped down to produce separate subject, author, corporate source, report number and accession number indexes. For each semi-monthly issue of STAR, a camera-ready printout of these five indexes is produced on the IBM 1403 chain printer. It should be noted that the book-form subject index only contains those entries which have been flagged by an asterisk. In addition, four cumulative indexes are produced by the same method, quarterly, semi-annually, third quarterly, and annually.

The retrospective search tape is arranged in linear file fashion by accession number followed by the full citation (without the abstract) and the index terms. The file contained approximately 100,000 items in July of 1964, with an average depth of indexing of 15 terms for a total of 1.5 million index entries. Copies of this magnetic tape file are duplicated and disseminated to 20 user groups. Updated information is provided monthly. Seven of these users are NASA centers and the other 13 are contractor information centers.

6.2.4 Request Processing

Figure 6-3 illustrates the typical operations involved in processing user requests for information services. There are generally three types of requests: requests for hard copy, requests for microfiche copy, and requests for literature searches.

6.2.4.1 Requests for Copies. The first operation in processing requests for hard copy is to validate the user's authority to receive the copy which may involve a check for security classification. If the request is fully identified and includes an accession number, the copy will be retrieved manually from stock or reproduced from microfiche if no stock is available. If the request is not fully identified, a reference look-up must be made in the card catalog files, book-form indexes, or computer index to identify the proper accession number for the requested copy.

If the copy is out of stock, the standard procedure is to manually retrieve a master microfiche copy and produce the hard copy from this on a Photostat model 1014 enlarger. The hard copy will be two pages per 8-1/2" x 11" sheet at 70 percent of original size. If the request was for a microfiche copy, this will be pulled directly from stock which is replenished as required. No perpetual inventory control is maintained of the stock copies except for a simple notation of whether the copy is in stock or not.

6.2.4.2 Requests for Bibliographies. Requests for bibliographies are handled either manually by use of the book-form index to STAR or by computer search. The requests are processed by a reference specialist who analyzes the question and constructs a query with the aid of a subject heading listing. The index terms assigned to the query along with their logical combination instructions, such as intersection, union or negation, are key-punched and verified. The questions are batched and processed twice daily against the entire citation file. The output of these searches is presently a list of citations printed out on the high-speed mechanical IBM 1403 printer. The majority of searches result in from 20 to 70 citations per query.

6.2.5 Selective Dissemination

The Facility is presently experimenting with a selective dissemination program being developed for them by IBM at Yorktown Heights, New York. They are profiling 500 individual NASA Scientists and Engineers for a test of the system. Since the SDI system is not yet operational, its results are unknown.

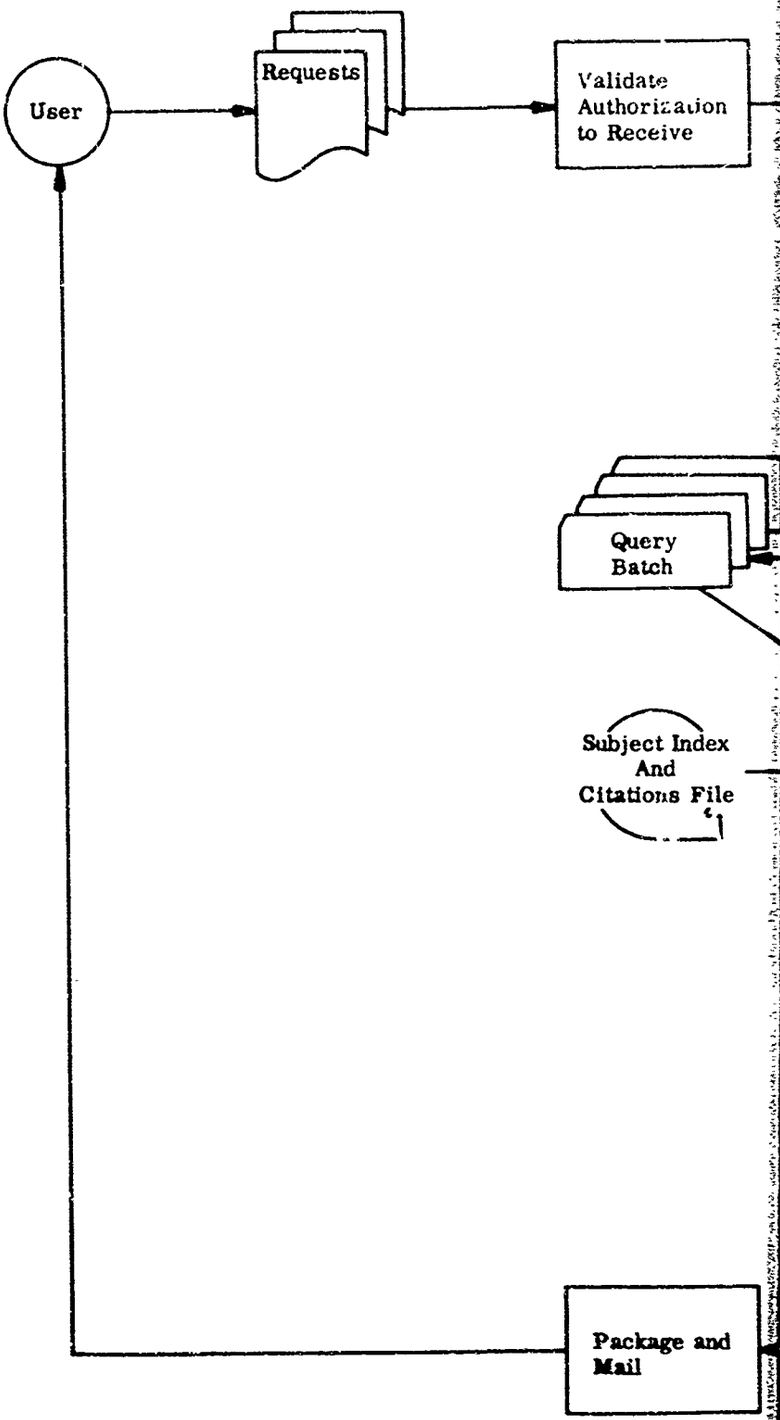
6.3 SATELLITE INFORMATION CENTER

The name, "Satellite Information Center," is used to represent the type of information center which operates in the reacquisition mode, i.e., which utilizes information products and services produced centrally by another information system. Since we described the NASA system as an example of a centralized mission-oriented information center, we will describe a typical "satellite" of the NASA Facility and the degree to which the services of the NASA Facility are integrated into the system of the satellite center. The system described herein is that of General Electric's Missile and Space Division Library, which is located at King of Prussia, Pennsylvania, with a branch at 32nd and Chestnut Streets, Philadelphia, Pennsylvania.[9] [10]

The GE-MSD Library services several thousand engineers and scientists of the Missile and Space Division of General Electric. The library is staffed by a total of 18 people.

6.3.1 Inputs and Outputs

Figure 6-4 illustrates the annual inputs to and outputs from the Library. The Library is divided into several physical facilities. At King of Prussia, there are two facilities. One is the Library which contains books and periodicals and the other is the Document Center which contains all government report materials, both classified and unclassified. At 32nd and Chestnut Streets in Philadelphia, a complete collection is maintained of unclassified technical reports furnished in microfiche form by the NASA Facility.


User
Requests
Validate
Authorization
to Receive
Query
Batch
Subject Index
And
Citations File
Package and
Mail

A

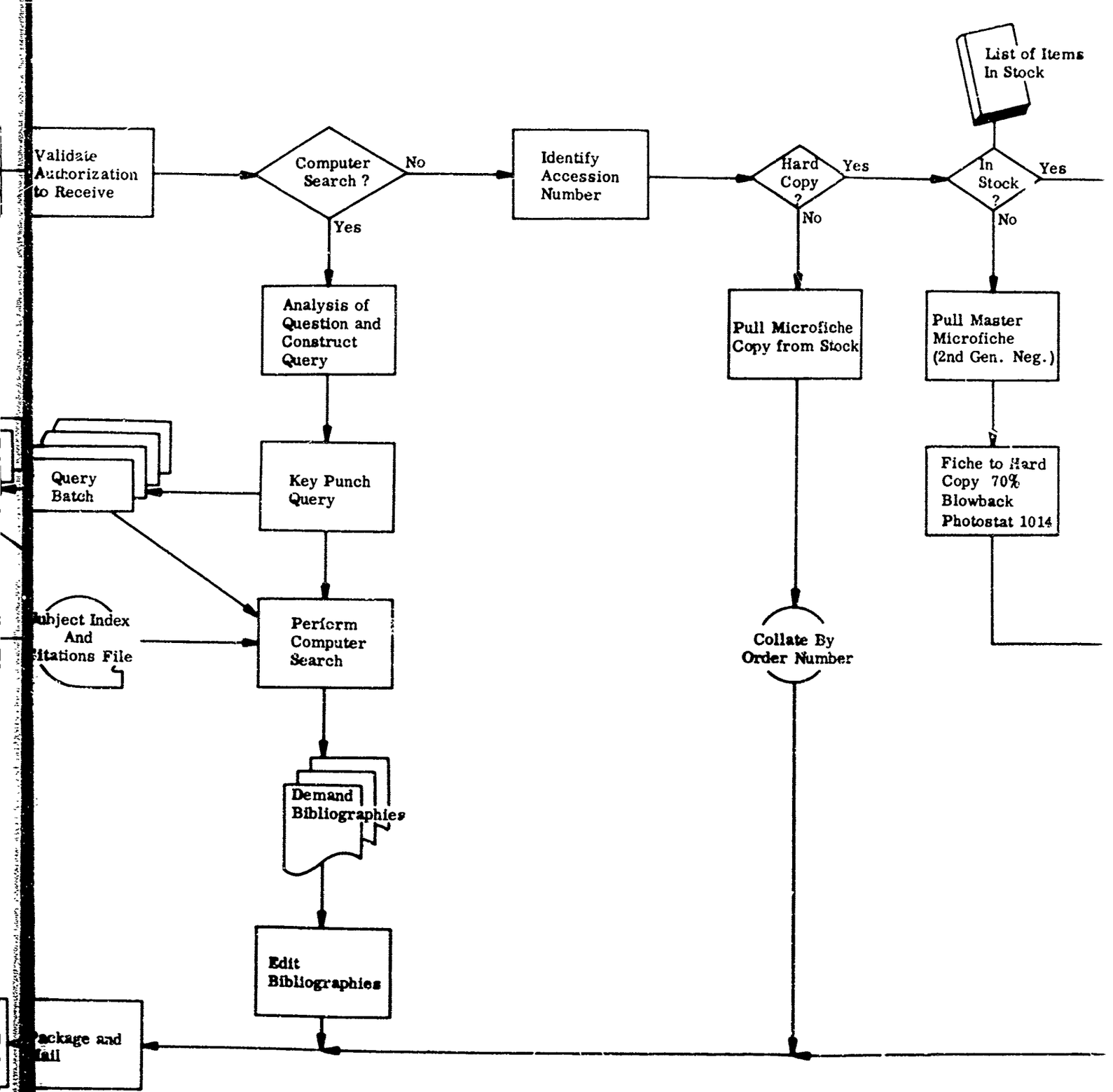


Figure 6-3. Mission-Oriented Informa
Request Processing

C

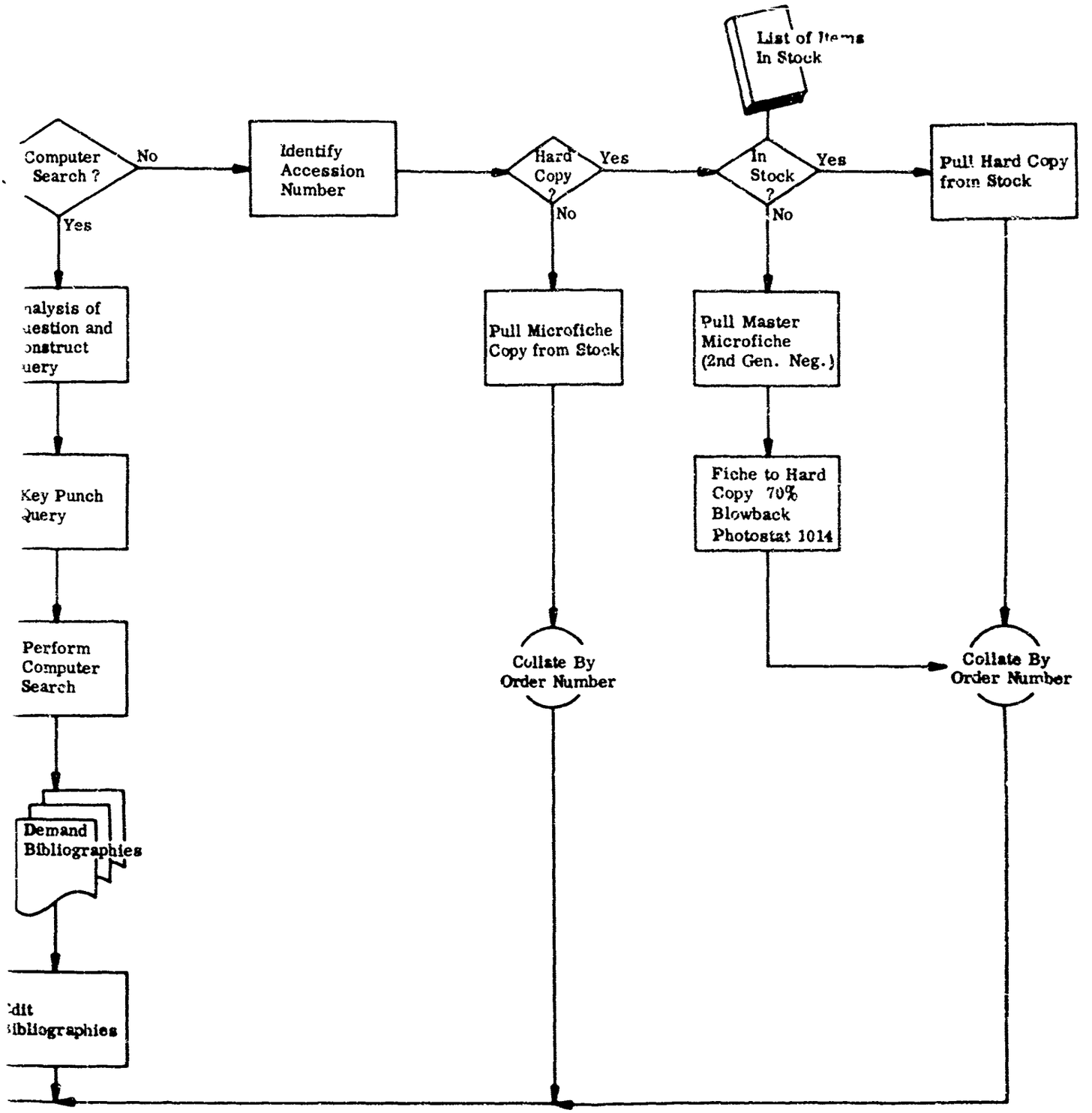


Figure 6-3. Mission-Oriented Information Center (NASA) Request Processing

C

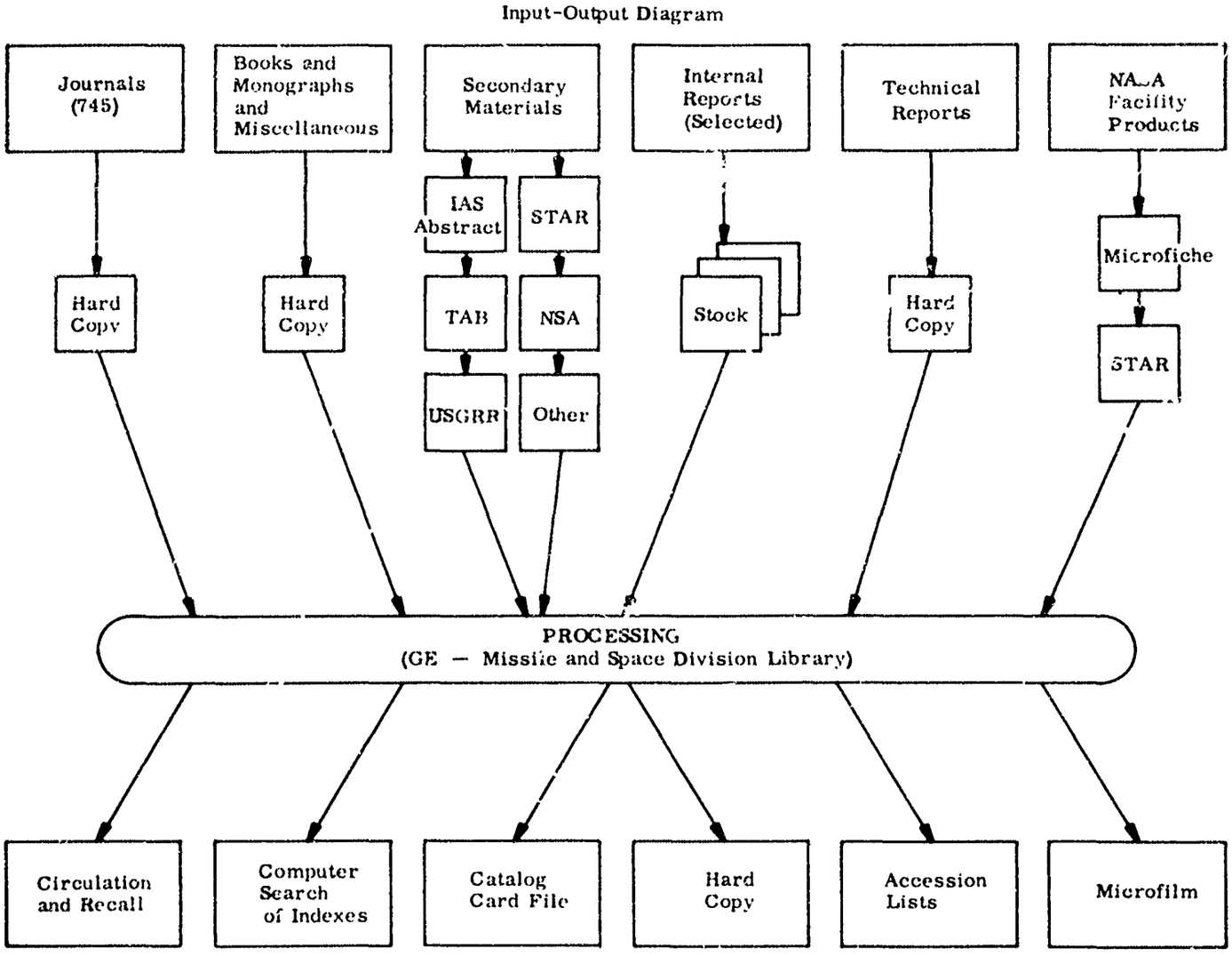


Figure 6-4. Satellite Information Center

Although the GE-MSD Library utilizes a computer (GE-225) for retrospective searching, they do not utilize the NASA Search Tapes. They do use, however, many of the secondary materials published in book form, such as the abstract and index journals, STAR, Technical Abstract Bulletin, Nuclear Science Abstracts, International Aerospace Abstracts, and others. They receive approximately 15,000 technical reports per year and subscribe to some 745 journals. The total collection at the GE Library now stands at 130,000 technical reports, plus 36,000 books and bound journals for a total of 166,000 volumes.

The outputs or services provided by the Library are depicted on the bottom of Figure 6-4. They provide an accessions list of new documents received by the Library, which list contains a full citation plus an abstract of each document. It is distributed to approximately 50 managers who presumably circulate it among their staff. There is no routine circulation of journals or abstract journals by the Library. In fact, journals are not even lent by the Library; instead, the Library will reproduce and provide a retention hard copy of any journal article to a bona fide requestor. Technical reports, however, are loaned by the Library and retention copies of reports are prepared only upon written authorization. The Library also has microfilm reader printers (Filmac 100 and Filmac 200) on which the user may prepare his own hard copy without authorization. The Library will provide bibliographic assistance to any GE user. Most of the requests can be handled by a search in either the catalog card file or in the secondary journals, such as Technical Abstract Bulletin or STAR. If a more extensive bibliography is needed, a computer search can be made of the deeper indexes which the Library has prepared and stored on the central GE-225 computer facility.

6.3.2 Storage and Announcement Process

Figure 6-5 illustrates the entire processing cycle for the GE system. After completion of the acquisitioning process, which primarily involves checking for duplicates and assigning a unique accession number, all technical reports are cataloged, abstracted and indexed. An author abstract will be used if available, or if not, a few lines from the summary or introduction will be utilized. At the time these reports are received, they do not have the NASA Facility accession number and frequently do not have an AD number (DDC). Also, these reports are received prior to their announcement in either the NASA

or DDC announcement journals -- the STAR and Technical Abstract Bulletin. Consequently, the GE-MSD library chooses to catalog, abstract, and index each document for its own purposes. It should be noted that 40 percent of the GE-MSD collection is company generated. Only a small portion of these reports are announced by AEC, NASA, or DDC, and at a much later date. Books and monographs are also processed in the same fashion, whereas journals are merely checked in and put on the shelves. The copy for cataloging and abstracting is typed and proofed and then made up into pages for the accession list. These same pages are also utilized for the printing of catalog cards which are printed on index quality stock.

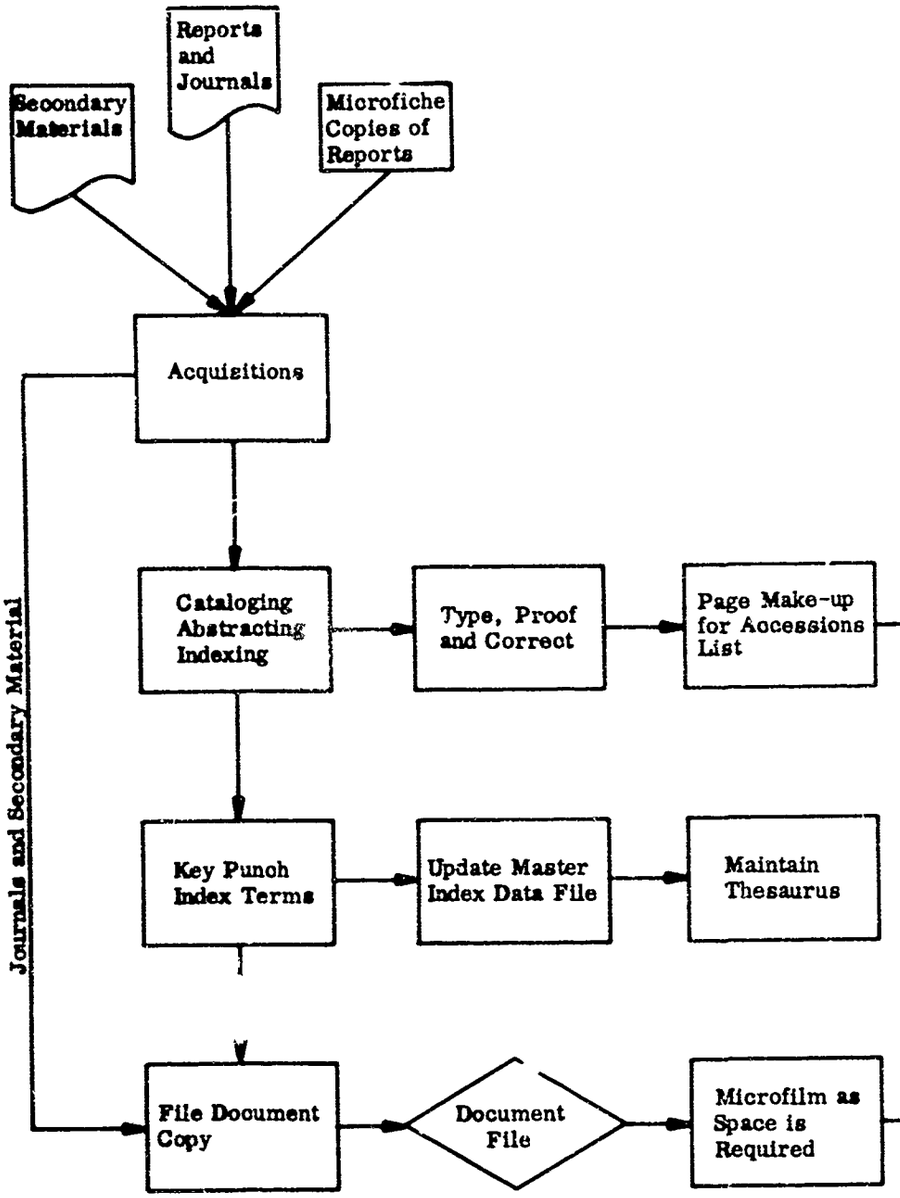
The indexing is done with the aid of a thesaurus which is based on the Defense Documentation Center Thesaurus. An average of eight index terms are assigned per document. The terms are not coded. They are keypunched in natural language and the computer index file is updated every several days.

Those terms which are not acceptable by the system are rejected and a decision made as to whether the thesaurus should be modified to include them.

Technical reports are maintained in hard copy form as long as storage facilities permit. Prior to the move to the King of Prussia facility, the Library would have a service company prepare a microfilm copy in jacket form of all documents older than one or two years. The quality of microfilming was rather low due to rotary microfilming at 24X and the use of acetate jackets. The original documents were destroyed after microfilming. No microfilming has been done in the last few years as they do not have as critical a space problem.

There is no routine dissemination of documents either by selective dissemination or other initial dissemination technique. There are no standard distribution lists or field of interest registers. The librarian, however, frequently will route new reports received by the Library to those engineers or managers he believes would be interested in receiving these documents.

The GE Library developed a preliminary system for mechanizing the catalog card file which would have eliminated the need for maintaining the card file. Essentially, the system would have utilized a Flexowriter for input, the paper tape by-product of which would be transferred to magnetic tape. The computer would then produce book-form catalogs which would be available within every technical operation of the Missile and Space


Reports and Journals
Secondary Materials
Microfiche Copies of Reports
Acquisitions
Journals and Secondary Material
Cataloging Abstracting Indexing
Type, Proof and Correct
Page Make-up for Accessions List
Key Punch Index Terms
Update Master Index Data File
Maintain Thesaurus
File Document Copy
Document File
Microfilm as Space is Required

A

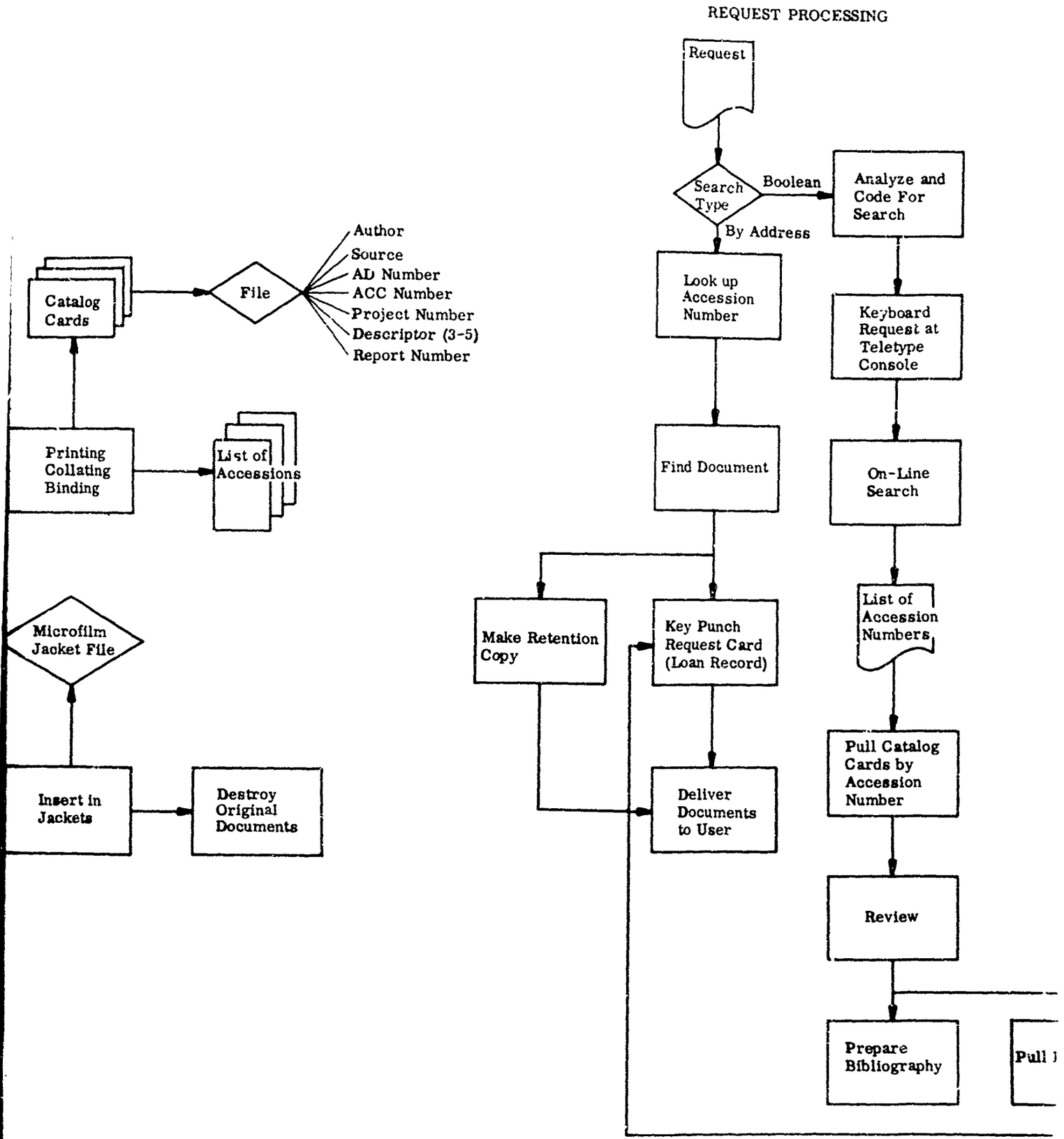


Figure 6-5. Satellite Information Center Proces
(GE - MSD Library)

B

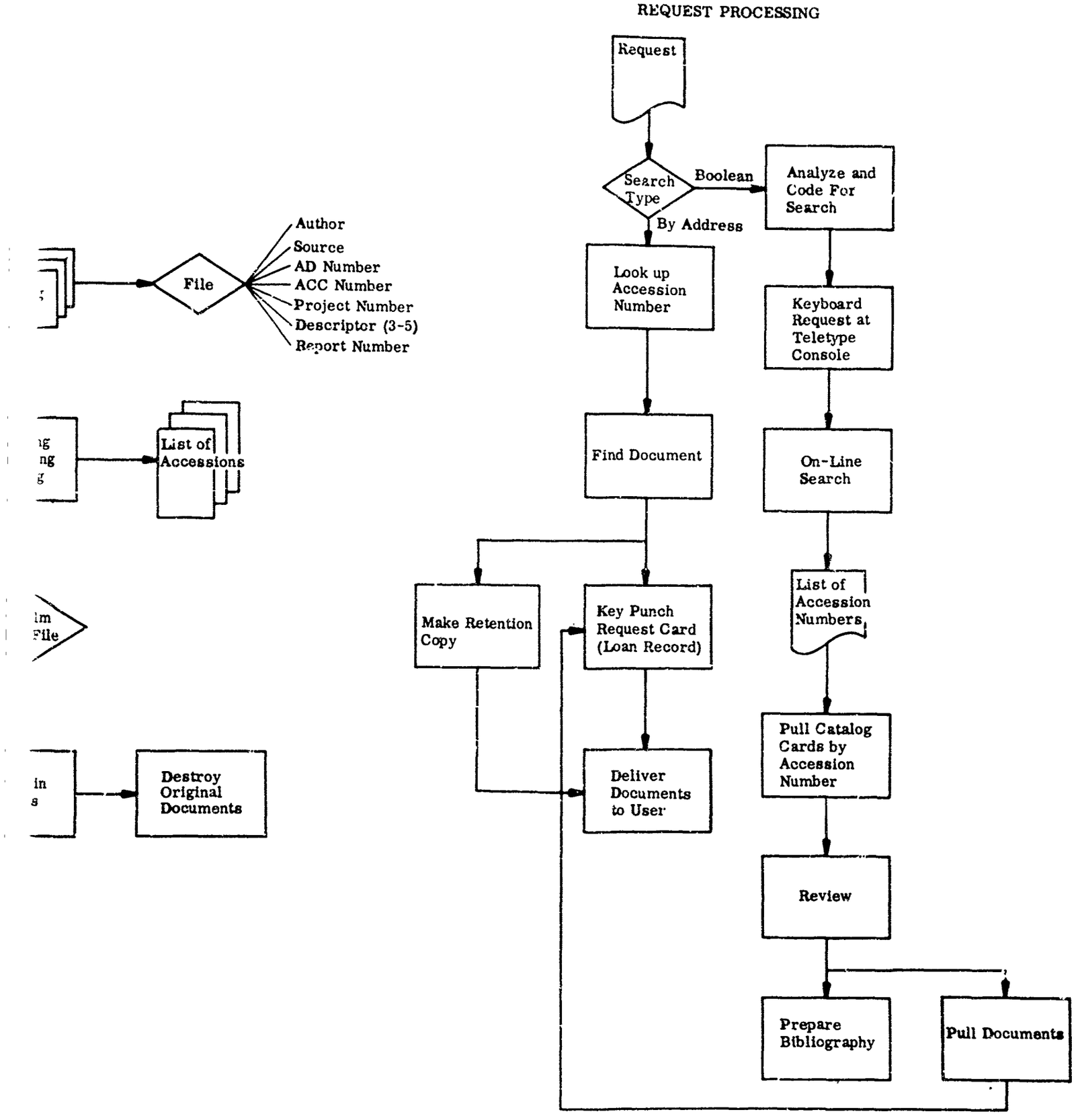


Figure 6-5. Satellite Information Center Processing Cycle (GE - MSD Library)

Division. The program was scrapped, however, allegedly due to the unavailability of an inexpensive input converter which would accept the paper tape input.

The book-form catalog idea has merit. It would be especially useful where many copies of a catalog card file are being maintained and updated on a manual basis. In this case, however, there is only one copy of the catalog card file and consequently, the publication of the book-form catalog would be more expensive than the present method. It would, however, afford more flexible and decentralized access to the index of the Library's collection.

6.3.3 Request Processing

As mentioned earlier, journals are not loaned by the Library. Therefore, if a user wants to see a particular article in a journal, he may either look at it in the Library or request to have a copy made for his permanent retention. For books, technical reports, and monographs, the Library has developed a mechanized circulation control system. This is a simple punch card system wherein the request card is punched and the accession number of the loan document is duplicated into the request card. Two copies are made, one of which is filed by employee number and the other by accession number. Overdue notices are prepared by machine every two weeks. Likewise, periodic inventories are greatly facilitated by these machine records.

Most reference requests are handled by utilizing the catalog card file or secondary materials, such as the abstract journals and their indexes. Where a complex search requirement is imposed, a request for computer search can be made. The GE-225 computer system is tied in directly to the Library by a communications link and a Teletype console keyboard for input-output. The computer can be accessed either by feeding the paper tape to the Teletype console or by keyboarding directly on the Teletype console. A designated time is available to the Library each day for requesting that a computer search be made on the central GE-225 computer. The Library has found that it pays to prepare the search question in advance and produce a clean paper tape. An average of ten searches are made in one batch each day.

The computer will respond first with the number of documents which meet the search criteria and the accession numbers of the first 10 documents. If the user would like additional citations, he will then respond by keying in the statement SEND MORE or SEND ALL. Since the output of the computer search is merely a list of accession numbers, it is necessary for the reference librarian to pull the catalog cards for each accession number. He will then review these to screen out those which are not really relevant and will then prepare a bibliography or pull the documents and prepare loan requests for sending them to the user.

6.3.4 Recapitulation and Analysis

6.3.4.1 Duplication of Descriptive Cataloging, Indexing, and Abstracting. The processes of descriptive cataloging, indexing and abstracting are generally the most time-consuming and expensive of all Library or information center operations. It is disturbing to observe that the GE Library finds it necessary to catalog, abstract, and index all technical reports it receives, in spite of the fact that many of these reports will eventually be cataloged, abstracted, and indexed by one of the major Government information centers. The problem, however, is that these reports are received much earlier by GE than are the corresponding products of the NASA Facility (STAR). The only choice available to the Library, therefore, is to do its own abstracting, indexing, and cataloging or to hold these documents in suspense until such time as the bibliographic products are available from the NASA Facility.

If the central (NASA) information services were compatible and more timely, this duplication of intellectual effort could be eliminated. These objectives appear to be difficult to achieve so the question is reduced to whether some better means can be developed for ameliorating the problems involved in the present situation. Perhaps a technique could be developed for rapidly and inexpensively producing a catalog card containing descriptive cataloging information and some form of abstract which could be used for an accessions list and for the catalog card file. One idea would be for the local library to make a copy of the title page and first page of the introduction or summary. These copies would be utilized until such time as the products of the NASA Facility or DDC were available.

6.3.4.2 Equipment and Compatibility Problems. It is interesting to note that while the GE Library employs a machine search system, it only recently decided to receive the NASA index tapes which are available, without charge, to contractor information facilities. Since these tapes are directly usable only on IBM 1401 computers, plans are now under way to convert these files to discs for use on the GE-225. The NASA index tapes are already in accession number order in linear file arrangement and contain the full bibliographic citation. A computer printout of these tapes could produce catalog cards containing the full citation and the list of descriptors which could be hand-checked against the catalog cards produced by the Library.

The book-form catalog project was also shelved because of code compatibility and format problems. The GE computer was unable to directly accept the output of the Flexowriter and it did not appear economical to obtain a converter solely for this purpose.

6.3.4.3 Duplication of Camera Microfilming. In the future, the GE Library should be able to request microfilm copies of those reports which they wish to "retire" to microfilm storage rather than to do their own camera microfilming. The advent of a U. S. Government Standard for preparing a microfiche copy of technical reports will facilitate this. They will receive a much higher quality product since the microfiche production at NASA, AEC, and DDC is done on planetary cameras to high quality specifications. It would even pay them to buy microfiche copies from some central source rather than to do their own microfilming because a duplicate microfiche is less expensive than preparing a camera negative and inserting it in a jacket.

6.4 TRAFFIC ROUTING CENTER (Army Chemical Information and Data System)

6.4.1 General

One of the fundamental issues in the planning of any new information program is the degree of centralization or decentralization desirable. At one extreme, there are proponents of a single national scientific and technical information center covering all scientific fields and servicing the needs of the entire country. The other extreme would involve a network of specialized information evaluation and dissemination centers, each covering a limited field of science and servicing only a small geographic area of the country. Neither extreme is entirely practical, nor has either extreme been followed

in practice. Many systems covering broad disciplines or missions have centralized certain functions such as descriptive cataloging, indexing, abstracting, and announcement. Functions such as reference services are provided on a decentralized basis at field offices, regional report centers, and by providing duplicate collections in microform.

There is a growing trend toward specialized information evaluation centers which not only control the literature in a given specialty but also provide for evaluation, correlation, and synthesis of the information they store. These centers allow for more detailed coverage of a specialized subject than is generally provided by a mission-oriented center such as the NASA Facility, and certainly more coverage than would be possible at a single national scientific and technical information center.

The problem, therefore, is how to make more effective use of the specialized information center. One concept being explored by the Army for its Chemical Information and Data System is that of the Traffic Routing Center (TRC) which would direct inquiries to an appropriate specialized information center by matching the general nature of the question against the profiles of the various Technical Information Centers (TICs).

6.4.2 Army Chemical Information and Data System (CIDS)

CIDS is an Army exploratory development program which was established in April of 1963. The objective of CIDS is to achieve more effective control over the growing mass of chemical data and information. There are, for example, over 2,500,000 chemical compounds which have been identified and approximately 70,000 new ones are being identified annually. The Army already has 60 Technical Information Centers (TICs) which collect, analyze, abstract, evaluate, disseminate, store, and retrieve information in specialized subject areas. It has been suggested that CIDS may consist of one or more Traffic Routing Centers which will store basic data on most of the 2,500,000 chemical compounds plus locator data on where to go for additional information. Each Traffic Routing Center would act as a switching and control center for a coordinated network of Technical Information Centers (TICs) linked together by a communications network.

Figure 6-6 illustrates the overall inputs and outputs to and from CIDS as a whole. The processing functions, however, would be allocated between the 60 Technical Information Centers and the Traffic Routing Centers. Figure 6-7 illustrates a possible network of Technical Information Centers tied together by communications and two Traffic Routing Centers.

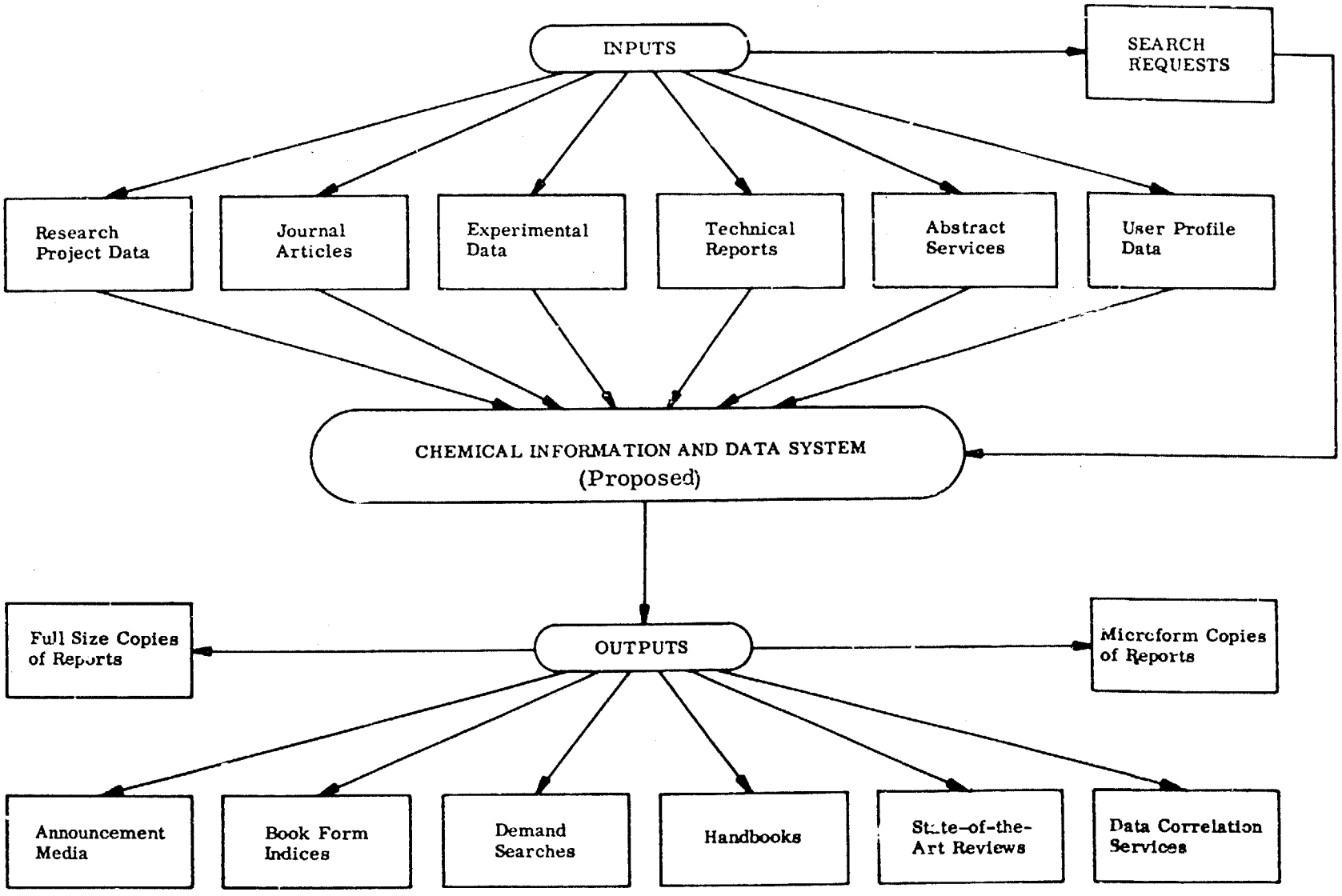


Figure 6-6. Chemical Information and Data System, Input-Output Diagram (Proposed)

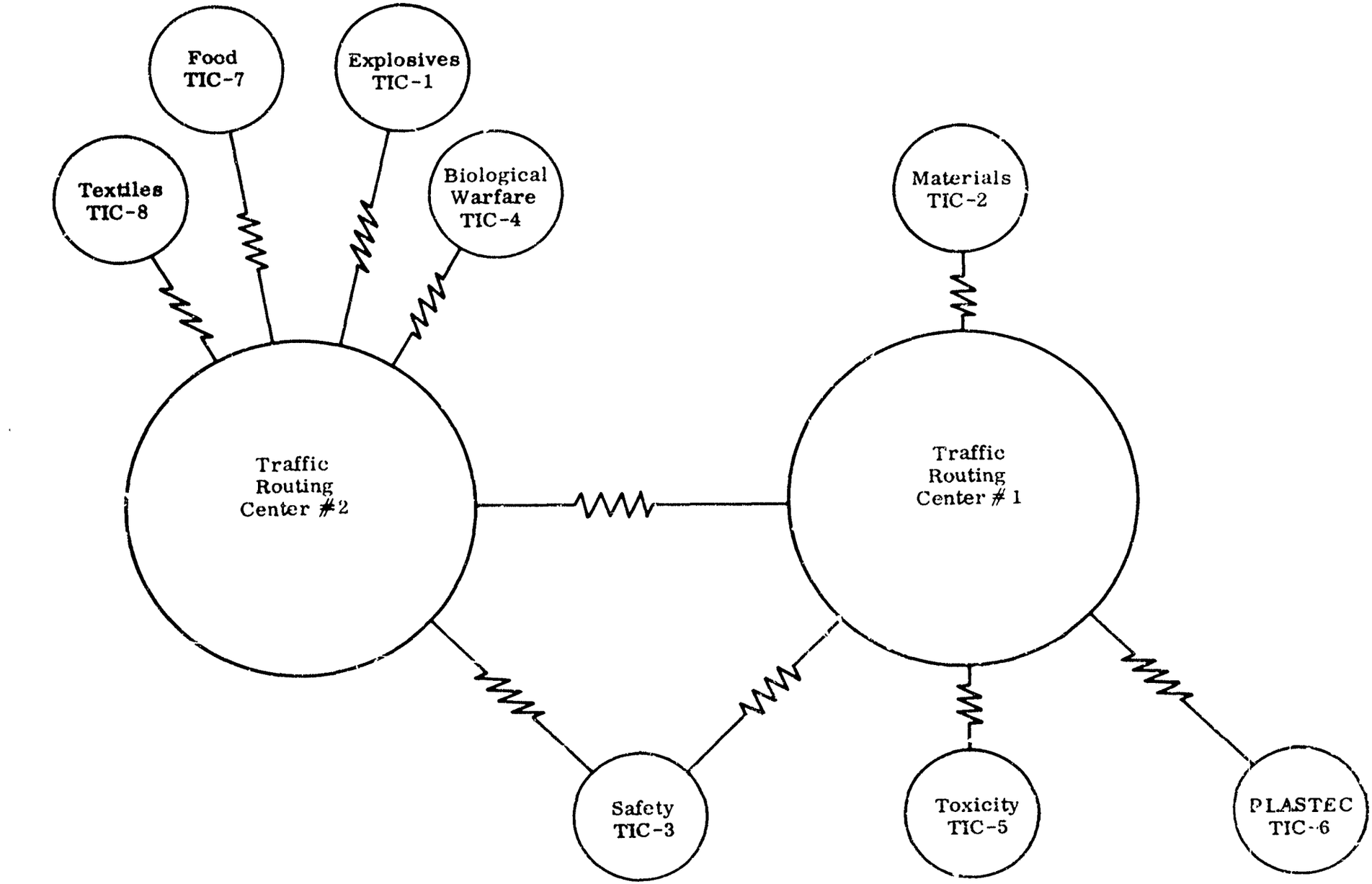


Figure 6-7. A Network of Technical Information Centers

6.4.2.1 CIDS Traffic Routing Center. The system design for a TRC has not been fully developed. Some of the files which might possibly be maintained by each TRC are described below.

6.4.2.1.1 Master File of Chemical Compounds. The TRC would coordinate at one location all of the common chemical information and data of general interest to Army programs. It is contemplated that some data would be stored on 2-1/2 to 3 million different chemical compounds. Precisely what type of data should be stored at the TRC is yet to be determined. There are at least 1,000 different characteristics which might be stored on each compound including, for example: structure, name, formula, density, melting point, boiling point, specific heat, solubility, toxicity, physiological effects, military applicability, security classification, bibliographic citations, test data, sources of data, releasability, and the like.

6.4.2.1.2 Index to Technical Information Centers. An important file to be maintained at the TRC would be an index to the information and data content at each of the TICs that are a part of CIDS. Since all of the available information would not be stored at the TRC, one of its functions is to direct a user to the information center or other source from which the desired information might be obtained. This concept is analogous to the functions of the National Referral Center at the Library of Congress. It would be necessary to store some type of index to the various files maintained at the TICs. The index would include compound identification data, location of supplemental information, releasability, security, and the like.

6.4.2.1.3 Index to Research in Progress. Another file which might possibly be maintained at the TRC is an index to research and progress in the field of chemistry. It is not enough to know what has been accomplished and reported on in the past. It is also necessary, for many users, to know what is being done in chemistry, i.e., who is doing what, where, and for whom. This is similar in concept to the functions of the Science Information Exchange operated by the Smithsonian Institution.

6.4.2.1.4 Register of Professional Personnel. The TRC might also be the repository for a master scientific manpower resource register in the field of chemistry. Such a register might be limited only to Army scientists, or it might include all DOD personnel,

or even DOD contractor personnel. The data about each person listed might include age, education, experience, fields of interest, projects (contract number) completed, and projects in progress. Such a file might serve as a basis for a Selective Dissemination of Information (SDI) system; however, there are many technical problems which would have to be considered before the feasibility of such a system could be determined.

6.4.2.2 User-CIDS Interface. The manner in which the user will interface with the system has not yet been established. The TRC concept offers a number of possibilities, however. Figure 6-8 illustrates some hypothetical interface relationships between the user and the various system components.

In this hypothetical example, the user prepares his query according to printed instructions which are available to him. If he is interested in specialized information available only in a particular TIC and he knows the information is stored there, he (User-D) will direct his query (Q_3) to that TIC. On the other hand, if he does not know this information is available, or where it is available, he (User-B) will direct his query (Q_2) to the Traffic Routing Center. The TRC will analyze his query (Q_2) and automatically determine that TIC_5 is best equipped to handle the problem and will route it to TIC_5 over the communications network. TIC_5 will analyze the query and communicate its response (R_2) directly to User B. The response may either be an answer, a request for more detail, or a request for an explanation of the request. If TIC_5 is actually unable to handle the query, it will feed the query back to the TRC so that it can be rerouted to the proper TIC (in this case TIC_4) so that the central index at the TRC, to the information files of all technical information centers can be updated. Another alternative would be for the user (User-C) to send his query (Q_5) to the nearest TIC (TIC_3) which will forward it to the TRC. The TRC then routes it to the appropriate TIC (TIC_4) which analyzes the query and transmits its response (R_5) directly to User C.

6.4.3 Recapitulation and Analysis

The CIDS Traffic Routing Center is still in the planning stage. The concept, however, appears to have merit for large scale missions that have numerous information processing or evaluation centers. It is similar in some respects to the system operated by the Science Information Exchange (SIE) of the Smithsonian Institution. The SIE is an

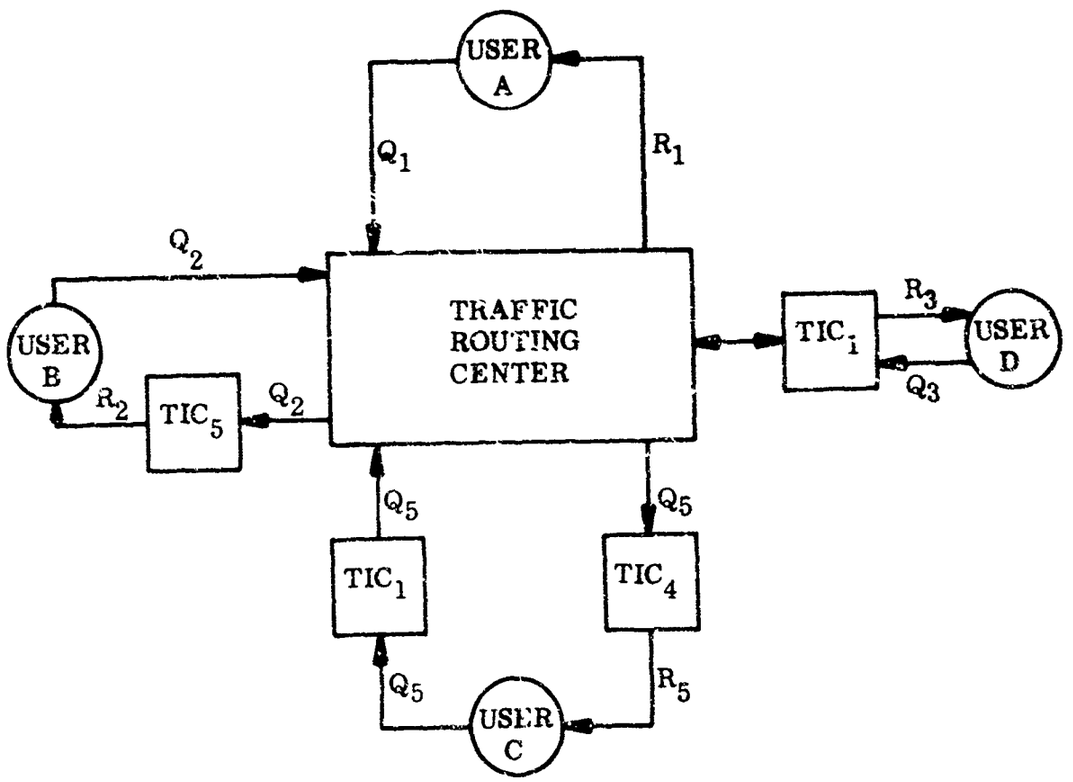


Figure 6-8. User — TIC — Traffic Routing Center Interfaces

index to research grants and can tell a requestor who is doing what kind of work in a given field. The TRC concept and SIE are primarily fact retrieval systems operating almost entirely in the retrospective mode. There is no announcement function or other type of current-awareness service planned in either case, at least not at the moment.

Being primarily a fact retrieval system, computers are indicated as the primary storage medium. Since many types of requests are unpredictable, it may be desirable to provide a general query language. With a query language, the requests can be stated and automatically compiled into search programs. Special programs for each "special" request do not have to be written.

Since there will be a network of information centers, an electronic data communications system will be required. There may also be a requirement for facsimile transmission using the same communication lines that are required for digital data communications.

The storage medium for the master file of chemical compounds is likely to require some form of random access equipment. Since there are already 2.5 million compounds and as many as a thousand properties for each, it would probably be inefficient to search serially through the entire file for each question. Random Access file structures for fact retrieval problems are described in Paragraph 8.4.3 and Appendix B.

6.5 ENGINEERING DATA CENTER

The Department of Defense has over 70 million engineering drawings on file. This collection is growing at the rate of six million drawings annually. In addition to these drawings, there are associated lists of material, indexes to drawings, as well as standards, specifications, test reports and reliability data. All of these constitute engineering data. One of the most significant decisions made by the Department of Defense is a requirement that all DOD contractors furnish DOD with engineering drawings from DOD projects on 35 mm microfilm.

The objective of this section of the report is to describe the functions and operation of a collection point within the military. This point receives, processes, and distributes engineering drawings to various user groups within the Defense establishment. The organization to be described herein is the Naval Air Technical Services Facility (NATSF). The Naval Air Technical Services Facility is the central repository for all engineering data owned by the Bureau of Naval Weapons. It services some 38 different Government activities.

6.5.1 User Community

The User Community which is serviced by NATSF is made up primarily of overhaul and repair (O & R) stations for particular weapons systems and other naval property, as well as "Naval Purchasing Activities." The people utilizing this data are involved in the functions of purchasing, maintenance, maintenance engineering, or reliability analysis and improvement. They deal with existing items rather than with the design of new systems. The NATSF system is not intended to service design engineers who might like to know what existing part meets particular design parameters. On the other hand, it is intended to service such requests as "give me all the drawings on wing assembly 17 of the B-27 aircraft." For this reason, the NATSF system is primarily a document IS&R system rather than a fact IS&R system, as would be the case with design engineering information.

6.5.2 Inputs and Outputs

Figure 6-9 illustrates the inputs to and the outputs from the Engineering Data Management Department of NATSF. The present form of input to the system is governed by Weapons Regulation 12 (WR12 - Notice 1) which specifies that all drawings shall be furnished on 35 mm microfilm along with a "slave" deck of punched cards, which contains the drawing number, manufacturer's code, roll number, frame number and model number, and also a data deck, which contains three subdecks - an index deck, a document deck, and a vendor deck.

An index deck shall be furnished for each aircraft, weapon, component, or equipment that is a line item under the contract. An index deck shall consist of one card for each of the following: the top assembly drawing, assembly drawings of assemblies and subassemblies down to the last assembly.

The document deck shall be prepared for each card included in the index deck. For each assembly, there shall be a document card following the applicable header card(s) for each of the following:

(1) Applicable assembly drawing.

(2) All drawings applicable to that assembly.

(3) All company specification control and source control drawings.

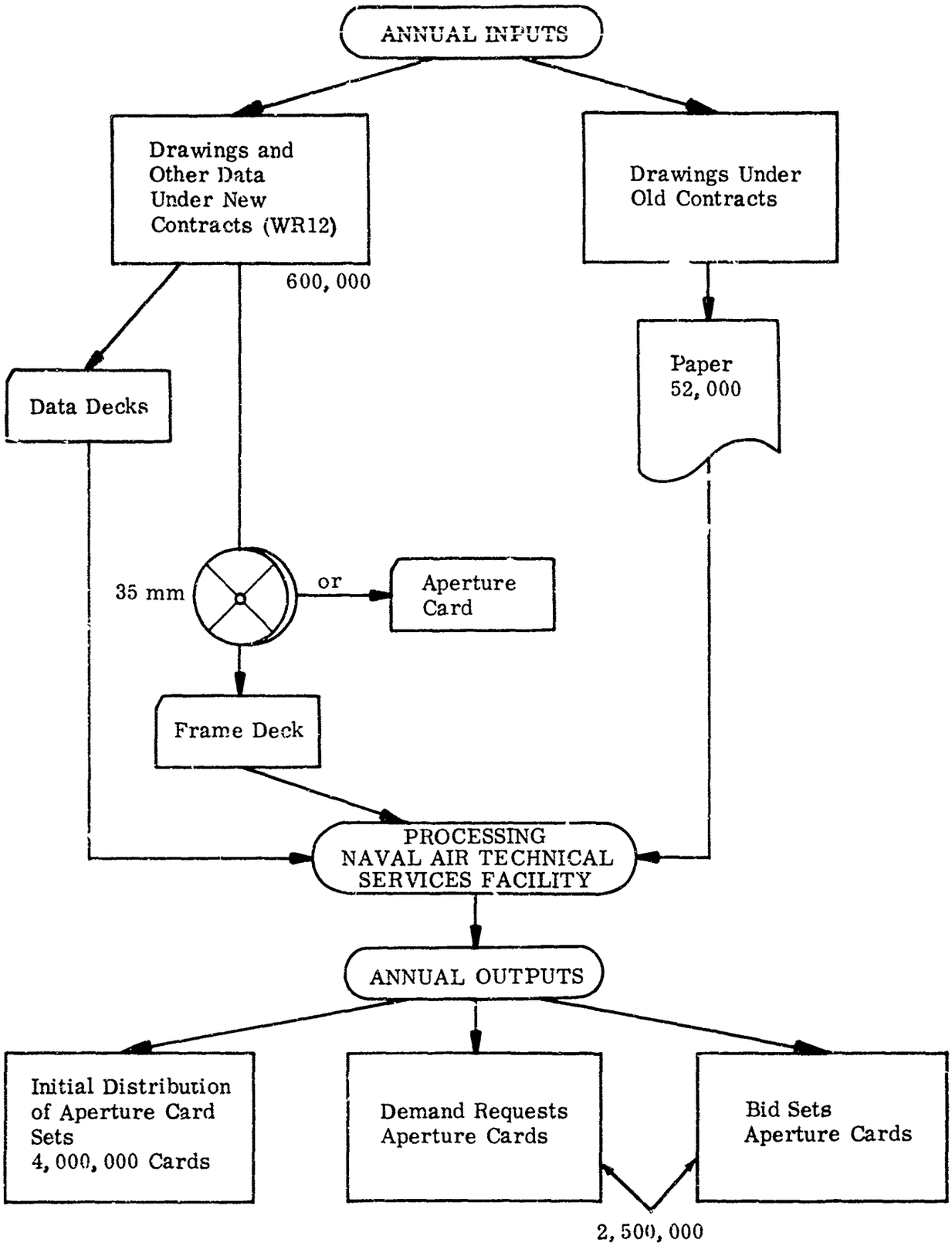


Figure 6-9. Input-Output Diagram

(4) All company parts specifications and standards applicable.

(5) All applicable vendor drawings where specification or source control drawings do not apply.

(6) All critical company specifications and standards applicable to the assembly.

(7) The government specification applicable to the assembly, if any.

The vendor deck shall consist of a card for each vendor on all specification control and source control drawings, and for each vendor where specification or source control drawings do not apply.

The specification deck shall have a card for all company, government, and industry specifications and standards used on the model represented by the index deck.

The lower portion of Figure 6-9 illustrates the outputs of the system. NATSF distributes complete sets of drawings to some 38 government activities. In particular, there are seven (O&R) stations, and each O&R station receives a complete set of drawings for all aircraft and equipments they maintain. NATSF distributes approximately 4,000,000 aperture cards per year on initial distribution.

In addition to the initial dissemination, NATSF disseminates approximately 2,500,000 aperture cards per year to meet demand requests for a single drawing of a line item, complete sets of a particular assembly or system or bid sets to support the purchasing function.

6.5.3 Files

NATSF maintains three main files of drawings or associated data. The first is the microfilm aperture file which is arranged by drawing number and contains drawings, materials lists, specifications, standards, and other data. The second file contains paper copies of drawings either in tracing or Vandyke form. There are approximately 2,000,000 drawings that are still in paper form. These are grouped by drawing size, A, B, C, D, etc., and then by drawing number. The third file contains paper copies of manufacturer's specifications and standards which for the most part are A size.

The paper drawing files were accumulated for the most part before WR12 went into effect. The drawings which are presently being received in paper form are generally from contracts which were entered into before WR12 was issued.

In addition to the drawing files, NATSF maintains enormous files of punched cards containing the data decks described above, plus the slave decks which are utilized to make complete sets of drawings.

6.5.4 Input Processing

Figure 6-10 illustrates the input and request processing which takes place at a typical engineering data center (NATSF). The camera negative received as input is duplicated on a roll-to-roll Kalvar duplicator (FME) and the camera negative is then saved for archival purposes. The resulting positive copy on Kalvar film is mounted into aperture cards on an automatic aperture card mounter (Filmsort). The Kalvar positive aperture cards are then duplicated into blank Kalvar cards in the quantity required for initial distribution as determined from the distribution list. The data from the input slave deck is then reproduced into each set of aperture cards and then each set is interpreted on an IBM interpreter. The duplicate copies are then disseminated and the Kalvar positives and slave deck are filed for future use.

6.5.5 Request Processing

6.5.5.1 Request for a Line Item. A normal demand request is serviced by pulling the aperture card, making an aperture card duplicate, reproducing and interpreting the punched data, and delivering the duplicate card. Where there is no aperture card on file, the tracing will be pulled and an entire sequence of operations from planetary camera through the preparation of a Kalvar negative duplicate aperture card will be followed to furnish the user with an aperture card and, at the same time, create an additional master positive aperture card and thereby destroy one more paper drawing. This is one method of handling the backlog problem where there are mixed paper and film files. The unit record filing advantages of aperture cards makes this "as requested" method of conversion possible. If no tracing is found, a search will be undertaken and if necessary a copy of the drawing will be requested from the manufacturer.

6.5.5.2 Request for a Set of Drawings. This type of request involves the identification of individual drawing numbers within a set. This is accomplished by searching the index deck by name of the assembly for the assembly number. Using the assembly number, the header card is located and all of the detail cards following this header card are pulled. This set is then reproduced into blank EAM cards or is listed on a tabulator. The drawings are then pulled, duplicated, reproduced, interpreted, and dispatched as shown in Figure 6-10.

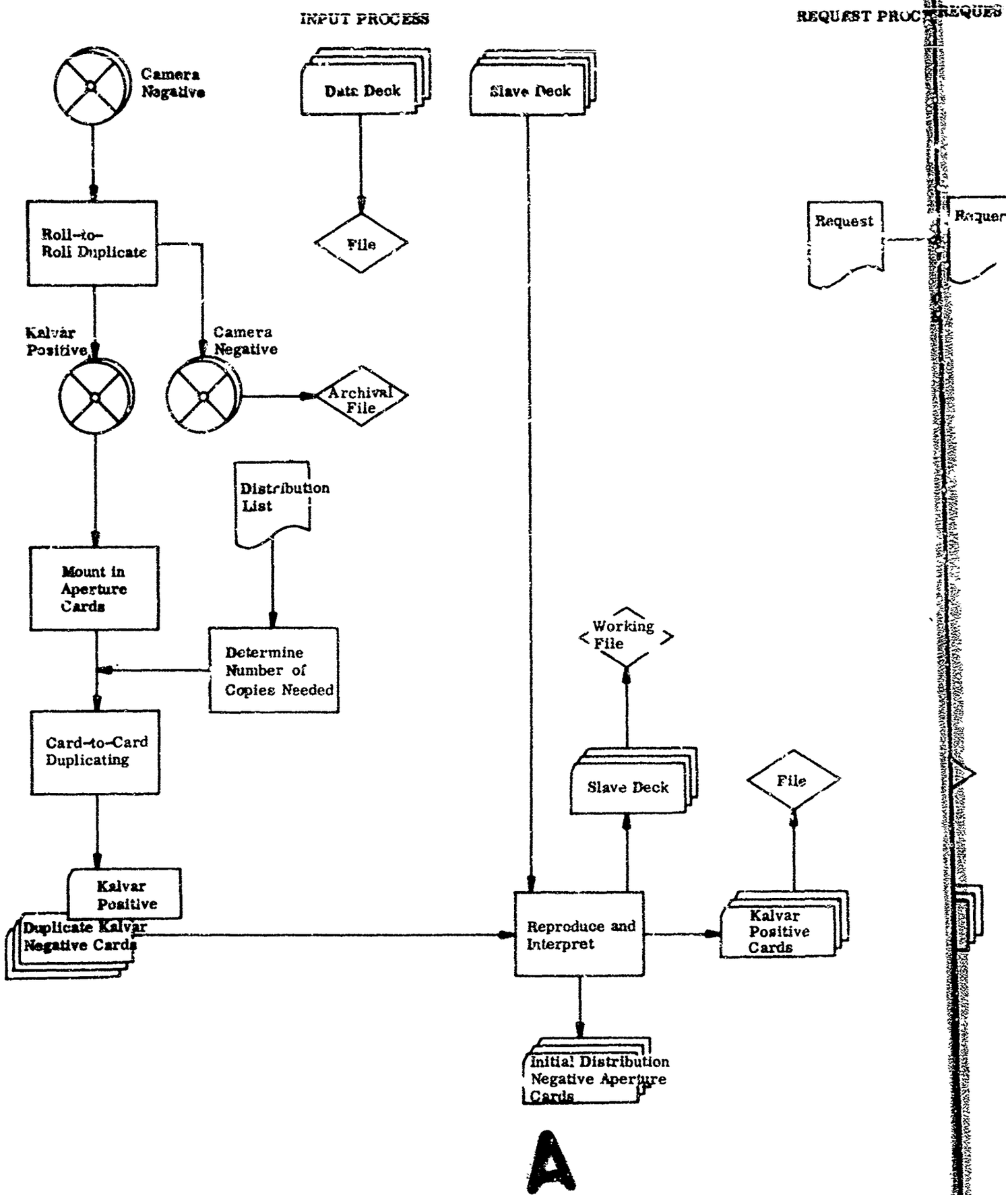
6.5.5.3 Request for Bid Set. The aperture card is becoming a popular medium for furnishing engineering drawings and data to prospective bidders on Government contracts. NATSF acts as a service center to various Government purchasing activities, in the preparation of bid sets. This task is handled like a line item if the request is for one drawing with the exception that multiple copies of the drawing are made. It is handled in the same fashion as a request for a complete set where a set of drawings is involved except that multiple copies of the entire set are made.

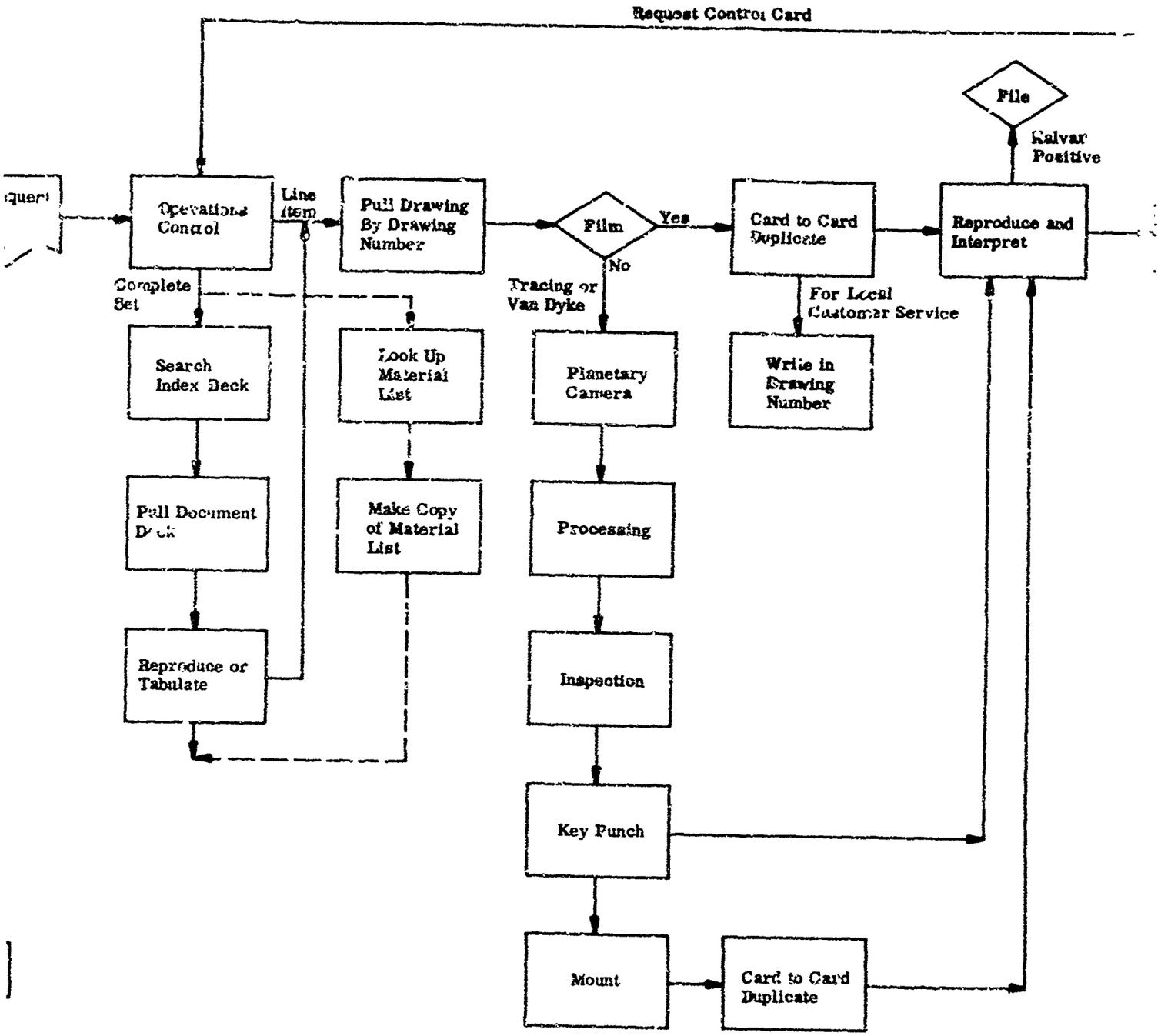
6.5.6 Recapitulation

The NATSF system is essentially a central service for receiving, duplicating, and disseminating copies of engineering drawings and associated data both on initial distribution and on request. The only indexes which are maintained by NATSF are nomenclature indexes and generation or top-down breakdown indexes which identify all of the drawings associated with a particular assembly or subassembly within a particular system or major component. The organizations serviced by NATSF also render demand copying services for individual drawings and even for sets of copies. However, the satellites of NATSF do not generally maintain a copy of the data deck and hence cannot readily furnish a copy of an entire set of drawings without the drawing numbers.

Neither NATSF nor its satellites, however, are able to aid the design engineer in his search for a manufactured part meeting certain criteria. This task must presently be done by referencing the manufacturers' catalogs or by other evolving systems for part selection. The design engineer may also refer to the Federal Supply Catalog or the Defense Industrial Supply Catalog under the appropriate class of items.

A point worth noting is that requests for copies of drawings are met entirely by on-demand copying. There is no pre-stocking of aperture cards for the purpose of


INPUT PROCESS
REQUEST PROCESS
Camera Negative
Data Deck
Slave Deck
Request
File
Roll-to-Roll Duplicate
Kalvar Positive
Camera Negative
Archival File
Distribution List
Mount in Aperture Cards
Determine Number of Copies Needed
Working File
Card-to-Card Duplicating
Slave Deck
File
Kalvar Positive
Duplicate Kalvar Negative Cards
Reproduce and Interpret
Kalvar Positive Cards
Initial Distribution Negative Aperture Cards
A



B

Figure 6-10. Engineering D
Input and Request Proc

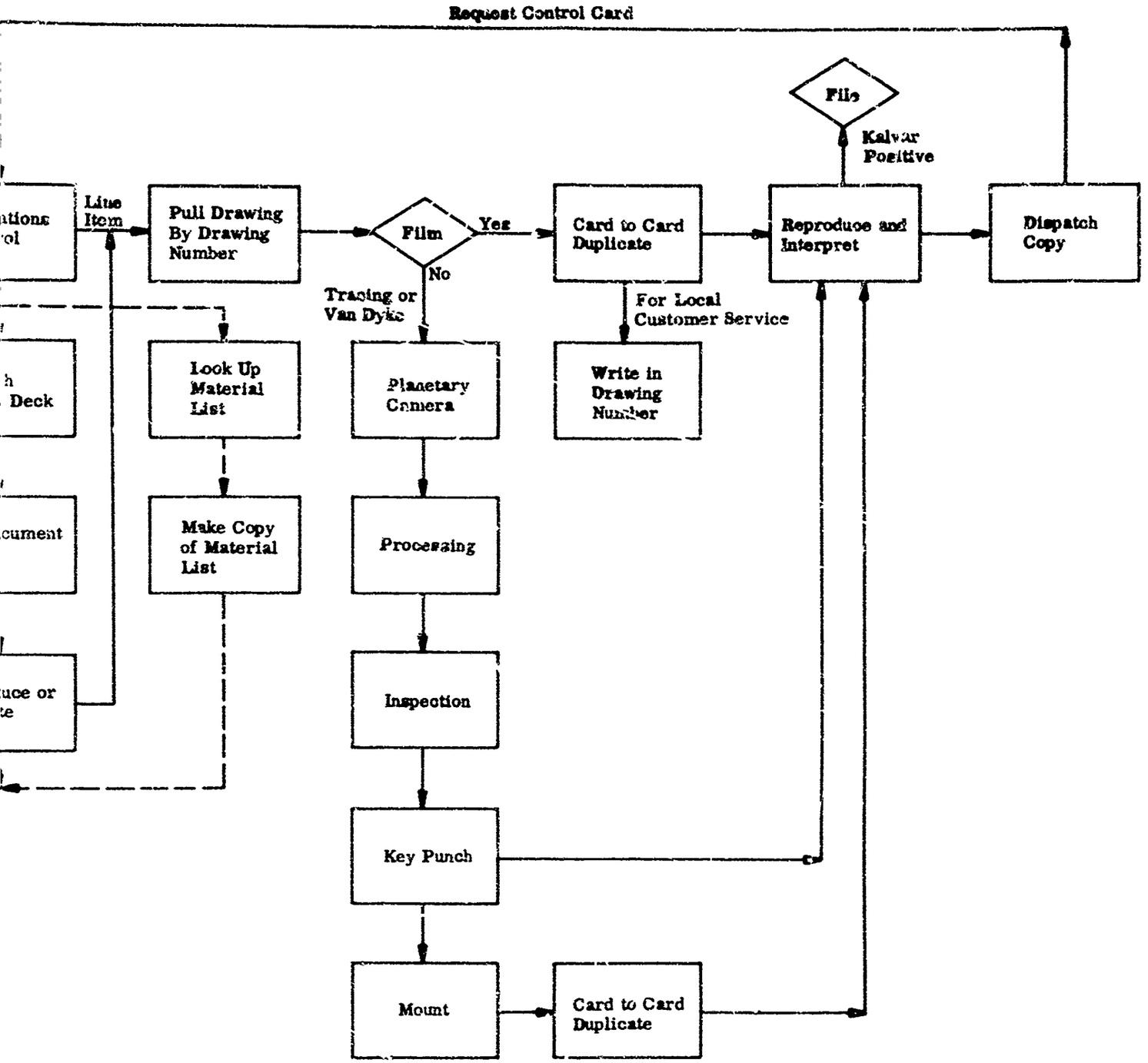


Figure 6-10. Engineering Data Center Input and Request Processing

meeting demand requests. There is a general policy of furnishing the requestor with aperture cards and not hard copy. The user is expected to be able to make his own hard copy from the aperture card. The activity within the file is very low since a total of only 2,000,000 requested copies (including many multiple copies) is produced each year from a file of approximately 4,000,000 drawings. Furthermore, the cost of producing one aperture card on demand is not too much higher than producing multiple copies for inventory or distribution.

6.6 REAL ESTATE TITLE SEARCHING SYSTEMS

6.6.1 General

The recording statutes were enacted to enable the buyer of real property to be sure he is obtaining an authentic and unencumbered title. Legal rights may be lost by failing to properly record a deed, mortgage, lien, or judgment. For example, if a prospective buyer of real estate thoroughly searches the records of deeds, mortages, etc., and finds no encumbrance, the law provides that his title is good against a party who has earlier obtained a deed or judgment on the same property but has failed to record same. The purpose of the recording statutes is to provide notice of ownership or claims to anyone who takes the trouble to examine the public records. The historical and even present difficulties in searching county real estate records created the need, or at least the market, for real estate title insurance. Title insurance companies have established filing systems to facilitate the making of searches which are more elaborate than those of the majority of county clerks' offices. The title company, in addition to making a search, insures the accuracy of its search up to the amount of the purchase price of the property.

6.6.2 Typical Search in a Recorder of Deeds Office

Figure 6-11 illustrates how an attorney (or layman) makes a search of the deed and mortgage records in a typical county recorder's office. Of course, in addition to deeds and mortgages, he must search the judgment and lien files as well as the delinquent tax records in the Department of Collections. The traditional method, which was the only method utilized until around 1925, was to search by the name of the grantor, grantee mortgagee, or mortgagor in large ledger books. The grantee-grantor indexes are generally arranged by letter of the alphabet and by year. In order to make a search

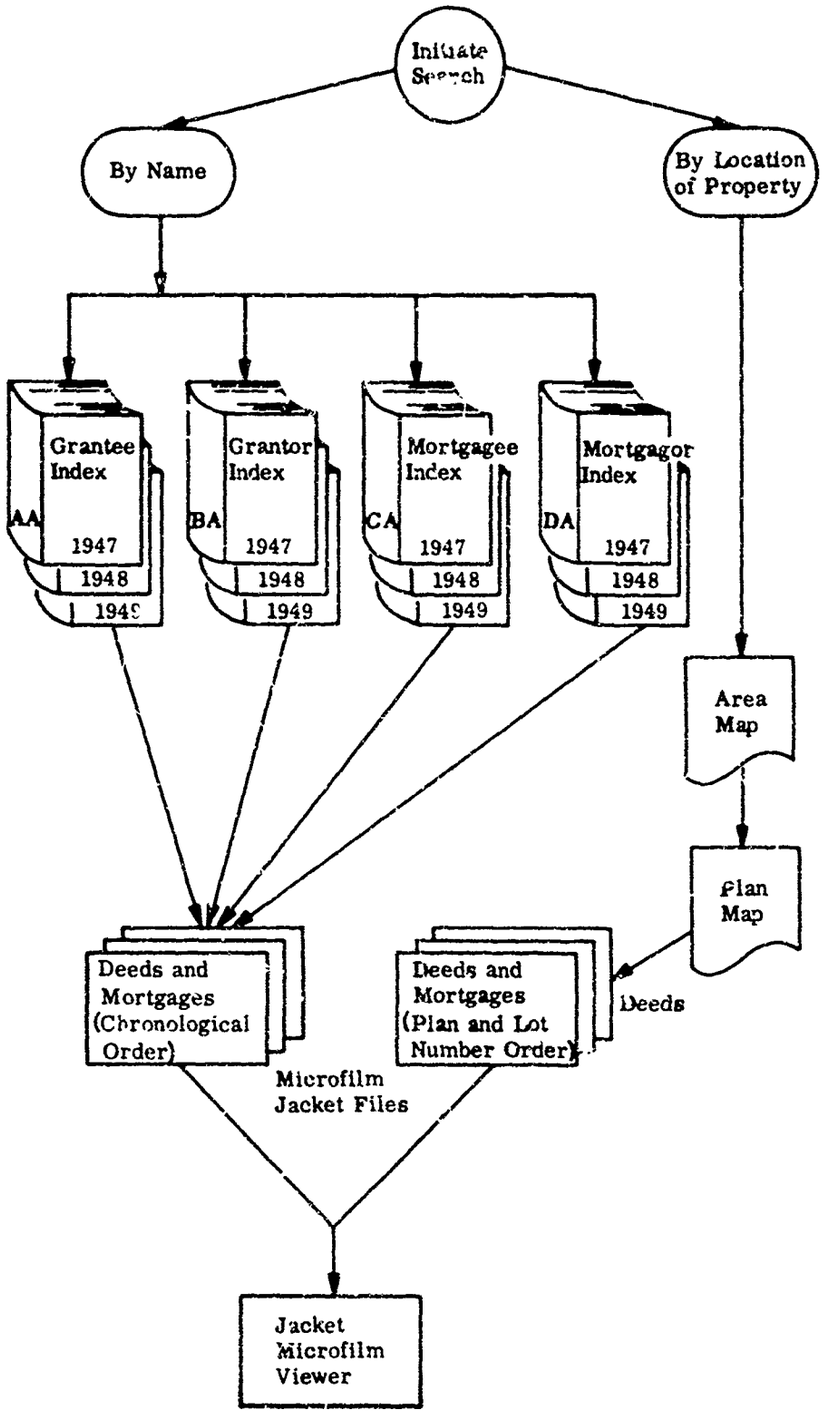


Figure 6-11. Title Search in Typical County Recorders' Office (Deeds and Mortgages)

you pull the most recent volume for a particular letter or group of letters; e.g., if you are looking for the name "Farino," you will look in the volume FA-FY, 1964. There may be 20 or 30 pages of FA's in this volume and you will have to scan the pages serially for the name "Farino." You will do this in a different volume for each year that you wish to search until you find all of the entries. The entry merely gives you a book and page number. You will then go to a microfilm jacket file which is arranged in book and page number (accession number) order, borrow the appropriate jacket and view the document on a microfilm reader.

The City of Philadelphia has a somewhat improved system which is used when you know the location of the property. In this case, you will first go to a map and identify the approximate location of the property and mark down the plan number. You will then ask a clerk to give you a copy of that plan on which you will locate the lot number of the property. You will then ask the clerk to give you the microfilm jacket containing all the transactions for that particular plan and lot number and will view this microfilm on a viewer. Obviously this is a much more effective system. It exists only for deeds, however, in the City of Philadelphia. For mortgages, you must search by the mortgagor-mortgagee index. For judgments, you must search a separate judgment index by the name of the seller of the property which must first be translated into a modified Soundex code.* Liens are filed by name or property. In addition to searching deeds, mortgages, judgments, and liens, the attorney must make a final search of delinquent tax payment records in the Department of Collections, as a delinquent tax does not become a lien unless it is two years past due. The delinquent tax records are extremely difficult to search because they are arranged on ledger sheets by street name and by legal description within street name rather than simply by street address.

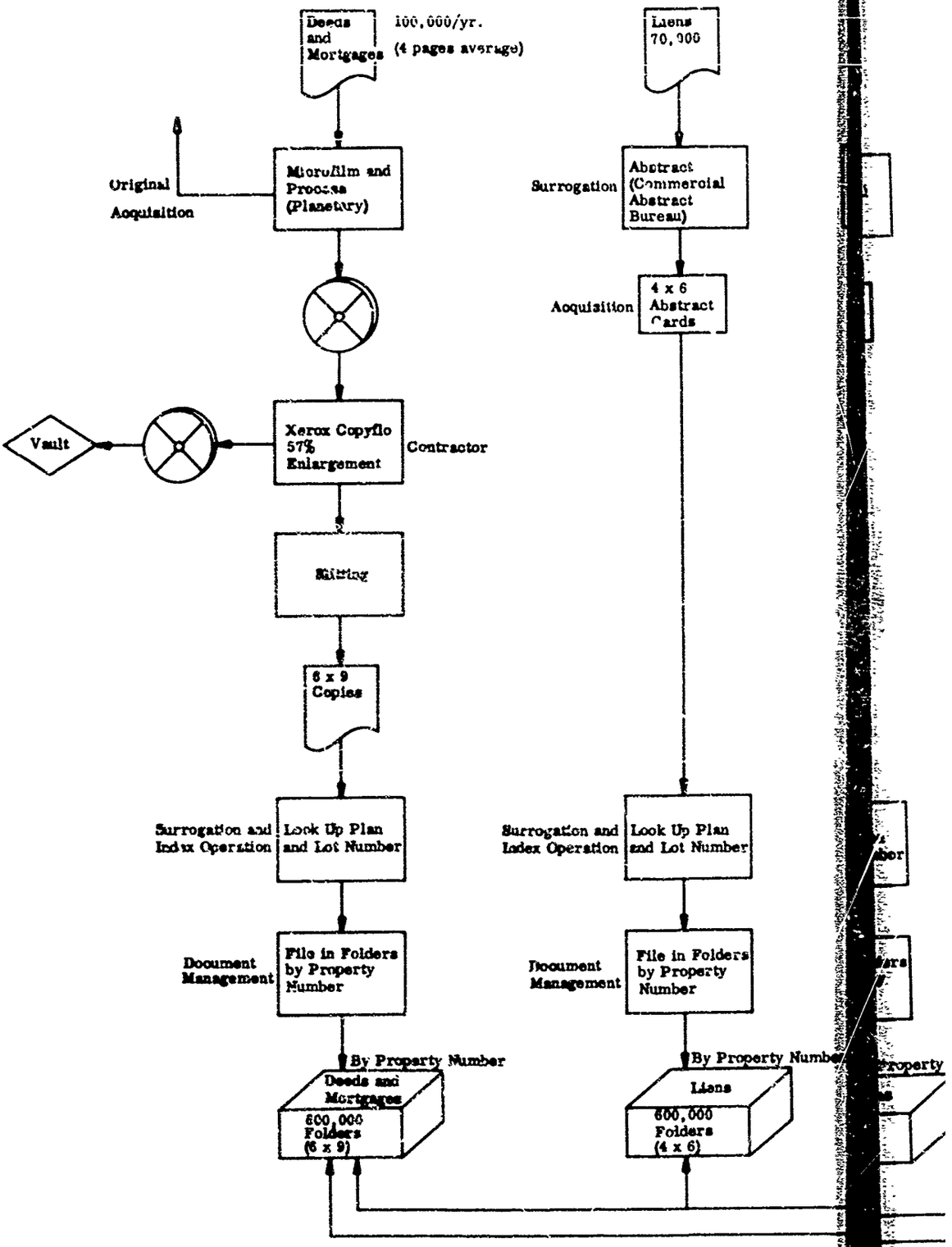
6.6.3 Typical Title Company System

Figure 6-12 illustrates the general system employed by a Title Insurance Company in Philadelphia.

6.6.3.1 Input Processing. It can be noted from Figure 6-12 that there are five primary input sources to the title company's files. These are deeds and mortgages, liens, judgments, reports of title (from earlier searches), and satisfactions of judgments and mortgages.

* Soundex is an indexing system developed by Remington Rand for filing groups of similar sounding names under the same code. For example, the name Moran is coded M-650 and Marahan is also coded M-650. This avoids not being able to find an item because of an error in spelling.

INPUT PROCESS



A

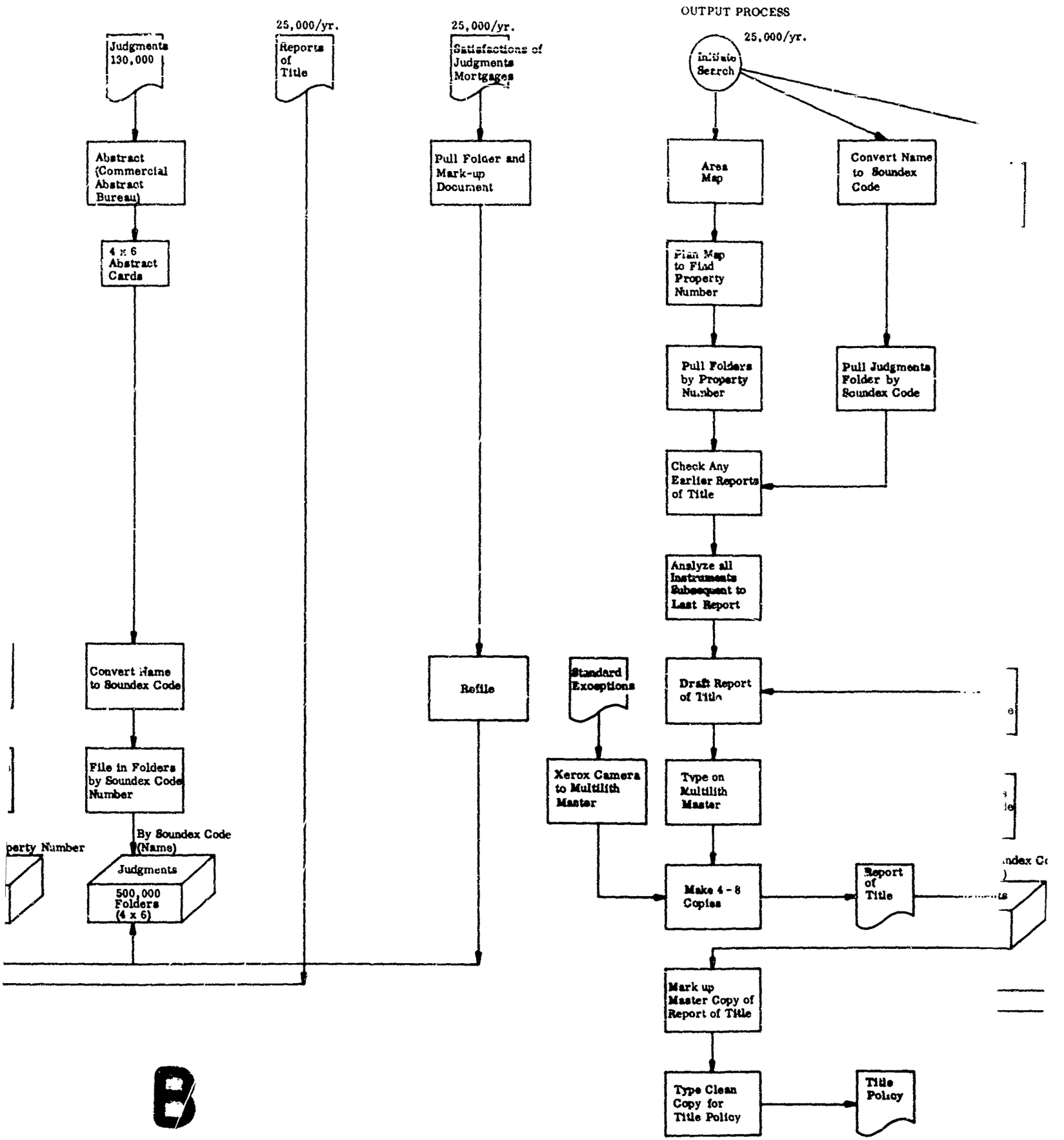


Figure 6-12. Typica

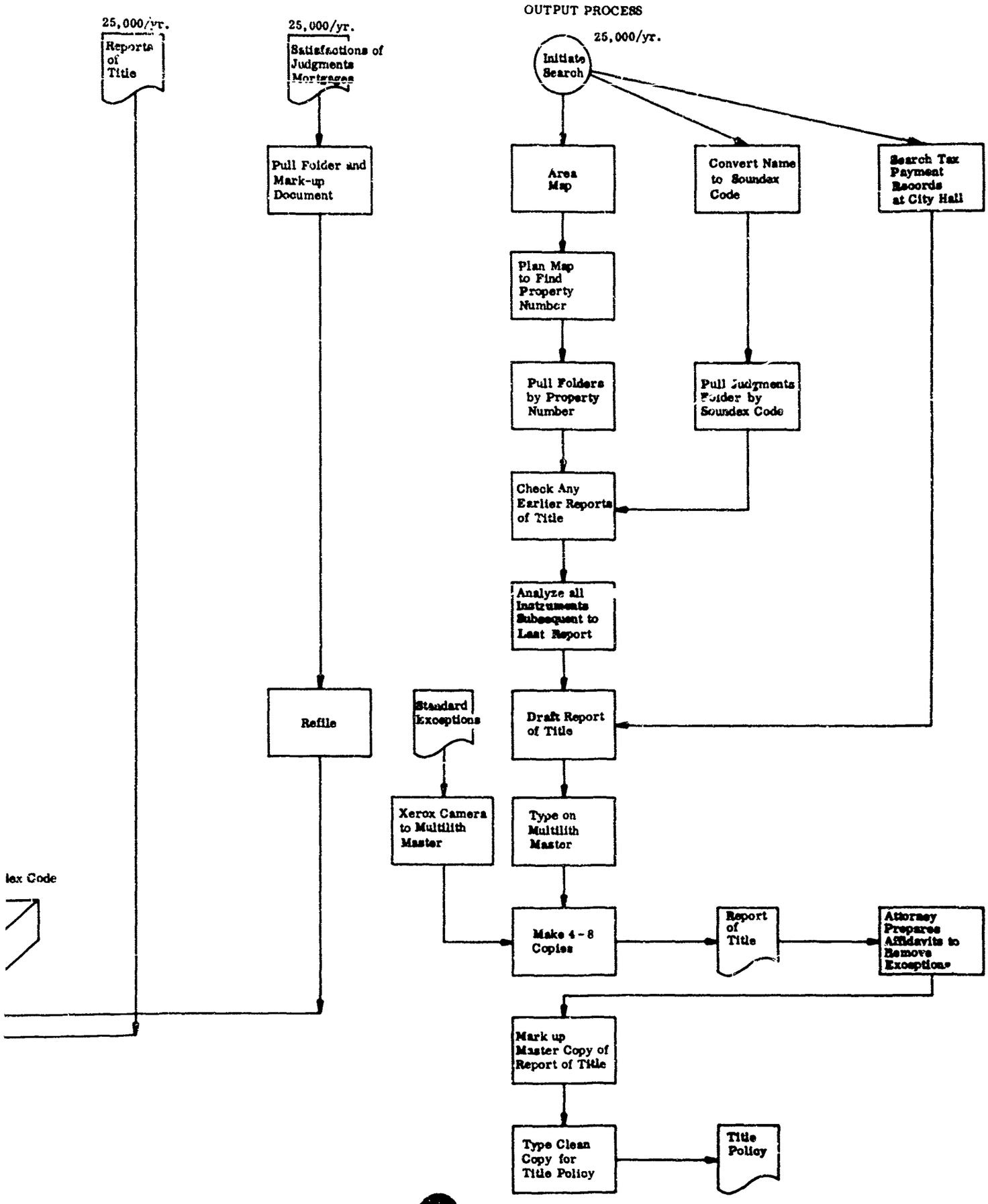


Figure 6-12. Typical Title Co. System

An important point to note is that this title company maintains all of its files in legible, although not full size (57 percent), hard copy.

Approximately 100,000 deeds and mortgages, averaging four pages each, are microfilmed each year on a planetary camera. This microfilming is done by the company at the city recorder of deeds office. The company utilizes the services of a contractor to produce 57 percent of full size enlargements from the microfilm using a Xerox Copyflo. The resulting 6 x 9 inch pages are slit and stapled and the original microfilm is sent to a vault for security. The plan and lot number of each document is determined by analysis of the legal description and reference to maps. The hard copies are then filed in folders by property (place and lot) number.

Liens and judgments are handled differently. The original records are abstracted by a commercial abstracting service which produces 4 x 6 abstract cards and sells these to various title companies. Liens are filed by property number and judgments are filed by a Soundex code derived from the name of the party against whom the judgment has been entered.

The total activity against all files is approximately 425,000 accesses per year, of which approximately 350,000 are for input and 75,000 (3 files x 25,000 searches) are for output or searching. This illustrates the generally typical high ratio of input costs to output costs in most information systems.

6.6.3.2 Output Processing. The typical output of the title search system is a report of title, which after review and modification based upon negotiations with the customer's attorney, will be incorporated in the title insurance policy. The search will always be by property number except for judgments and delinquent tax records which are always by name, or derived code. The property number is obtained by reference to maps and plans and the relevant folders are manually retrieved from the files. The first thing that the searcher will look at is the latest report of title in the file, if there is any. This will reduce the amount of work he will be required to do since he need only bring it up to date. The searcher must be able to analyze the legal descriptions of the documents, which frequently are stated in metes and bounds, as well as their legal form.

The report of title is typed on a paper offset master, and generally only four copies are produced. If the particular title has a standard exception such as the easements, which are frequently reserved against all properties in a particular housing development, these will be incorporated into the report of title by placing them on multilith masters by means of a Xerox plate-making camera.

The one aspect of title searching which is not done at the title company, in this instance, is the so-called "date-down" search for delinquent taxes. This search is made on the day before the settlement by employees of the title company who are stationed at the Department of Collections of the City. It does not presently pay the title company to keep an up-to-date record of taxes because of the extensive number of transactions and the fact that there is no machine language record of tax payments and delinquencies.

6.6.4 A Punched Card System for Improved Grantor-Grantee Indexes

The primary difficulty in searching a grantor-grantee index of the type produced by the City of Philadelphia is the fact that they are not in strict alphabetic order and are not cumulated over a period of years. To cope with this problem, a punched card system was developed for the Norfolk County Registry of Deeds, at Deedham, Mass.[38] By utilizing punched cards, each index entry becomes a unit record which can be sorted, merged, and listed automatically on tabulating equipment. Figure 6-13 illustrates this system. Microfilming and Xerox Copyflo enlarging is utilized to produce a permanent hard copy of the document which is bound in a record book. The document is abstracted and coded. The descriptive abstract and coding is then keyboarded at an IBM 826 which produces a typed description card and punched index card. The typed description cards are used to cover the period before a printed listing of the index is available. The grantee-grantor index cards are sorted into alphabetical order, merged with header cards for such things as common surnames, and listed on an IBM 407 tabulator. These cards are then filed and re-cumulated every five and ten years. The resulting listings are bound into a grantor index and a grantee index which contain some 350,000 lines per year. The five year index for the years 1956 through 1960 was printed, proofread, microfilmed, and bound and put on the shelves by January 30, 1961. This cumulative index totaled 20 books and allegedly would have required five typists a full year to prepare.[38]

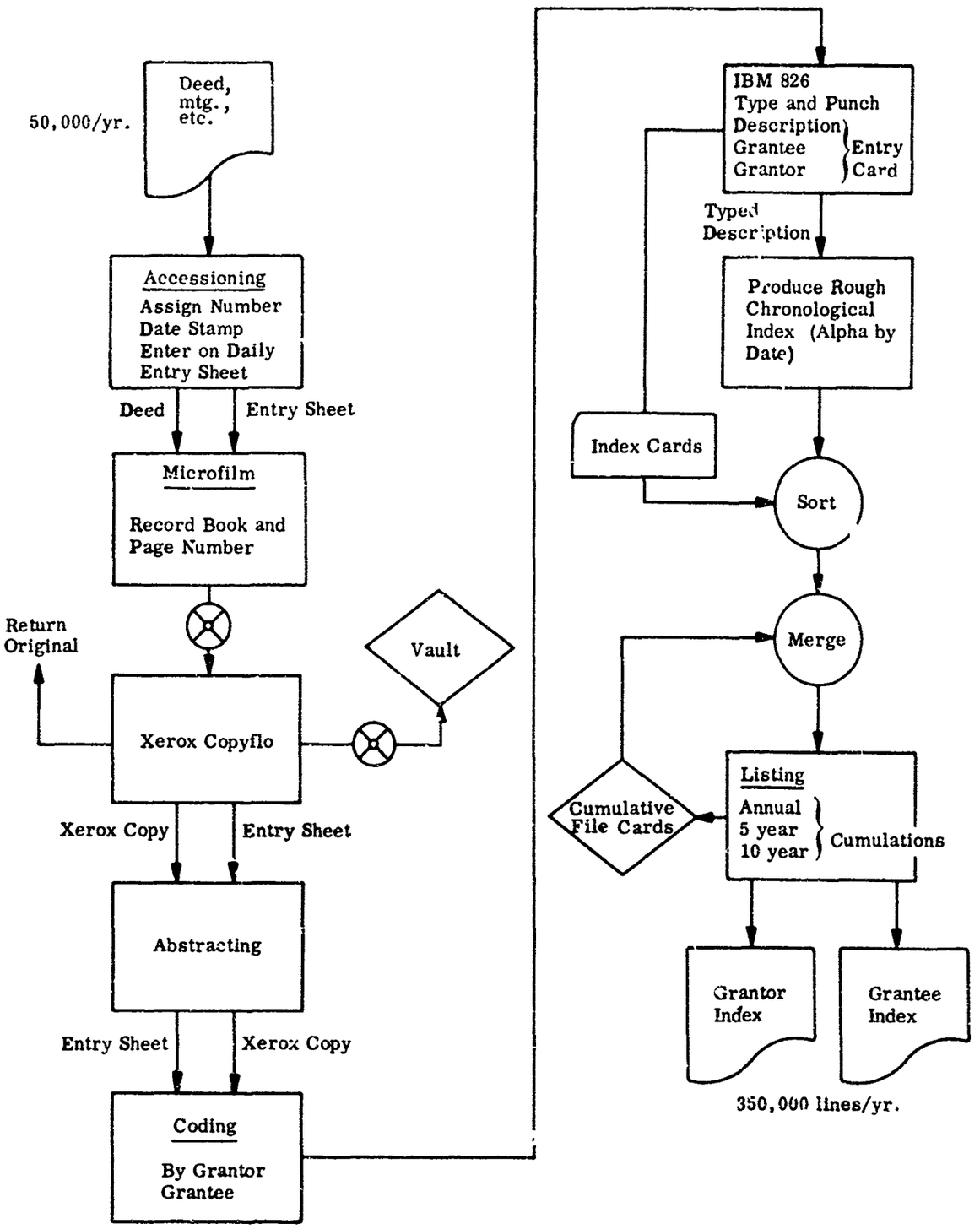


Figure 6-13. Punched Card System for Grantee-Grantor Indexes
Registry of Deeds, Norfolk County, Massachusetts

6.6.5 Computer System for Real Estate Tax Searching

None of the above organizations have attempted to modernize the method of making delinquent tax searches. There are two primary reasons for this fact. First, the particular county or city offices have not yet mechanized their accounting operations and second, the title insurance companies have only been considering the use of,or market for,delinquent tax data as related to real estate transactions.

The County of Los Angeles utilizes a Honeywell 800 computer in the County Assessors Office for its tax accounting functions. As a result, the County Assessors Office is able to furnish the Title Insurance Company of Los Angeles with magnetic tape containing new assessor role information such as parcel number, amounts of taxes, assessed valuations, assessed owners' name,and tax descriptions.[2] (See Figure 6-14.) Once the magnetic tapes have been converted to a format which is compatible with its H-800 computer, the Title Insurance Company of Los Angeles is able to utilize this information for its own delinquent tax searches for incorporation in title reports as well as for a new tax searching service which it renders to its various customers. The primary customers for this type of information are lending institutions,such as mortgage companies, insurance companies, banks and finance companies. The master tax data file which is contained on 37 reels of magnetic tape is updated daily. As a by-product of the up-dating run, the computer performs a tax data search preparing reports both for the company and for its tax service contract customers. The system handles an average of 10,000 transactions a day with a peak load of as much as 160,000 transactions per day.[2]

6.6.6 Recapitulation and Analysis

In many cases the title companies have preempted the field of title insurance so that individual attorneys will rarely search the public records, but rather utilize the services of the title companies. In view of this, the county clerks offices in those communities where title companies are most active have been slow to modernize their systems. Where title companies are inactive or considered to be engaged in the unlawful practice of law, the county clerk's systems are somewhat more effective. The title companies still appear to have a preference for eye legible copy over microfilm, in spite of its higher cost.

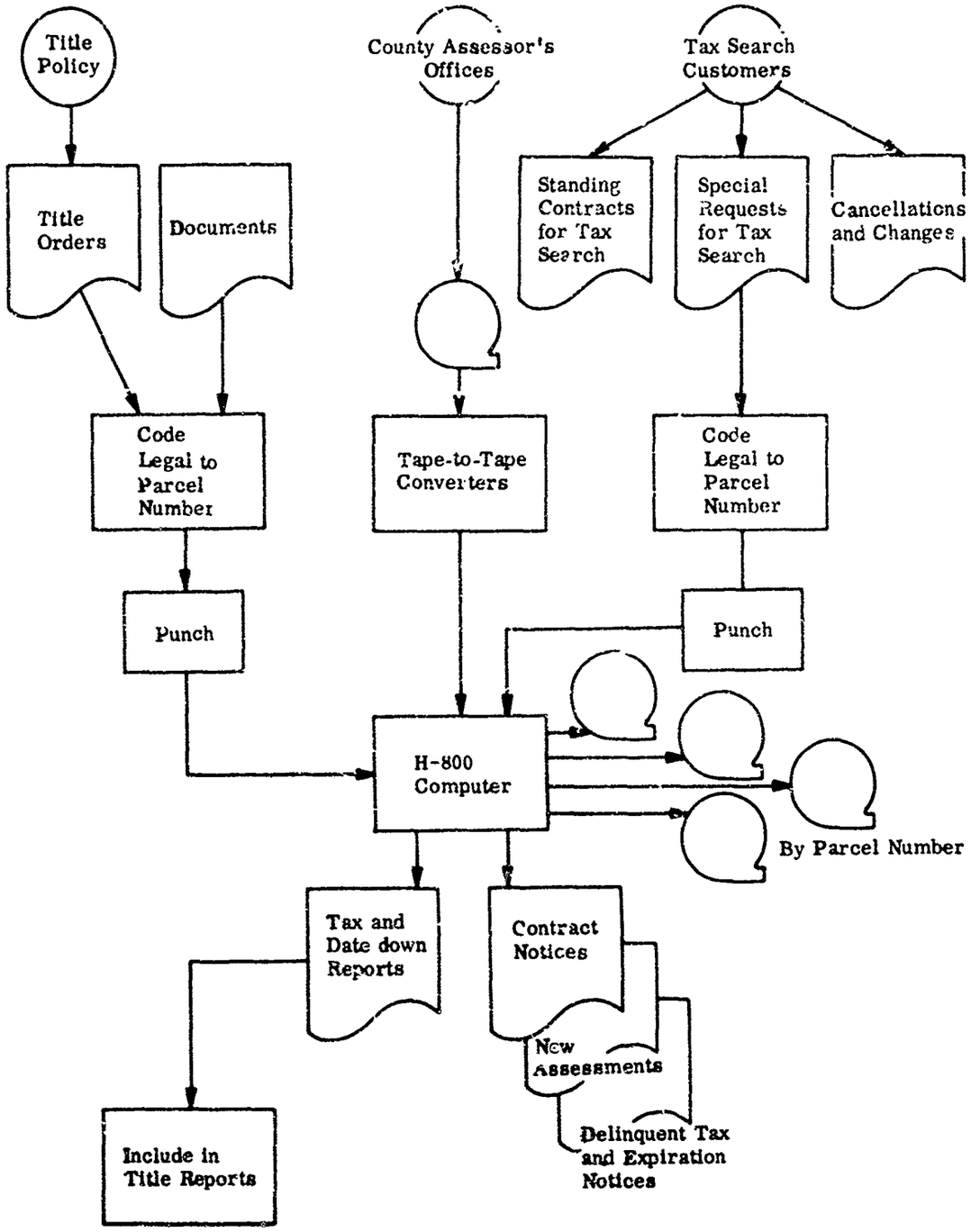


Figure 6-14. Computer System for Real Estate Tax Searching

The searching of real estate tax payment records by the Title Insurance Company of Los Angeles and by others such as the Chicago Title and Trust Company, has been made possible by the mechanization of these same records by the municipalities themselves.

The input costs for creating the up-to-date files maintained by each title company are quite high. In areas where competition among title companies is keen, this cost is becoming prohibitive. For example, in Miami, Florida, there are 10 title insurance companies, each inputting 100 percent of the total records to its own system. Since each company can only obtain an average of 10 percent of the business, this is an expensive duplication of effort. These companies are now investigating the possibility of establishing a central service bureau which would maintain the title plant. The service bureau would presumably be owned and supported jointly by the ten companies. This may, however, be an opportunity for an independent venture in other jurisdictions wherein the service bureau could offer to perform the services now performed by each individual title company at far lower cost than they presently incur by doing the job themselves. The precedent for such an enterprise exists in the commercial title abstracting companies which abstract judgments and liens and sell these abstracts to all title companies. In some jurisdictions, a commercial microfilming company will do the microfilming of the official documents at the county recorder's office and sell microfilm or Xerox copies to the various title companies.

SECTION VII. ECONOMIC ANALYSIS

7.1 INFORMATION CENTER OPERATING COSTS

Very few information centers have adequate cost accounting systems and, as a result, there is not much data available on the costs of information center operations. Fortunately, cost accounting data has been obtained from two information centers, which shall be referred to as Center A and Center B.

Neither Center A nor B can be considered to be "typical". For that matter no information center can be considered typical. They are both primarily document centers which involve that group of functions designated in Section V as System 7 — origination, acquisition, surrogation, announcement, index operation, document management, and end-use.

Center A utilizes a general-purpose computer for index operation, including bibliography preparation and other request processing, whereas Center B performs these functions manually with the aid of a catalog card file. Both centers supply copies of documents from printed inventories as well as by reproducing full size blow-back copies from roll microfilm on demand. Center A does most of its own printing, whereas Center B contracts out all of its printing to another organization.

Table 7-1 compares the annual operating costs of Center A and Center B by system function. Three separate aspects of the cost of each system function are presented: unit cost, total dollars, and percent of overall center costs. The unit costs are based on the number of titles processed, copies prepared, document requests handled bibliographies handled or catalog cards handled. The actual workload or number of units processed varied within each function. The workload represented by the cost analysis of Table 7-1 is summarized in Table 7-2.

TABLE 7-1. COMPARISON OF INFORMATION CENTER OPERATING COSTS (BY SYSTEM FUNCTION)

	Center A			Center B		
	Unit Cost	Total Dollars	% of Overall Center Cost	Unit Cost	Total Dollars	% of Overall Center Cost
INPUT PROCESSING AND ANNOUNCEMENT						
Acquisition	$ 1.80	70,939	1.8	6.13	209,127*	12.1
1. Mail Receipt and Delivery	0.21	8,708		0.10	4,177	
2. Exam, Selection, Accessioning	1.59	62,231		6.03	204,950	
Surrogation	15.74	541,768	13.9	5.49	131,561	7.6
3. Descriptive Cataloging	4.74	162,843		2.77	63,222	
4. Analysis, Abstracting, Indexing, Editing, etc.	11.00	378,925		2.72	68,339	
Announcement	13.10	417,442	10.7	9.44	272,598	15.8
5. Prepare and Proof Catalog and Journal Copy	5.11	164,779		3.09	92,490	
6. Paste up and Print Announcement Journals	6.54	200,210		4.83	134,660	
7. Prepare Card Negatives and Reproduce	1.45	45,754		1.52	45,448	
8. Miscellaneous Distribution	-	6,699		-		
Document Storage and Pre-Stocking	13.25	457,252	11.7	26.60	356,190	20.7
9. Printing and Binding Stock Copies	30.93	207,584		57.34	280,371	
10. Microphotography, Developing, and Inspection	0.60 2.94	101,363		2.30 1.96	26,254	
11. Prepare Film Copies	0.25	148,305		0.48	49,565	
Input of Index Data	6.90	211,180	5.4	1.03	30,882	1.8
12. ADP Input	6.90	211,180		-	-	-
13. Catalog Input and Maintenance				1.03	30,882	
Sub Total - Input Processing and Announcement	55.49 (30,613 Titles)	1,698,581	43.5	36.00 (27,894 Titles)	1,000,358	580

* Includes the cost of buying printed stock copies.

TABLE 7-1 COMPARISON OF INFORMATION CENTER OPERATING COSTS (BY SYSTEM FUNCTION) (Continued)

	Center A			Center B		
	Unit Cost	Total Dollars	% of Overall Center Cost	Unit Cost	Total Dollars	% of Overall Center Cost
REQUEST PROCESSING						
Requests for Document Copies (907K)	$ 1.76	1,597,679	40.8	1.02	505,911	29.4
1. Mail Receipt and Dispatch	0.03	33,227		0.94	23,850	
2. Identification Search	(ADP) 0.28	276,894		0.15	77,929	
3. Special Processing	0.75	215,657		0.20	106,035	
4. On Demand Copying	1.74	736,837		1.97	165,194	
5. Storage of Stock Copies	0.06	31,631		0.03	12,681	
6. Assemble and Process for Mailing	0.33	303,423		0.24	120,222	
Bibliography Request Processing	104.44	612,637	15.7	‡	216,250	12.6
7. Mail Receipt and Dispatch	0.37	2,192		0.03	6,108	
8. Interpret and Prepare for Search	20.94	122,839		0.20	43,807	
9. ADP Search	65.79	337,382		-	-	
10. Manual Search	-			0.40	87,365	
11. Prepare Comprehensive Bibliographies	-	56,883		1712.50	68,498	
12. Store, Maintain and Pull Catalog Cards	0.09	93,341		-	10,472	
Sub Total - Request Processing		2,210,316	56.5		722,161	42.0
Grand Total ††		3,908,897	100.0		1,722,519	100.0

‡ { 216,788 Reference Identification Requests
252 Bibliographies Prepared

†† Excludes certain additional operations performed by these agencies.

TABLE 7-2. ANNUAL WORKLOAD

	CENTER A	CENTER B
INPUT		
No. Reports Accessioned	30,613	27,894
OUTPUT		
Requests Processed	1,000,000	530,000
Copies Furnished	907,000	498,000
From Stock	483,000	414,000
Reproduced	424,000	84,000
Bibliographies	5,600	250
Document Reference Services	Incl. in Request Processing	215,000

7.1.1 Input Processing and Announcement

The total unit cost of input processing and announcement was $55.49 per title at Center A, and $36.00 at Center B. The higher acquisition cost at Center B is due primarily to the fact that they frequently buy a quantity of copies from the original source for pre-stocking purposes and this cost is included in acquisitions. The higher cost of surrogation in Center A is due in part to the deeper subject indexing required for a computer search, to the higher percentage of titles which must be abstracted, and to a generally higher cost structure within that center. The higher cost of document storage and pre-stocking within Center B is due to the fact that they contract out all of their printing, whereas Center A does its own printing at a considerably lower cost The input cost at Center A of the index data to the index store is lower than that at Center B, which does not utilize a computer. The relatively high unit cost of this function within Center A may be the result of a low average utilization of the computer equipment.

7.1.2 Output-Request Processing Costs

While both centers primarily provide copies of documents upon request, Center A also provides a rather extensive retrospective search service, whereas Center B provides only a modest retrospective search service since it can only manually perform this time-consuming function. It is therefore impossible to compare the costs of bibliography request processing since this function is not comparable within these two centers. The unit cost of handling requests for copies of documents appears to be lower in Center B than in Center A. however, this is due to the fact that in Center B a higher percentage of copies are furnished from inventory, the cost of which is included under input processing.

7.2 RELATIVE OVERALL COSTS OF VARIOUS SYSTEM FUNCTIONS

It should be noted that both of these information centers are relatively large in terms of the volume of output produced. It can be stated that they both service a national market. With this thought in mind, it is significant to observe that the so-called input processing costs which might be construed as fixed costs, are close to 50 percent of the total costs of these centers. This figure is slightly exaggerated, however, since the cost of pre-stocking has been included in both cases.

Table 7-3 presents a breakdown of the total costs for Center A and Center B in terms of personnel costs, computer rentals, other equipment rentals, postage, contracts (including printing), supplies and equipment, and General Administrative (G&A) expenses. It is interesting to note that direct personnel costs account for almost 50 percent of total cost. When the personnel cost portion of G&A is added, this figure is closer to 60 percent. Costs which were incurred for services purchased from outside sources for computer rentals, contracts, supplies, and equipment accounted for about 28 percent of the total. In Center B, this figure is closer to 35 percent, even without computer rental since it contracts out all of its printing.

7.3 FACTORS AFFECTING UNIT COST

In analyzing alternative system concepts or in comparing the effectiveness of existing systems, it is generally easier to compare unit operating costs rather than the total cost of a given function. There are several factors which are common to nearly all

TABLE 7-3. COMPARISON OF INFORMATION CENTER OPERATING COSTS BY TYPE OF COST

	CENTER A		CENTER B	
	$	%	$	%
Personnel	1,868,169	47.8	762,852	44.3
Computer Rentals	389,700	10.0	--	--
Reproduction Equipment Rentals	152,119	3.9	64,000	3.7
Postage	140,000	3.6	56,000	3.3
Contracts (incl. Printing)	164,150	4.2	489,000	28.4
Supplies and Equipment	373,093	9.5	43,000	2.5
Sub Total	3,087,231		1,414,852	
G&A	821,666	21.0	307,667	17.8
Total	3,908,897	100.0	1,722,519	100.0

unit cost computations. For example, fixed charges, such as rentals, depreciation, and maintenance, must be spread over the production load. Hence, the amount of the fixed charge, the productivity of each machine or workplace, and the production load including peak load variations enter into the unit cost computations.

The unit cost of information system functions is particularly sensitive to variation in the performance characteristics required for each of the system functions. For example, higher typographic quality of an announcement journal will increase the unit costs per title of the announcement function. Similarly, higher intellectual quality of indexing and abstracting, which is achieved by utilizing more highly trained people and by providing deeper indexing or informative abstracting, will raise the unit cost of the surrogation function.

The unit cost of performing a retrospective search is extremely sensitive to a number of interdependent factors. For example, the timeliness with which a response is required affects the mode of transmission, the file structure, and the ability to batch a number of questions. The size of the file affects unit cost either in terms of the amount

of the file which must be serially scanned or the amount of random access memory required. The complexity of the question affects the size of the batch, the amount of computation required, and the sophistication of the program. The processing speed and tape speed of the equipment will affect the optimum batch size and possible average question complexity.

7.4 ANALYSIS OF COSTS OF RETROSPECTIVE SEARCH

Generally, six elements of cost are considered in performing a retrospective search:

(1) Analyze the question.

(2) Structure the question.

(3) Encode the question.

(4) Search.

(5) Review and analyze.

(6) Deliver response.

The first four of these constitute the "look-up" aspect, the fifth (review) the "look-at" aspect and the sixth (deliver) the "take-away" aspect. Table 7-4 illustrates the relative costs of these functions for an average question based on data reported by the American Society for Metals (ASM) Documentation Service.[34] Table 7-5, which is also based on data from the American Society for Metals Documentation Service, illustrates the effect of batching and the number of hits on search cost per question. Obviously, the absolute figures will depend upon the effectiveness of the program and the performance characteristics of the machine itself. It is interesting to note that the overall cost figure of $129 for an average retrospective search using the ASM system as seen in Table 7-4, is not too far from the average of $104 per retrospective search for Center A as seen in Table 7-1.

7.4.1 Comparison of Several Retrieval Methods

Since the performance requirements of every information center is unique, and these requirements have a substantial effect on system design, it is difficult, if not impossible, to compare one operating information retrieval system to another because of these differences in the specific requirements of each system.

TABLE 7-4. COST OF A RETROSPECTIVE SEARCH — AVERAGE COMPLEXITY*

(1958-61 file; 5 simultaneous questions; 2000-card output or 400 answers per question).

1.	Analyzing	$ 3.130	
2.	Structuring	22.500	$25.880
3.	Automatic encoding	.250	
4.	Searching		
	a. Computer	38.670	
	b. Other machine operations	1.670	
	c. Expediting and recordkeeping	1.660	
5.	Review		$103.250
	a. 1st review	20.000	
	b. 2nd review	11.250	
6.	Transmitting answers		
	a. Photocopies	18.000	
	b. Assembling and mailing	12.000	
	TOTAL:	$129.130	

Cost per question of low complexity with same searching strategy — $105.310

Cost per question of high complexity with same searching strategy — $149.930

* Taken from a report by the American Society for Metals Documentation Service[34]

TABLE 7-5. EFFECT OF BATCHING AND NUMBER OF HITS ON COST OF RETROSPECTIVE SEARCHES*

	Number of Hits				
	None	1000	2000	4000	6000
Total minutes to search	240	265	290	340	390
Cost at $40/hour	$160.00	176.67	193.33	226.67	260.00
Number of questions in computer	Number of Hits (Cost per question)				
One	$160.00	176.67	193.33	226.67	260.00
Two	80.00	88.33	96.67	113.33	130.00
Five	32.00	35.33	38.67	45.33	52.00
Ten	16.00	17.67	19.33	22.67	26.00
Twenty	8.00	8.83	9.67	11.33	13.00
Fifty	3.20	3.53	3.87	4.53	5.20

* Taken from a report by the American Society for Metals Documentation Service[34]

In order to effectively compare the performance of various retrieval methods, it is first necessary to define a problem or problems which are to be handled by the various methods to be considered. The following discussion illustrates the type of trade-off analysis that can be made to compare various retrieval methods once a problem has been defined. The time and cost for handling a fact retrieval type of IS&R system by each of four types of retrieval methods, magnetic disc, magnetic tape, roll microfilm, and magnetic cards will be compared.

7.4.1.1 Sample Problem Definition. The sample problem to be analyzed involves a file with the following characteristics:

Daily input volume	=	17,000 records
Record size	=	120 characters
Retention period	=	12 months
File size	=	approx. 550,000,000 characters
Daily inquiries	=	1,750

The problem concerns an inactive file in that no maintenance or alteration is performed on a record once it enters the file. The only deletions occur as a result of a

TABLE 7-6. COMPARATIVE RETRIEVAL COSTS AND RESPONSE TIMES

Retrieval Method	Response Time	File Maintenance Time	Annual Cost		Total
			Equipment	Manpower	
Roll Microfilm (Indexed by computer)	15 Computer Minutes 29 Man-Hours	11 Computer Minutes	$65,000	$23,000	$88,000
Magnetic Tape	5 Computer Hours	20 Computer Minutes	$79,000	---	$79,000
Magnetic Discs	8 Computer Minutes	40 Computer Minutes	$317,000	---	$717,000
Magnetic Cards	34 Computer Minutes	3.5 Computer Hours	$84,000	----	$84,000

regular purge at the end of the retention period. The queries to the file do not entail Boolean logic since they simply ask for copies of records with a certain key. The input to the file has not been presorted according to the key. The input records are generated in a computer and are in machine-readable form.

7.4.1.2 Roll Microfilm. The problem can be solved using roll microfilm, but the machine-readable records must first be converted into microfilm. The conversion into microfilm can be accomplished by using a rented magnetic tape-to-microfilm recorder such as the General Dynamics/Electronics SC-4020. Another prerequisite is that a directory, as a link between the queries and the records, must be compiled by a computer. The reason that a manually prepared directory is not feasible is because the input to the file is not sorted by the key used for querying. A computer can maintain within itself a directory which points from a specific record key to the exact record address, stated in terms of the microfilm magazine and roll frame where the record is located. Queries would be processed in the computer and would produce a series of such addresses for further manual look-up.

The response time for 1,700 queries (allowing one minute for finding and reproducing the frame which contains the desired record) would be approximately 29 man-hours. Technically, the response time can be lowered to any desired level by performing the search operation in parallel. For instance, five reader printers would reduce the response time to six elapsed hours. Only about 15 minutes of computer time would be required to process the queries against the directory.

The computer file maintenance required to produce microfilm by means of the SC-4020 is only 30-odd seconds since the equipment is rated at 60,000 characters per second and the input is 17,000 records times 120 characters. Maintenance to update the directory would require about 11 minutes of computer time.

The computer cost is predicated on a computer which costs $60 per hour. The 26 minutes for maintenance and querying over 260 days would equal $6,760 annually. One hundred and fifty thousand frames per year at five cents per frame for the SC-4020 amounts to $7,500. Five reader-printers (with keyboard search) at $4,650 each is $23,250 which, if amortized over 42 months, is $6,640 annually. The total annual

equipment cost is $20,900. The supply cost for 1,750 response copies at 10 cents per copy over 260 days is $45,500, or more than 50 percent of the cost. This cost could probably be reduced by using a smaller sized sheet for the copy.

Manpower cost of 29 hours at $3 per hour over 260 days is $22,620. The total annual cost including equipment and manpower is $88,000.

A disadvantage of the microfilm approach for fact retrieval is that any significant manipulation of the data (e.g., summaries, averages, percentages, etc.) will undoubtedly require a conversion of the data back into machine-readable form.

7.4.1.3 Magnetic Tape. The similarity shoud be noted between the searching of magnetic tape and the searching of graphic information and indices on roll microfilm, such as the Rapid Selector or the FMA File Search.[3] In each case, the records in the file contain both the data of the record and the index values associated with the record. In each case, the entire file must be searched sequentially in order to compare the index values of each record with the values contained in the query. When a match occurs, the record is simply selected and written in a convenient form for the requestor after which the search continues. The queries can usually be batched, to one degree or another, in order to diminish the amount of searching time per query.

Batching queries for the sample problem with the file records in random order is not economically feasible. Too many comparisons within the computer would be required. The file is too long to be passed more than once, and a random file would require that each record in the file be compared with each of the 1,750 daily queries. To avoid execessive comparisons, the file and the queries must be ordered by the same key so that only a single set of comparisons need be made for each record in the file.

The response time would include an average of 10 minutes for computer sorting of queries. Assuming that there are 10 million characters on each tape reel, and that the query answers are distributed over all the reels, reading 55 reels (at five minutes per reel) would take approximately 4.5 hours to pass, or a total of almost five hours.

File maintenance for 17,000 records of 120 characters each would consist mainly of input sorting and set-up time, and would take approximately 30 minutes.

The primary cost of searching by magnetic tape is the tape passing time required by the serial search. Five hours of computer time at $60 per hour for 260 days is $78,000. The cost of tape reels and filing cabinets increases this figure to $79,000.

7.4.1.4 Magnetic Discs. Magnetic disc random access devices have been on the market for several years. Although extremely expensive and sometimes unreliable, discs have proven their worth for processing tasks which require a large number of random accesses, each in a very short period of time. For the sample problem, discs become economically unattractive because the disc would be used basically for long-term storage rather than for short-term processing, and the rapid accessing speeds of the disc would be worthless. Several methods for structuring the file are described in Appendix B.

Disc response time would drop from hours to minutes showing the dramatic benefits of rapid access from discs. Using an average of 100 milliseconds (access time for some discs is rated as low as 20 milliseconds) for each access and an average of 1.2 seeks for each record (using the Chaining Method described in Appendix B), the 1,750 queries can be processed in 8.5 minutes of computer time, allowing five minutes for set-up time.

File maintenance would be greater than query processing because an average of 1.2 accesses would be required for each of the 17,000 daily records for a total of 40 minutes with set-up time.

The annual cost is excessive because of the expensive nature of the discs. It would require three IBM 1302's to retain all the characters for 12 months. At $355,500 each, the total cost is $1,066,500. Amortized over 42 months, the annual cost is $305,000. The computer time of 48 minutes at $60 per hour for 260 days is $12,480, for a total of $317,480.

7.4.1.5 Magnetic Cards. Contrary to magnetic discs, the design emphasis of magnetic cards is less on speed of access and more on size of storage capacity. Until recently the NCR CRAM was the only magnetic card device on the market. Now there are two

others available for consideration: the RCA Model 3488 (formerly known as RACE) and the IBM Model 2321 Data Cell Drive. One major advantage of the magnetic cards over discs for the standard problem is that the cards are removable and can be replaced by other cards for other problems, whereas the discs are stationary and can be used only for the one problem. Consequently, the cost for the card drives can be shared with other problems. The methods for structuring the files would be the same as for discs and are described in Appendix B.

The response time for magnetic cards is much greater than for discs, but is well within reason on operating systems. An average of 500 milliseconds is the rated time for the IBM Model 2321 (the RCA Model 3488 is closer to 350 milliseconds). For 1,750 queries at two seeks each, the searching time would be close to 34 computer minutes, including set-up time.

File maintenance time would be substantially more because each of the 17,000 records would have to be placed in a chain on the random access cards. At 500 milliseconds for 1.5 seeks each, the time would be approximately 3.5 hours.

The annual cost of equipment includes four hours of computer time at $60 per hour for 260 days, or a total of $62,400. Using the RCA equipment for illustration, two Model 3488's would retain all the desired characters. The cost would be for one Control Unit ($32,500) and one Retrieval Unit ($135,000) and one Expansion Unit ($65,000) for a total of $232,500. At least two other comparable problems could be used to share the cost, lowering it to $77,500. Amortization over 42 months lowers the cost to an annual figure of $22,150. The total equipment cost would be $84,550.

7.4.1.6 Conclusions. Search system costs and characteristics vary too widely for any general conclusions to be drawn from a single sample problem, but several comments can be made:

(1) The roll microfilm solution at first appears attractive but involves implicit administrative and scheduling problems which are not apparent in Table 7-6.

(2) The serial search, which is an inherent part of magnetic tape systems, necessitates a delayed response time which can be intolerable in some search situations.

(3) The premium costs for magnetic discs exist because of the complexity of the equipment required to obtain their rapid response time. Discs are seldom justified for information retrieval systems unless a very large amount of file maintenance is required.

(4) Magnetic cards appear to offer the most flexible and economic approach for future large retrieval systems because the retention cost per character is low, because random access file structures can be utilized, and because the cards can be replaced by other cards in order to spread costs over the operation of several files.

7.5 COPY FULFILLMENT COSTS

In considering copy fulfillment costs, it is useful to distinguish between initial dissemination and demand requests for copies. Initial dissemination includes the first printing of a document by the publisher as well as the initial dissemination by a centralized documentation service, which might be considered to be a secondary distribution from the viewpoint of the publisher.

A number of comparative cost charts are presented in the subsequent paragraphs. The cost data included on these charts has been extracted from a variety of sources; in particular, from the cost accounting data obtained from Center A and Center B (see Tables 7-1 and 7-3), and from price lists and quotations obtained from commercial microfilming service companies in the Philadelphia area. A detailed analysis of individual equipment costs and productivity of individual subsystems has not been made. Consequently, the cost data presented in the following tables is not adequate for budgetary purposes, however, it should be useful for comparison of the relative cost of various methods.

7.5.1 Effect of Number of Pages Per Document

Table 7-7 presents a cost comparison between five different methods of initial dissemination, i.e., 16 mm. roll microfilm, 16 mm. microfilm in magazines, 3 x 5 microfiche, 3 x 5 microfilm jacket, and full size copy (offset). The problem assumed is the distribution of an average of 50 copies of 10,000 15-page documents. As would probably be expected, 16 mm. roll microfilm is the least expensive mode of dissemination and full size copy the most expensive. Microfiche and jackets are surprisingly close in cost.

TABLE 7-7. COST COMPARISON OF INITIAL DISSEMINATION METHODS

(10,000 — 15 page documents — 50 copies)

OPERATIONS	16 mm Roll	16 mm Magazine	3 x 5 Microfiche	3 x 5 Jacket	Full Size (offset)
MASTER COST					
Planetary Microfilming Process and inspection 10,000 x 15 x .015	2,250	2,250	2,250	2,250	2,250
Roll-to-Roll Duplicate 10,000 x 15 x .002	300	300	300	300	300
Title and Strip-Up (Fiche) (50¢)			5,000		
Copyflo to Multilith 6¢/page					9,000
Total Cost of Master	$ 2,550	$ 2,550	$ 7,550	$ 2,550	$11,550
RELEASE PRINTS/SET					
Roll-to-Roll Duplicate	300	300		300	
Multilith, collate, and staple at 60¢/page $\frac{10,000 \times 15 \times .54}{50}$					1,620
Magazines and loading cost $\frac{150,000 \times 1.50}{2000}$		112			
Fiche-to-Fiche Duplicate 10,000 at 10¢ (3 x 5)			1,000		
Jackets — 4¢				900	
Inserting 4¢ (2 sleeves)					
Label $\frac{1¢}{9¢}$ x 10,000					
Total Cost per Release Print	300	412	1,000	1,200	1,620
Cost for 50 Prints	$15,000	$20,600	$50,000	$60,000	$81,000
Total Cost of Master + 50 Prints	$17,550	$23,150	$57,550	$62,550	$92,550

With an initial distribution volume of 50 copies per title, one would expect that microfiche should be considerably less expensive per copy than the microfilm jacket method. The reason for this is that microfiche becomes more efficient with a greater average number of pages per microfiche card.

Table 7-8 presents a cost comparison of the same five methods for 10,000 50-page documents with a run of 50 copies per document. The microfiche becomes an attractive contender in this situation because 50 pages can easily be packed onto one microfiche with room to spare. The jacket cost is high because of the labor of inserting film into five sleeves plus the cost of the 16 mm. film duplicate, jackets, and labels.

7.5.2 Effect of Number of Copies Disseminated on Total Cost

To examine the effect of number of copies on the total cost of initial dissemination by various methods, the graph shown in Figure 7-1 was prepared. This graph is based on the data presented in Table 7-8 for the 50 page document. The breakeven point between jackets and microfiche for a 50 page document is somewhere around three or four copies. On the same basis, the breakeven point between magazines and microfiche is somewhere around 21 or 22 copies.

7.5.3 Unitized Microforms

The aperture card has gained a considerable foothold in the unitized microform field, particularly in the field of engineering drawings where the document is generally one page. The aperture card is by far the most expensive type of unitized microfilm both from the point of view of creating the master as well as the creating of duplicate files. For this reason, microfilm jackets have generally been utilized wherever a unitized record of a multiple page document was desired. The jacket afforded the advantage of relatively easy updating. The disadvantage of the jacket, however, is that it is not easy or inexpensive to make duplicate sets of jacket files. Engineering difficulties have prevented the production of a jacket-to-jacket (or microfiche) duplicator. Instead, multiple files of jackets are produced by making duplicate 16 mm. rolls and inserting them into blank jackets. The advent of the microfiche provided a good alternative to the jacket, especially in those instances where there was an initial distribution of a number of copies. Up until the present, the microfiche has been more expensive than the jacket for a one copy system.

TABLE 7-8. COST COMPARISON OF INITIAL DISSEMINATION METHODS

(10,000 — 50 page documents — 50 copies)

OPERATIONS	16 mm Roll	16 mm Magazine	4 x 6 Microfiche	4 x 6 Jacket	Full Size (offset)
MASTER COST					
Planetary Microfilming Process and inspection 10,000 x 50 x .015	7,500	7,500	7,500	7,500	7,500
Roll-to-Roll Duplicate $5 \times 10^5 \times .002$	1,000	1,000	1,000	1,000	1,000
Title and Strip-Up (Fiche) (50¢)			5,000		
Copyflo to Multilith 6¢/page					30,000
Total Cost of Master	$ 8,500	$ 8,500	$13,500	$ 8,500	$38,500
RELEASE PRINTS/SET					
Roll-to-Roll Duplicate	1,000	1,000		1,000	
Magazines and loading cost $\frac{500,000 \times 1.50}{2000}$		375			
Fiche-to-Fiche Duplicate 10,000 at 12¢ (4 x 6)			1,200		
Jackets 5 sleeves x 2 = 10¢ Jacket = 5¢ Label = 1¢ 16¢ x 10,000				1,600	
Multilith, collate, and staple $\frac{500,000 \times .54}{50}$					5,400
Total Cost Per Release Print	1,000	1,375	1,200	2,600	5,400
Cost for 50 Prints	50,000	68,750	60,000	130,000	270,000
Total Master + 50 Prints	$58,500	$77,250	$73,500	$138,500	$308,500

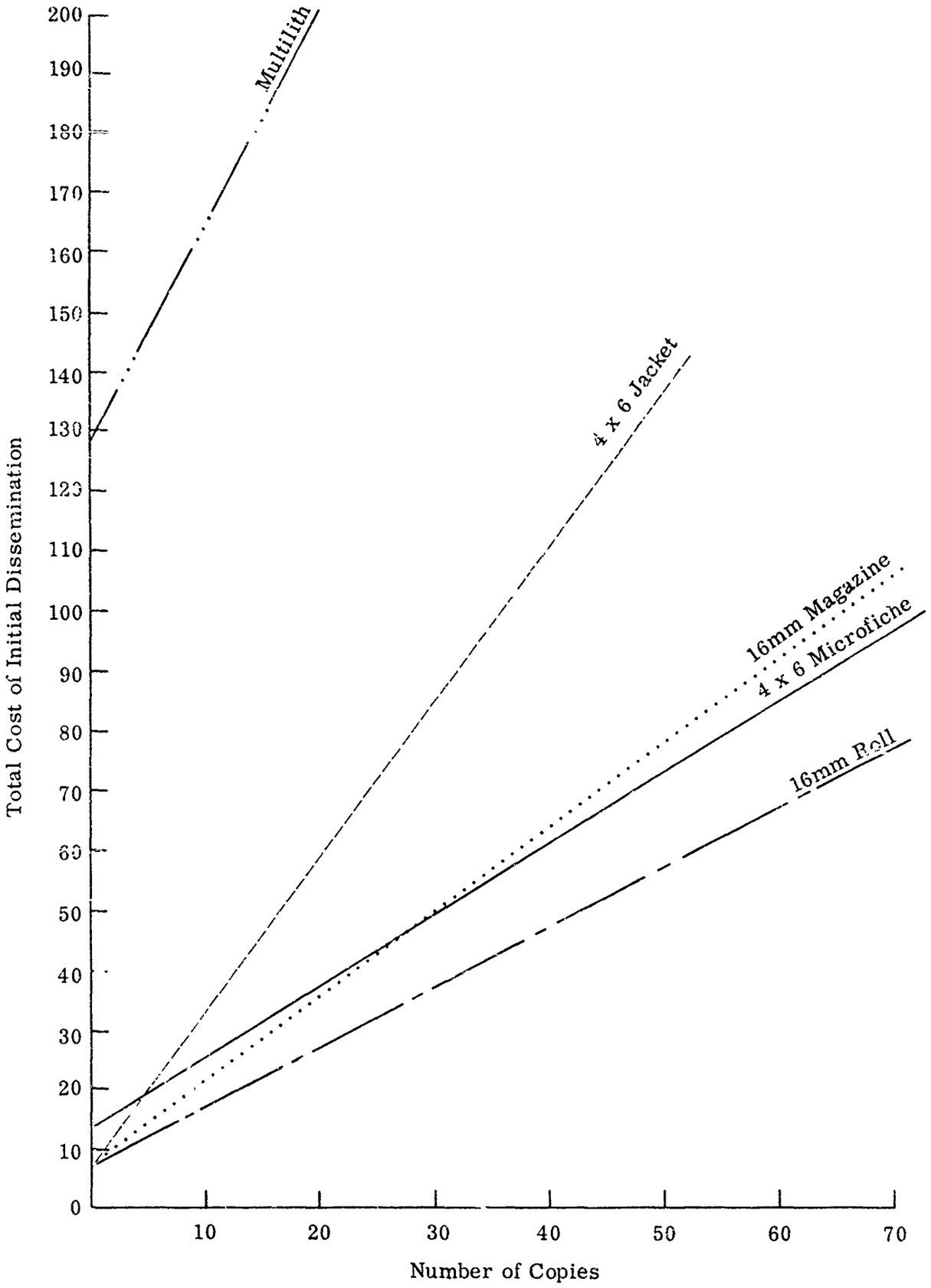


Figure 7-1. Effect of Number of Copies on Total Cost of Initial Dissemination for 10,000 — 50-page documents

This is due to the problems of stripping up a master microfiche and of making an eye-legible title. Where these costs could be spread over a number of copies, the microfiche became less expensive than the jacket because of its ease of reproduction. There are new developments underway in many quarters for reducing the cost of preparing a master microfiche. Such developments include step and repeat cameras and pressure sensitive adhesives applied directly to the 16 mm. film for stripping up the 16 mm. microfilm onto an acetate sheet. This latter approach affords the additional advantage of easy updating. It is expected that the microfiche will gain in popularity and will eventually be competitive with the jacket even for a single copy installation. It is likely that the microfiche will also begin to have application in engineering drawing and data applications, wherein the drawings and bill of materials might be combined onto a single microfiche thereby making publication of sets of copies much less expensive than individual aperture cards. For example, the drawings for an entire subsystem could be combined on a single microfiche. This microfiche could then be duplicated for about 12 cents, which is little more than the cost of duplicating one aperture card.

7.5.4 Request Copy Fulfillment

There are a number of methods for satisfying individual requests for documents. Prior to the advent of Xerography, the only practical method was to maintain an inventory of printed copies of all documents in the collection; the inventory was replenished by reprinting as required. To avoid the problems associated with maintaining inventories of rarely requested items, some organizations have gone to the other extreme, wherein all requests for copies are serviced by making a replica copy (usually by Xerography) from a microform master, and no inventory of printed copies is maintained. Today, most organizations, which have a copy fulfillment problem, operate somewhere in between these two extremes, i.e., they maintain a stock of the more popular items and make individual replica copies by Xerography of the less popular items.

Whether the policy is pre-stock or on-demand copying, there are a host of system variations that affect the cost of production. For example, some document centers provide reduced size copies, two pages up on an 8-1/2 x 11 sheet (70 percent blowback) or four pages up (50 percent blowback). These methods can provide a savings in printing or copying costs of from 50 to 75 percent. Other centers prefer to furnish the user with a

reproducible microform copy such as an aperture card or microfiche. If the user then wishes to make hard copy, he can make it from the microform.

7.5.4.1 Pre-Stocking By an Overrun on Initial Dissemination. The cost of pre-stocking depends on the method and quality of printing. In addition, the cost depends on whether the stock is obtained as an overrun from a printing for other purposes, e.g., initial dissemination, since this would save additional set-up costs such as mounting and unmounting of paper plates. The quality of offset reproduction from paper plates is usually higher than the quality of a Xerox copy. The quality is a function of the plate making process employed. Since the cost of on-demand copying to pre-stock printing is compared, it is assumed that the quality of pre-stock copies need not be higher than that of the on-demand copy. Consequently, the printing method assumes the making of offset paper plates from a microfilm image by means of the Xerox Copyflo, and then offset duplication on 50 lb. offset paper.

Figure 7-2 illustrates the costs for supplying single copies on demand of 50-page documents as compared with the cost of pre-stocking various quantities of copies. The solid line represents the cost of preparing on-demand copies of 50-page documents. The unit cost figure used in preparing this figure was 3.5 cents per page for on-demand copying, based on the experience of Center A. This figure includes finding the strip of 35 mm. microfilm, splicing, Xerox Copyflo rental, paper, labor, cutting and stapling, unsplicing and refiling the 35 mm. film. The printing costs, also based on the experience of Center A, amount to 60 cents per page for 50 copies, including preparation of paper plates, mounting and unmounting of paper plates, supplies, labor, press time, collating, and stapling. It was assumed that one-third of this cost is a fixed charge which does not vary with run size and that two-thirds is variable with run size. Consequently, for an overrun, a cost of only 40 cents per page is incurred for obtaining the 50 copies for pre-stocking since no additional fixed charges are incurred because of the overrun. A lower figure of $2.50 per page per thousand copies is utilized for quantities above 50 copies, this figure is based on commercial printing estimates.

It is also necessary to consider the cost of financing the inventory as well as storing the unsold stock or inventory of copies. Obviously, if the minimum number of copies required to "break even" is not sold within a reasonable number of years, the storage

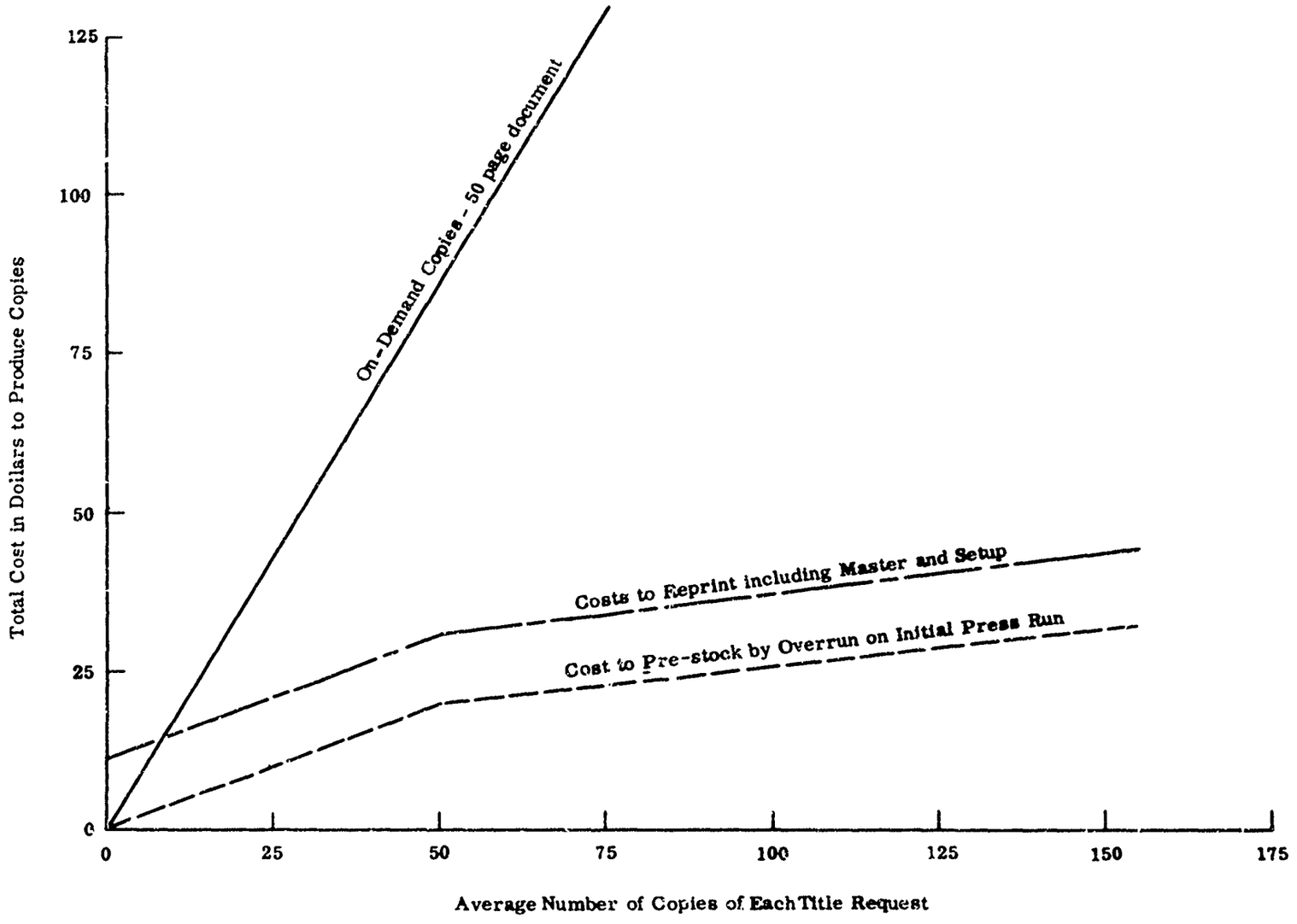


Figure 7-2. On-Demand Copying Costs Vs. Pre-stocking For Various Numbers of Copies

and financing costs will offset the advantage of pre-stocking. Consider the following hypothetical example for determining storage costs.

(1) Annual cost of storage per copy is $.03.

(2) Twenty-five percent of the total stock is sold within an average of one year from receipt of initial stock.

(3) All unsold stock is destroyed routinely after five years.

The storage cost (Sc) per document request can be computed as follows:

$$S_c = \frac{3/4\ P_a\ (5\ \text{yrs.} \times \frac{0.03}{\text{yrs.}}) + 1/4\ P_a\ (1\ \text{yr.} \times \frac{0.03}{\text{yrs.}})}{R_a}$$

$$S_c = \frac{3\ R_a\ (.15) + R_a\ (.03)}{R_a}$$

$$S_c = 0.48/\text{request.}$$

where P_a = Volume of copies pre-stocked annually.

R_a = Volume of requests annually.

While the above figures are hypothetical, they illustrate the need for considering the costs of storage and financing. It is obviously necessary that some time limit must be established before routine destruction of excess inventory, otherwise storage and financing costs will become larger each year as the inventory grows.

7.5.4.2 Reprinting Policy as A Function of Document Age or of Demand History. The question of whether or not to reprint a particular document is analogous to whether or not it should be pre-stocked in the first place, and in what quantity.

The reprinting question is somewhat easier than pre-stocking since demand history is at least hypothetically available. In the case of large files, where no detailed inventory records are maintained, reprint policies are often based on the age of the document. For example, the Patent Office, until a few years ago, would reprint 50 copies of any of the 3,000,000 U. S. patents which was out of stock upon the receipt of an order for a single copy. It now follows the policy of reprinting all items above Patent Number 2,500,000 and only those which have a proven record of high demand below 2,500,000. Other organizations maintain a firm no reprint policy. The rationale behind this viewpoint is that the demand for a document diminishes rapidly with its age. A number of studies have confirmed this viewpoint for a variety of document collections.[25]

Where the demand history for an individual document is obtainable, it is possible to make predictions of future demand and hence make better decisions on the question of reprinting Frequently, the demand history is only maintained for some short period after announcement since this is generally the period of greatest demand. In those document centers which begin with about 10 copies of the document obtained from the source, they are able to obtain some demand data before they must decide whether to fulfill future requests from inventory or by on-demand copying.

7.5.5 Lower Cost Printing

A number of developments are underway which are likely to have a significant effect on the cost of short-run printing and thereby accelerate the trend toward pre-stocking. The factor which made printing comparatively expensive for a short run of 10 or 15 copies was the set-up costs, e.g., mounting and unmounting printing plates. This relatively high set-up cost was due to the necessity of manually performing these functions. A highly mechanized offset duplicator has been introduced by Addressograph-Multigraph Corporation which is known as the A/M 2575 Tandem Duplicator. This machine, which costs approximately $10,000 is capable of printing on both sides of the page in tandem. According to the manufacturer's claims, one operator can eject the old masters, mount two new masters, run clean-up copies, and run 60 clean copies in 52 seconds. There are presently only a few A/M 2575 Duplicators in use. It is expected, however, that their use will become more widespread and that the cost of short run printing will drop to about 25 cents per page per 50 copies, if not less.

As pointed out earlier in this Section, it is possible to put reduced size images on a paper plate and then print these on a duplicator such as the A/M 2575. At two pages up, both front and back, this machine can easily deliver 60 clean copies of four pages every 52 seconds or approximately 13 seconds per page per 60 copies.

7.5.6 Lower Costs for On-Demand Copying

The concept of supplying reduced sized copies was first applied to on-demand copying in order to reduce per page costs. This was particularly important in the case of copies prepared on silver halide photographic paper in order to reduce the high paper

cost per copy. As pointed out above, however, these advantages can be offset by similar savings obtainable by providing reduced size printed copies applying the same principle.

Other techniques for reducing the cost of on-demand copies have been to purchase the copying equipment, e.g., Xerox Copyflo and obtain maximum utilization by a three-shift operation. The same principle has been applied under rental contracts. However, when the maximum rental applies, it is nearly three times the minimum rental. Another cost saving technique which has been applied to reducing the cost of on-demand copying via the Xerox Copyflo has been to utilize the largest size web of paper obtainable and place the documents with their longest side across the width of the web, thereby obtaining more copies per lineal foot of processing.

SECTION VIII. HARDWARE CONSIDERATIONS

8.1 IS&R SYSTEM FUNCTIONS AND ASSOCIATED EQUIPMENT

As described in Section V there are eight basic functions of all IS&R systems from which all such systems can be assembled, i.e., origination, acquisition, surrogation, announcement, index operation, document management, correlation, and end-use. This section describes the hardware implications of these system functions. It is not meant to be a catalog or evaluation of specific hardware, but rather a description of the functional requirements and the general type of hardware which has been or can be used to meet these functional requirements.

Table 8-1 summarizes the type of equipment which is utilized in the five typical applications described in Section VI for each system function as well as other types of equipment which may possibly be utilized for these functions.

One thing is reasonably clear. There is no piece of equipment or even complex of hardware and software which can perform all of the major functions involved in a total IS&R system. Rather, an IS&R system requires a collection of seemingly unrelated pieces of equipment tied together by a well-documented set of manual systems and procedures, plus computer programs where general-purpose computers are used.

Document retrieval systems generally involve some form of index storage and retrieval device which may or may not be a computer, plus a wide variety of document replication equipment. Fact retrieval systems are likely to require extensive digital processing capability, random access storage, inquiry consoles, communications equipment, and, possibly, even some facsimile equipment. It is doubtful that equipment for searching combined graphic and digital files will ever be highly effective for document or fact retrieval for the reasons described in Paragraph 8.3.

8.1.1 Origination

The origination function involves the initial publication of a report, journal article or other recorded form of information. The types of equipment normally

TABLE 8-1. HARDWARE IMPLICATION

FUNCTION	NASA	GE-MSD	TRC (in planning
Origination			
Acquisition	Manual	Manual	Manual
Surrogation	Manual	Manual	Army chemica Computer Light pen Auto. characte
Announcement	LCC-S Justowriters Photon photocomposer IBM 1410-1401 computer IBM 1403 printer Process camera Platemaking equipment Offset presses Make-up tables Diazo copier (proofs)	Typewriter Offset press	
Index Operations	Key punches LCC-S Justowriters IBM 1401-1410 Systematics tape-to-card converter	Keypunches GE-225 Teletype console	Computer Random acces Army chemica Automatic cha reader Data link Switching equi Display device Query console
Document Management	Planetary microfilm Cameras Unipro processor Strip-up tables Densitometer & microscope FME-roll-to-roll duplicator Ozalid super ozamatics Arc-vac platemaking machines Photostat 1014 Xerox 914 Filmac 200 R - (Inspection)	Xerox 914 Filmac 100 Filmac 200 Microfilm Service Bureau Jacket readers EAM equipment	Computer Random acces Data link Switching equi Display device
Correlation			
End-Use	Reader-printer	Manual	Display device

A

TABLE 8-1. HARDWARE IMPLICATIONS OF VARIOUS IS&R APPLICATIONS BY SYSTEM F...ONS OF

NASA	GE-MSD	TRC (in planning stage)	ENG. DATA CENTER	TITLE SEARCH SYSTEMS	C ...ing stage
...ual	Manual	Manual	Manual	Microfilm planetary ca...	...ical typew...
...ual	Manual	Army chemical typewriter Computer Light pen Auto. character readers	Manual	Manual	...cter read...
...S Justowriters ...n photocomposer ...1410–1401 computer ...1403 printer ...ss camera ...making equipment ...presses ...-up tables ...copier (proofs)	Typewriter Offset press				
...unches ...S Justowriters ...1401–1410 ...matics tape-to-card ...erter	Keypunches G... 225 Teletype console	Computer Random access storage Army chemical typewriter Automatic character reader Data link Switching equipment Display devices Query console	EAM equipment Manual card files	Typewriters EAM equipment Computer IBM 826	...cess stora... ...ical typew... ...character... ...quipment... ...rices ...ole
...tary microfilm ...eras ...o processor ...up tables ...ometer & microscope ...roll-to-roll duplica- ...super ozamatics ...ac platemaking ...ines ...tat 1014 ...914 ...: 200 R - (Inspection)	Xerox 914 Filmac 100 Filmac 200 Microfilm Service Bureau Jacket readers EAM equipment	Computer Random access storage Data link Switching equipment Display devices	Planetary microfilm cameras Aperture card mounters Roll-to-roll duplicator Card-to-card duplicator Film processors EAM equipment Reader-printers	Xerox Copyflo Offset duplicator Jacket reader Microfilm duplicator Xerox platemaking camer...	...cess stora... ...quipment... ...rices
...-printer	Manual	Display devices	Reader-printers	Jacket reader Reader-printer	...ices

B

OF VARIOUS IS&R APPLICATIONS BY SYSTEM FUNCTION

stage)	ENG. DATA CENTER	TITLE SEARCHING SYSTEMS	OTHER POSSIBLE HARDWARE
			Keyboard devices Photocomposers Computers Platemaking equipment Printing equipment
	Manual	Microfilm planetary cameras	
typewriter readers	Manual	Manual	Light pen Character readers Computers
			Electronic composers (ZIP) (LINOTRON) Sequential card cameras
storage typewriter cter ient	EAM equipment Manual card files	Typewriters EAM equipment Computer IBM 826	Random access storage Termatrex Microcite Microfilm selection devices Facsimile Displays
storage ient	Planetary microfilm cameras Aperture card mounters Roll-to-roll duplicator Card-to-card duplicator Film processors EAM equipment Reader-printers	Xerox Copyflo Offset duplicator Jacket reader Microfilm duplicator Xerox platemaking camera	Automatic offset duplicator Roll-to-fiche printers Facsimile
			Computers Random access storage Microfilm selection devices Query console Displays Fascimile
	Reader-printers	Jacket reader Reader-printer	Copiers Computer printers CRT Displays

C

involved in publication are Linotype and Monotype keyboards and casters, proof presses, platemaking equipment, and printing presses. The high cost of typesetting combined with new developments in electronic composition techniques is likely to result in the establishment of cooperative publication facilities utilizing high-speed, automatic equipment. Such a facility would publish hundreds of journals for a number of publishers. The equipment utilized would include tape typewriters, computers, electronic photocomposers and possibly optical page readers for transforming typewriter draft copy into machine language. Figure 8-1 illustrates a hypothetical computer-based publishing production facility. Figure 8-2 illustrates a hypothetical non-computer based publishing facility.

8.1.2 Acquisition

The acquisition function includes the acquiring of documents (either by purchase or exchange), evaluation, selection, duplicate checking, and accessions. These functions involve mostly intellectual and some clerical effort. Because of the high level of intellectual effort involved, hardware is not generally utilized.

The purchase function, however, can be aided by standard punched card equipment. In addition, a device known as Photoclerk, developed by Dr. Ralph Shaw, when he was director of the Department of Agriculture Library*, has been used to aide the acquisition function. Photoclerk uses a photographic process to capture a strip of information required to place the order, and by the use of a mask adds certain standard information such as the name and address of the purchaser, purchase order number, and the like.

The function of checking for duplicates, while usually performed by reference to a card catalog, can be aided by a simple computer search or look-up using the descriptive cataloging information. The disadvantage of using the computer for verification is that it requires processing the suspect duplicate through at least part of the cataloging function and through translation to machine readable media via keyboard entry (either on or off-line).

---

* Now known as the National Agricultural Library.

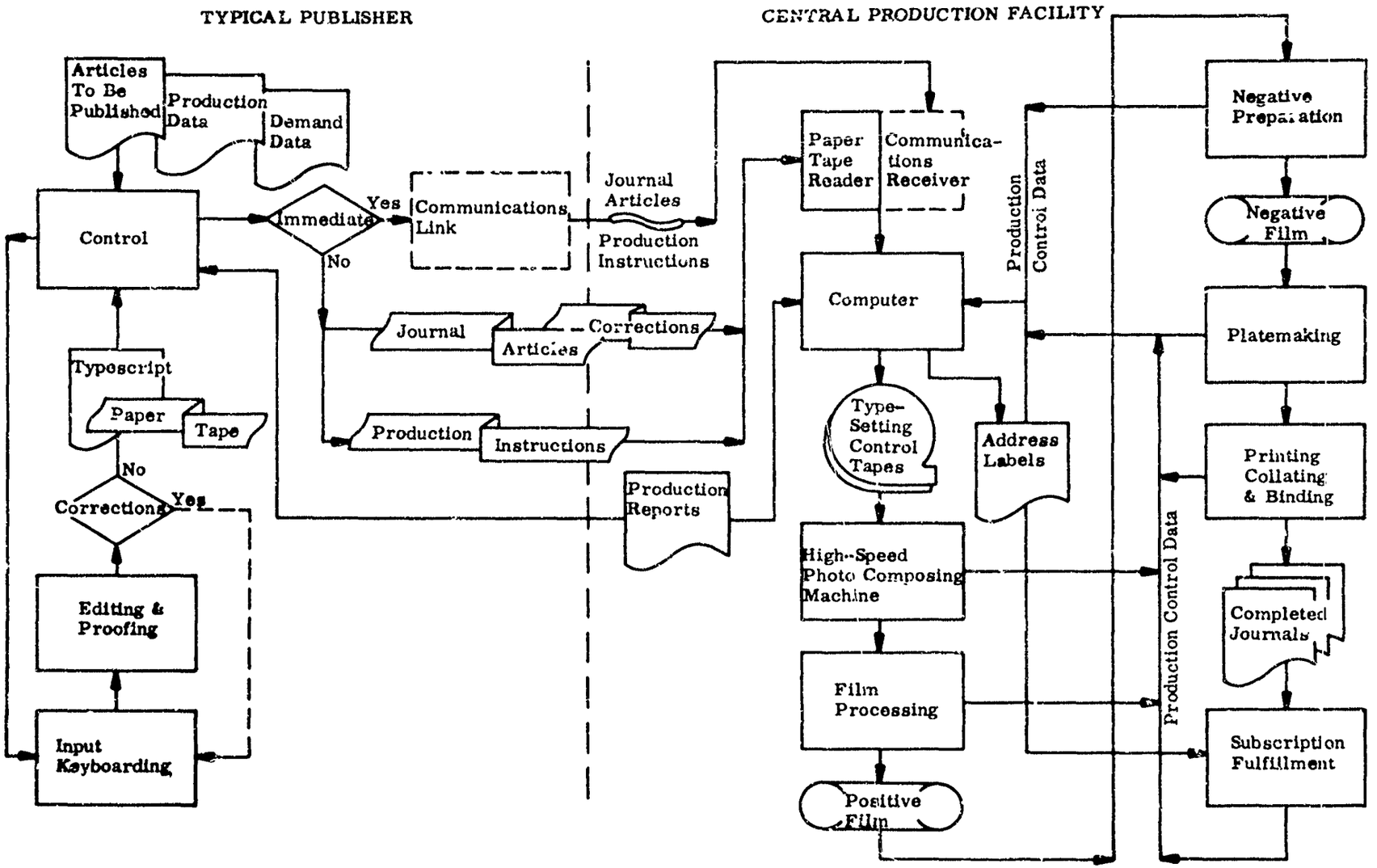


Figure 8-1. Hypothesized Computer Based Publishing Production Facility

8.1.3 Surrogation

The functions of cataloging, abstracting, and indexing are generally considered to be intellectual functions and do not directly involve any hardware. Much research is being done, however, to mechanize these functions, e.g., automatic indexing and abstracting, which is discussed in Paragraphs 3.3, 9.2, and 9.6.

The major element of hardware required to perform automatic indexing and abstracting is a general-purpose computer. A primary difficulty in automatic indexing and abstracting has been storing the full text into the computer. Consequently, if automatic indexing and/or abstracting is to be economically feasible, it will be necessary to capture the text as a by-product of publication or by effective optical page readers. These machines will have to be capable of handling a wide variety of type fonts, type sizes, and page sizes in order to accept the wide variety of input forms received by most document information centers. A number of companies are working on the development of such machines.

In those cases where the full text is prepared for machine input on a tape typewriter or received on tape as a by-product of the type-setting function, there may also be a requirement for code converters and for high-speed, paper tape readers.

8.1.4 Announcement

The announcement function helps to serve current-awareness needs by announcing newly obtained documents through the medium of announcement journals and book-form indexes. These publications involve a number of special system requirements which significantly affect the choice of hardware. These requirements frequently include the following:

(1) Short throughput time (two weeks).
(2) Unit records ranging from 175 to 1,000 characters each.
(3) Sorting and merging.
(4) Updating.
(5) Repeated entry of text in multiple locations.

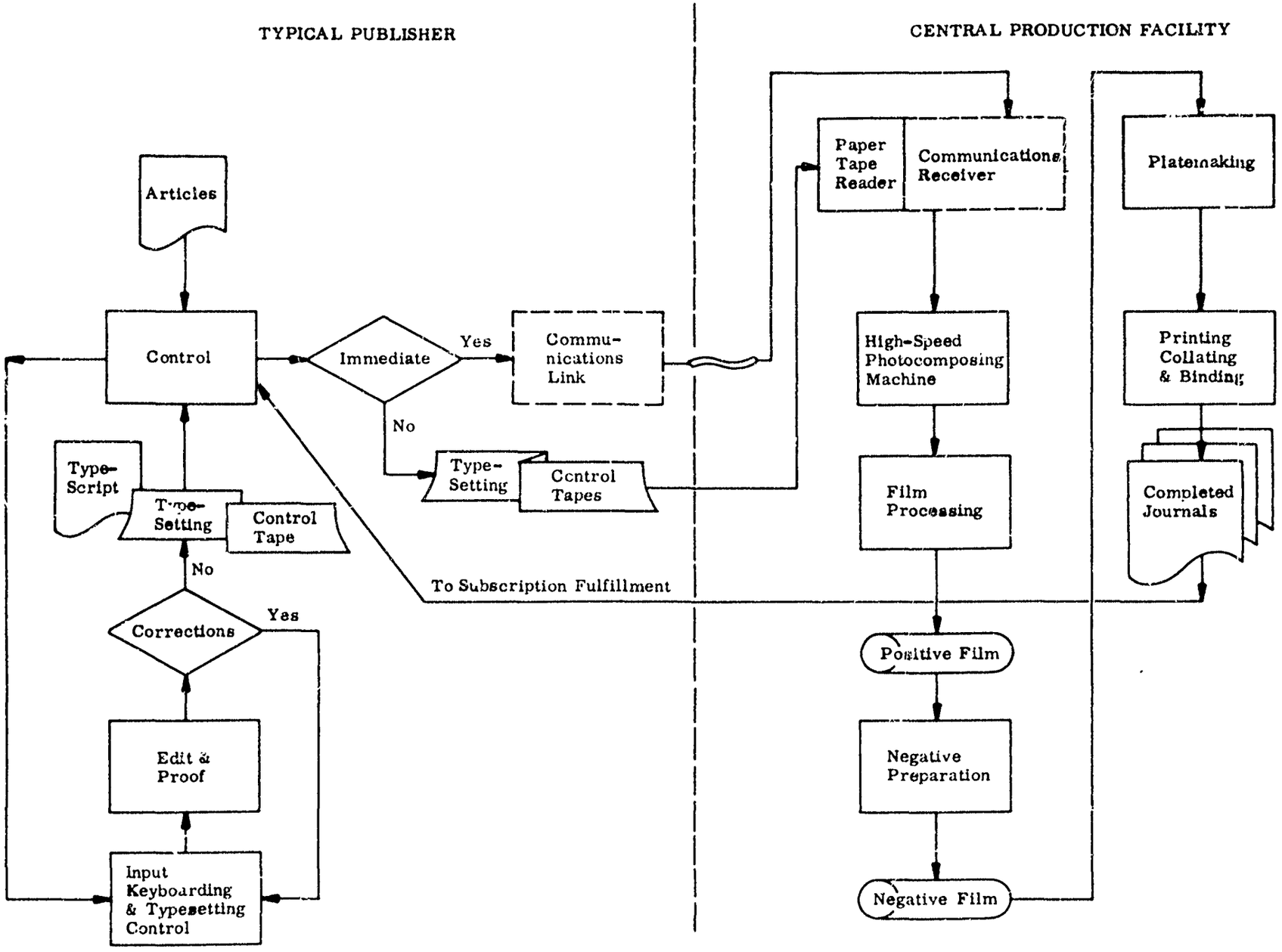


Figure 8-2. Hypothesized Non-Computer Based Publishing Production Facility

(6) Cumulation (merging of weekly editions into monthly cumulations, etc.).

(7) Reasonable typographic quality.

The time limitations on the publication of an announcement journal generally necessitate that the keyboard data transcription function be done in parallel with the surrogation function. This in turn necessitates the processing of unit records and the arrangement of these records into publication format just prior to publication. There are essentially three basic methods by which this can be done. The simplest method is to manually arrange and paste-up camera ready copy. This, however, is generally unsatisfactory where there is a requirement for cumulative indexes such as would necessitate either recomposing or tearing down and rearranging the unit records. The second method is known as sequential card composition. Briefly described, the method involves typing one, two, or three lines of copy in a designated position on a tabulating card, which may be punched for automating sorting and merging operations. At the time of publication, the cards are counted and separated into column subdecks; each column subdeck is then run through a sequential card camera such as a Listomatic, Fotolist or Composoline camera Figure 8-3 illustrates the cards with a single line of typing on each and the resulting sequential card camera negative output. The breaking down of the text into single or multiple lines on individual cards, which can be machine sorted, merged and photographed greatly simplifies the updating and cumulation problem. An example of an announcement journal prepared by this technique is Nuclear Science Abstracts, which is produced by the Atomic Energy Commission-Technical Information Division using IBM typewriters, Varitypers, and a Listomatic sequential card camera.

The third technique involves computer assisted typesetting. The computer is highly suitable for handling the problems of sorting, merging, multiple entry, updating, and cumulation. A major problem associated with computer typesetting, however, has been the unavailability of suitable high-speed output printers of sufficient typographic quality. Many applications have used mechanical computer printers of low typographic quality. The first application to seriously attack this problem was the National Library of Medicine's MEDLARS system for producing Index Medicus.[14] This system which is similar to the generalized system illustrated in Figure 8-1 involves (1) the input keyboard

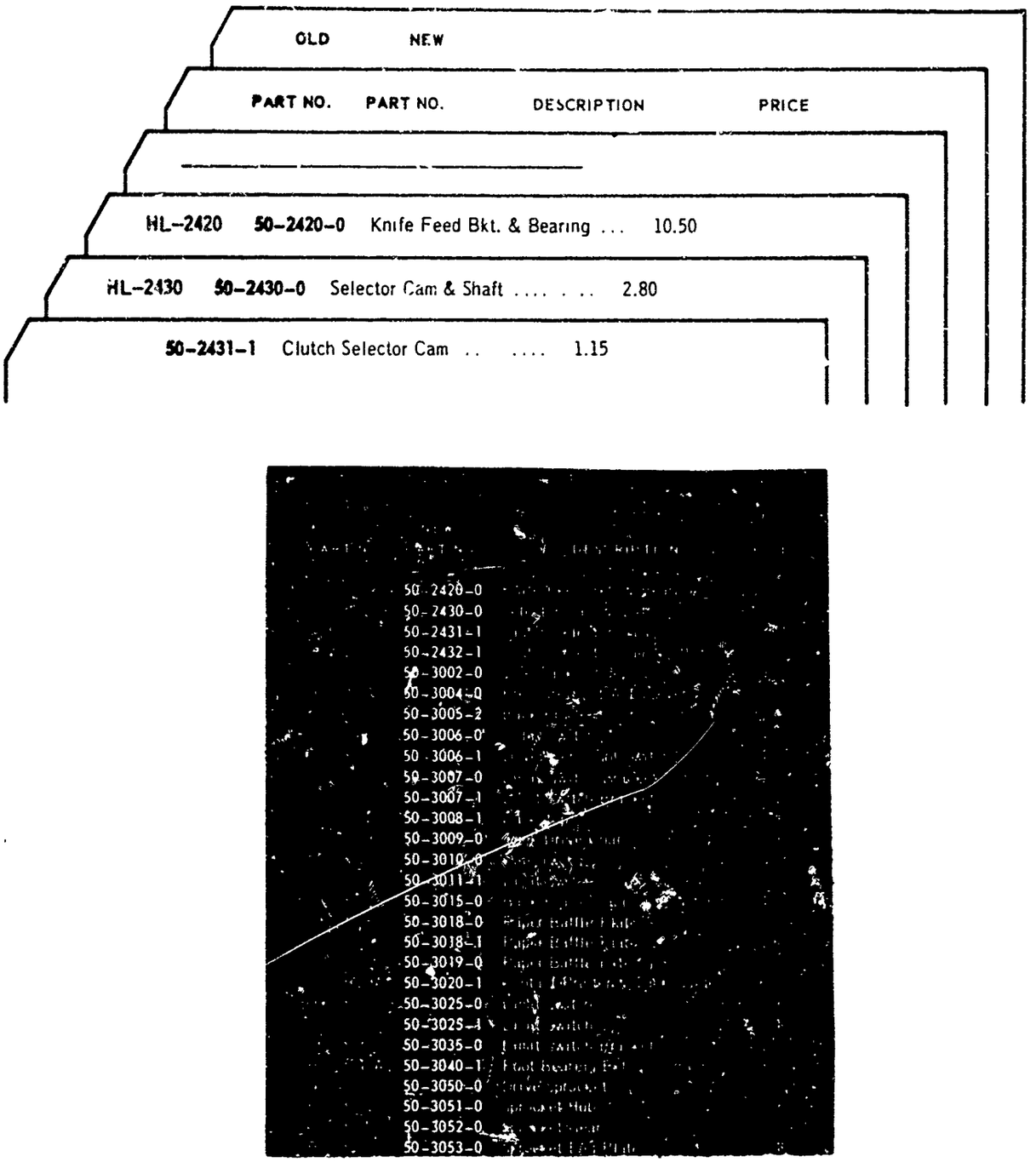


Figure 8-3. Sequential Card Composition

data transcription of citations on a Flexowriter, (2) the manipulation and regeneration of text for multiple entry on a Honeywell 800 computer, and (3) a photocomposition procedure using a high-speed graphic arts quality composer known as GRACE. GRACE was developed for the MEDLARS application by the Photon Corporation under a contract with the General Electric Company information systems operation, prime contractor for the MEDLARS system. GRACE can produce up to approximately 440 characters per second with a character repertoire of 226 alphanumeric symbols.

Where the information system also involves computerized index storage and retrieval, it is generally desirable to capture the input to this function as a by-product of the announcement function. This can be accomplished within any of the three techniques described above by using a paper tape producing keyboard, e.g., a Flexowriter, Justowriter, Dura Mach 10, etc. Whether the composing for the announcement journal should be done before or after input of this information to the computer depends upon the specific requirements of the particular announcement journal. As described above, Index Medicus is produced on the output side of the computer. This is desirable because of the multiple entry requirement (average of three entries per item), the large volume of entries processed per year (150,000), and the small average size of each item (175 characters). The NASA Announcement Journal, Scientific and Technical Aerospace Review (STAR), is produced partially on the input side of the computer and partially on the output side. The abstract section involves unit records of approximately 1,000 characters each, which are entered once only and are not cumulated. Consequently, the manipulation function is rather minor and the abstract tape goes directly from the keyboard to the Photon. The indexes to STAR, however, involve multiple entry, updating, cumulation, and considerable sorting and merging. Consequently, these are processed by the computer, and presently printed on an IBM 1403 mechanical computer printer. It is expected that NASA will eventually produce these indexes on a graphic arts quality device, such as GRACE.

8.1.5 Index Operation

Index operation normally involves the functions of storing and searching index data. As demonstrated by this report, this is only one of the many functions of typical information systems. It is neither the most difficult nor the most expensive of the various

functions. The intellectual functions associated with surrogation are usually the most expensive and troublesome. Index operation may utilize a wide variety of equipment and techniques including card catalog files, edge-notched and interior-notched cards, special-purpose mechanical and electronic searching devices, and general-purpose computers.

8.1.5.1 Card Catalog Files. Although catalog cards are typeset and printed by the Library of Congress for wide dissemination, the average local library has, until recently, generally utilized only a typewriter for this function. Recently, however, there have been a number of innovations applied to the problem of producing multiple copies of a catalog card so that a card can be filed under each of various headings such as subject and cross reference headings, author, title, source, report number, and the like. These innovations include the use of such devices as Programmatic Flexowriters, computers and various duplicating techniques for making multiple copies such as spirit hectography (Ditto), offset duplication, and Xerography. A special-purpose device, known as KROSFILER, was developed by Itek Corporation for the Air Force Cambridge Research Laboratories' library.[40] [18] This device accepts a machine interpretable encoding of the descriptive cataloging and produces an exploded tape with the information rearranged. This exploded tape is fed to a Flexowriter which prepares a set of reformatted cards so that the tracing will be in the appropriate place on each card of the set. This result can also be accomplished by a general-purpose computer.

8.1.5.2 Edge-Notched and Interior-Notched Cards. A variety of edge-notched cards and interior-notched cards is frequently used for retrospective searching (see Paragraphs 4.2.2.1 and 4.2.2.2). The hardware involved in these systems is usually special-purpose punching and drilling equipment for placing the appropriate holes or notches in the cards such as required in the Termatrex system manufactured by Jonkers Business Machines Company.

In addition to the punching and drilling devices, there are a variety of devices to assist the search process including needles, light tables with cursors for reading coordinates, and random card filing equipment such as that produced by Acme Visible Corp., Randomatic Systems Inc., and Mosler Safe Corporation.

8.1.5.3 Miscellaneous Devices. A number of electromechanical and electronic devices, other than computers, have been developed for searching index files. Examples of these are the IBM 101 Statistical Sorting Machine, the IBM 9900 Special Index Analyzer (also known as COMAC), the General Electric Tape Comparator and the Heatwole developed by Herner & Co., which utilizes audio tape. In addition, there are a number of microfilm retrieval devices which have been developed for searching combined index and document files, including the IBM Walnut system, Magnavue, File-Search, Miracode, Minicard, Rapid Selector, and others.[1] [3] (See Table 8-2.) Thus far none of these machines has proven to be a commercial success or particularly effective in searching for documents by subject terms. The reasons for this lack of success are primarily related to file structure considerations which are discussed in some detail in Paragraph 8.3.

8.1.5.4 Computerized Indexes. There are a number of applications utilizing computers for searching an index file. The equipment considerations will depend upon various factors such as the size of the file, the depth of indexing, the response time requirement, and the volume and complexity of questions. The various elements of equipment utilized include a general-purpose computer, input devices for reading in the question (e.g., paper tape reader, card reader or on-line console keyboard), memory devices (e.g., magnetic tape, magnetic discs, magnetic cards), and output devices (e.g., high-speed printer, console typewriter, or display). A subject of considerable research and development is the so-called associative or content-addressable memory. A content-addressable memory is one which is addressed by the name of the data stored there, (e.g., automobiles) rather than by a directory look-up which locates the place in memory where information about automobiles is stored. It is believed that such memories may have particular application to the retrospective search problem. The subject of computer search systems is discussed in more detail in Paragraphs 9.4, 8.2, 8.3 and 8.4.

8.1.6 Document Management

The operations within the document management function are described in Paragraph 5.5.5 and cost comparisons between the various techniques are included in Paragraph 7.5. The operations within the document management function include document dissemination, document storage and retrieval.

8.1.6.1 Document Dissemination. Document dissemination can be made either in full-size hard copy or in one of the various microforms. The equipment associated with full-size document dissemination usually involves some form of plate-making equipment, which frequently is photographic or Xerographic, and some form of printing or duplicating equipment. Where microform dissemination is made, the equipment requirements include planetary microfilm cameras including a variety of special cameras, such as a step and repeat camera, film processors and various types of microfilm duplicating equipment (e.g., aperture card duplicators, roll-to-roll duplicators, microfiche duplicators and roll-to-aperture card or microfiche duplicators). This type of equipment is manufactured by such companies as Recordak Corporation, Ozalid Corporation, Technifax Corporation, Minnesota Mining and Manufacturing Company, Photo Devices Corporation, Kalvar Corporation, Bell and Howell Company, and others. A guide to microreproduction equipment is published by the National Microfilm Association.[4]

8.1.6.2 Document Storage and Retrieval. As mentioned above, there are devices which are designed to store and retrieve micro images of documents by coded index terms which are usually recorded on film in proximity to the document image. These devices (see Table 8-2), which have not found significant application, are discussed in some detail in a state-of-the-art report by the National Bureau of Standards entitled "Information Selection Systems – Retrieving Replica Copies."[3] This particular report also discusses automatic devices for retrieving micro images by direct address (see Table 8-3).

When the document address is known, there is usually no better method for retrieving the document than sending a human being to extract the document from a file. There is at least one application, the U.S. Patent Office patent copy sales function, in which the number of requests for on-demand copies is significantly large that automatic retrieval by document address appears to be feasible.

Where no automatic equipment is to be employed for retrieval, the question of whether the document will be stored in full size or in microform will depend upon two considerations: space and frequency of on-demand copying. Where space is not at

a premium and a small volume of on-demand copying is expected, the documents will usually be stored in hard copy form. Sometimes microfilming is done for security reasons, or for dissemination, in which case a working full size copy may also be desirable for on-demand copying.

8.1.6.3 On-Demand Copying. Where the system is required to supply copies of documents upon request, there is a choice between filling these requests from stock or making the copies on demand. The economics of these two approaches are discussed in Paragraph 7.5.2. The equipment utilized in providing on-demand copies typically includes the Xerox Copyflo, microfiche to hard copy printers, ordinary office copiers, and a wide variety of reader-printers. Where microform copies are supplied in response to a request, these also may be furnished either from stock or on demand. There are a few machines which are designed primarily for making single microform copies on demand, such as the microfiche-to-microfiche copiers by Kalvar Corporation, and aperture-card to-aperture-card printers by 3M and IBM.

8.1.7 Correlations

The function of producing correlations, such as state-of-the-art reports from a multitude of source documents, is basically an intellectual process and does not involve any particular equipment. The production of data correlations, as can be performed by most fact retrieval systems, is a function which thus far has been within the domain of the general-purpose computer because of the logical decision-making facilities which would be too expensive to wire into a special-purpose device having an undefined market. The precise configuration of the computer and its peripheral equipment will depend upon the nature of the problem and on file structuring.

8.2 USER FUNCTIONS AND ASSOCIATED EQUIPMENT

The following paragraphs discuss the equipment becoming available to facilitate the users' functions. None of these devices can operate independently of the system to which they are attached. Thus, these paragraphs also touch on the evolving on-line system concept.

TABLE 8-2. DOCUMENT

	Manufacturer	Process	Input	Output	
Rapid Selector	NBS	Machine scans with photoelectric cell and prints on a "bit."	Terms on punched cards. Put on 35 mm film. Search criteria on card input.	"Hits" copie on microfil	" copied crofilm.
File Search	FMA	Up to 6 questions scanned at a time, like the Rapid Selector.	Cards set up index coding Cards set up search logic	Viewing Hard copy Microfilm	ng copy ilm
FLIP	Benson-Lehner	Simply a viewer which scans with logic.	16 mm film keyboard logic	Viewing onl	ng only
Minicard	Eastman Kodak	Scanning of partial inverted file of 2nd generation positive	16 x 37 mm film chips — index on paper tape.	Hard copy only	copy
Filmorex	Filmorex	Photoelectric scanning	3 x 5 cards shingled for index code	An abstract hard copy o viewing	stract: copy or g
Miracode	Recordak	An advanced Lodestar	Flat bed camera with 9 selector slides 3 terms keyed into machine	Viewing or hardcopy	g or py

ENT RETRIEVAL DEVICES WHICH SEARCH ON IMAGE INDEXES

	PRO	CON	Capacity	Search Speed (frames per minute)	Index Info	Reduction	C
ied lm.	Can copy without stopping. Provides a trial search.	No browsing.	40,000 frames per reel.	6,000 fpm	240 binary bits	8:1	(
	Faster, more flexible	Fairly expensive	32,000 frames per reel	6,400 fpm	56 alpha 84 decimal	25:1	$
ly	Very fast scanning Large capacity	No recorder on market No hard copy Minor logic	72,000 frames per reel	24,000 fpm	32 coded binary bits	?	$
	Deep indexing	Very slow, publish only 1st page.	2,000 chips on a skewer 12 frames per chip	1000-1200 fpm	252-2730 binary bits	60:1	$ p
t: or	Cheap	Search only 3 terms Limit to depth of indexing	?	600 fpm	25 6-digit numbers	10:1	$
	Cheap for limited systems.	Superficial depth of indexing	Any no. of cartridges	10'/sec.	6-15 terms	23:1	$

B

IEVAL DEVICES WHICH SEARCH ON IMAGE INDEXES

O	CON	Capacity	Search Speed (frames per minute)	Index Info	Reduction	Cost
n copy with- stopping. ovides a al search.	No browsing.	40,000 frames per reel.	6,000 fpm	240 binary bits	8:1	(Prototype)
ster, more xible	Fairly expensive	32,000 frames per reel	6,400 fpm	56 alpha 84 decimal	25:1	$150,000
ery fast anning rge pacity	No recorder on market No hard copy Minor logic	72,000 frames per reel	24,000 fpm	32 coded binary bits.	?	$45,000
ep dexing	Very slow, publish only 1st page.	2,000 chips on a skewer 12 frames per chip	1000-1200 fpm	252-2730 binary bits	60:1	$2 million per system.
eap	Search only 3 terms Limit to depth of indexing	?	600 fpm	25 6-digit numbers	10:1	$7-25,000
eap for mited stems.	Superficial depth of indexing	Any no. of cartridges	10'/sec.	6-15 terms	23:1	$30,000

C

8.2.1 User Function

The user, when in contact with an IS&R system, performs three functions:

Query (Lookup) — he formulates (and reformulates) the search question(s).

Look at (or Browse) — he reviews selected material in order to determine its relevance (for a specific task, for current awareness or for both).

Take away — he obtains a copy of the information for use away from the IS&R system interface.

8.2.2 User Requirements

The principal user requirements on the system while performing these functions (as discussed in Section II) are:

(1) Response time (between submission of the query and presentation of selected material).

(2) Completeness and currency (up-to-date material).

(3) Relevance and specificity.

(4) Legibility and flexibility in format.

(5) Take away facility.

The equipment systems under discussion here are designed to (1) reduce the response time (sometimes called "turn-around" time), without reducing legibility, and (2) provide take away copy. The other requirements are not affected by user interface equipment, but only by central equipment and procedures.

The query function involves the entry of a search question into the system in such form that it is intelligible to the system. This may be done either through the auspices of an intermediary, e.g., a reference specialist, or by putting the user in direct communication with the system. In either event, negotiation of the search question is an iterative process inasmuch as the user can rarely state the exact parameters of his question on the first try. In other words, it is necessary that the user be able to maintain a dialogue with the system. The intermediary or reference

TABLE 8-3. DOCUMENT RETRIEVAL DEVICES WHICH SEARCH ON ADDRESSES

	WALNUT (Experimental)	VERAC (Experimental)	MEDIA (Commercial)	CRIS (Experimental)	LODESTAR (Commercial)
Manufacturer	IBM	AVCO	Magnavox	Info. for Industry	Recordak
Media	strips of microfilm	microphoto-graphic plate	film chip (16 x 32 mm)	microfilm scroll Mylar and Kalfax	16 mm. microfilm
Storage Format	bins	microphoto-graphic plates	200/cartridge	400' long scroll: 17" wide	roll in a cartridge
Searching	Mechanical selection of strip from bin	Mechanical selection of image on plate	Cartridge: manual. Chip: electronic at 600/min.	Electronic address on scroll	Motor drives roll until address is found
Image I/P	Image converted from 35 mm at 1500/hr.	Camera System	100' rolls of microfilm 240/min.	Contact printing from microfilm	Regular roll microfilm
Search I/P	Address on punched cards	Keyboard	Keyboard	Keyboard	Keyboard
O/P Form	Aperture cards 4 images	CRT display on micro-film	Display on hard copy	Display on aperture card	Display on hard copy
Image Capacity	990,000	1,000,000	400-600/capsule. Any no. of capsules	500,000/ scroll	cartridge
Reduction	35:1	70:1	30:1	?	19:1 - 24:1
Response	5 sec.	.3-2 sec.	10 sec.	20 sec. (avg)	10 sec. (max.)
Cost	(proprietary)	(prototype)	$35,000	(prototype)	$4,600

specialist is generally more familiar with the language of the system and may be able to assist the user in properly stating his question. On the other hand, since the reference specialist does not really know what the user is looking for, a bias may be introduced as a result of the reference specialist's own background and experience, and perhaps, misunderstanding of the problem.

Without equipment, the response time is in the order of one day to several weeks. This long response time not only delays progress on the task for which the information is needed, but also hampers the user's formulation of the search question.

8.2.3 Query-Response Equipment

8.2.3.1 Look-up and Look-at. The query response equipment is usually some form of a console connected on-line to a computer. The on-line system permits response times of the order of a few seconds. Where several users' consoles are tied to the same computer system, a time-shared mode of operation is needed. This requires a sophisticated executive program and interrupt facilities in the computer.

To be useful for look-at or browsing the computer files must contain an index with at least citations and possibly brief abstracts. Accession numbers alone would not permit browsing and therefore do not contribute to the look-at process.

The equipment which has been used or is contemplated in on-line I&R research studies ranges from basic keyboard (Teletype) units to sophisticated cathode ray tube displays.

Teletype permits only alphanumeric information exchange (or very rudimentary graphs). This is sufficient, however, for many search procedures. The abstract or document would be provided separately. Cathode ray devices or facsimile methods permit display of more extensive texts rapidly, and of graphics, if graphics are digitally represented in the file.

8.2.3.2 Take-Away. Once the user has looked at and evaluated the output of the search function and decided which documents he would like to have, the next problem is the delivery of them to him, with dispatch at a reasonable cost. The two methods which are most commonly utilized today are to supply the copy from a central document store, from stock

or by on-demand copying or on a decentralized basis by having previously provided the user with a reproducible set of microfilm for all documents in the store so that he may make a copy on an ordinary reader-printer or microfilm-to-hard-copy printer such as the Xerox 1824. When communication and terminal equipment costs are low enough, the copies can be delivered from the central store to the user by means of facsimile devices. This may be useful with on-line systems (regardless of whether the user has a microfilm copy available to him within his own organization), because of the desirability of seeing the document before a search is terminated.

8.3 RELATIONSHIP BETWEEN INDEX AND DOCUMENT FILES

8.3.1 General

The following paragraphs describe the relationship between the two major storage or file elements found in all search systems. Also described is the relationship between the pieces of hardware which contain the two file elements:

(1) The document file is the heart of the search system because it contains the ultimate goal of each search. The file consists of either full text documents or abstracts of the documents.

(2) The index file supplies the means by which the search is effected. The file is made up of index terms in the form of subject headings, descriptors, or key words. Each index term is connected to its associated document either by physical juxtaposition or by an address.

There are two fundamental ways by which the two files can be related.

(1) The files can be combined into a joint file. Each record in the file contains the full text or abstract of the document as well as representations of all the index terms associated with the document. With combined files, the entire file must generally be searched to be certain that all documents relevant to the query have been scanned. When a relevant document has been found, it is immediately available for output. The only exception to searching the entire file is when the file is classified into mutually exclusive subfiles. Searches within a class may then be initiated in the appropriate subfile.

(2) The files can be separated into two files. The document file is arranged in a predetermined order, such as numerically by accession number or alphabetically by citation. The index file can be arranged in several ways (see Paragraph 8.4 for a detailed description) and is used exclusively for searching. Each search may produce a series of addresses which point to specific documents in the document file. An extra step is required to locate the documents by means of the address.

8.3.2 Early Search Systems

In the early days of designing IS&R systems, cards were used for retaining the index information. The index and document files were separated by necessity for the following reasons:

(1) Tabulating cards contained insufficient room for anything more than the index data and, possibly, a citation.

(2) Edge-notched cards had room for an abstract, but the source document had to be filed elsewhere.

(3) Peek-a-boo cards contained only a hole which represented an address code which, in turn, pointed to a document.

Search system designers during the early days dreamed of concocting an integrated system which would combine the index with the document so that no extra steps would be necessary once the proper combinations of index terms had been found. A number of experimental systems were developed during the 1950's which employed

microfilm to retain the document images, and which retained the indexes in binary form adjacent to the document. The Rapid Selector was one of the first search systems; it evolved into the FMA FileSearch. The large and complex Minicard system employed film chips instead of roll microfilm. FLIP, Filmorex, and Miracode are later entries in the field. Although Miracode is too new to have been thoroughly evaluated, none of the systems has been entirely successful. A number of inherent problems have plagued them, stemming from the following:

(1) Since the entire file needed searching, extremely rapid scanning speeds were required, creating difficult engineering problems.

(2) It was cumbersome to carry the whole document along during every search.

(3) In a special-purpose machine, it was expensive to build in the logic which would permit the batching of several queries. The alternatives were increased equipment cost or inefficiency due to repeated passes over the same data.

(4) The index file was rigidly structured in serial form and could not be manipulated to ease the searching problem under varied criteria.

(5) The system was so specialized around the searching function that it offered no by-products from its large store of information, such as accession lists and other current-awareness products.

8.3.3 Computer Search Systems

As computer tape transports become more efficient, search system designers began experimenting with combined files on a computer. The Western Reserve experiment, described in Paragrapn 9.4.9, is an example. There were several advantages to such a system:

(1) The most important advantage was that the basic citation and related information was now in machine-readable form and could be used to produce a number of by-products sometimes more valuable than the search itself. Examples of these by-products are accession lists of new acquisitions, permuted title and subject heading indexes, statistical and accounting aids, as well as catalog and punched cards for external manipulation.

(2) The computer permitted a large input query batching factor, making the serial search more efficient.

(3) The logic available in a general-purpose computer permitted the utilization of much more complex questions.

(4) The computer could use a compiler thereby accommodating query languages rather than being limited to coded queries.

The disadvantages of a computer based search system using combined files were still numerous. A serial search was required, which was time consuming because the index and the record had to be passed. Also, the index file retained its rigid structure. A new problem appeared: the conversion of full text or abstracts into machine-readable form, a long and expensive process. The output from the system also was slow and inefficient.

Experiments were conducted which used full text input, but which separated the index file from the document file in the computer for the purpose of creating a more flexible index file. The University of Pittsburgh system, described in Paragraph 9.4.11, is an example. This system eliminated the serial search through the entire text, but input and output problems persisted.

It appears that a lesson has been learned in the field of IS&R. The most efficient search systems retain the index file in the computer with just enough data to facilitate all the advantages and by-products that can be produced by electronic logic. These systems retain the documents separately in graphic form where inexpensive storage, easy output, and high resolution are attainable. A number of automatic devices (see Figure 8-2) have been designed using this concept whereby the computer search produces an address as a search product, and the device uses the address to locate the document.

8.4 FILE STRUCTURE

8.4.1 Early File Structures

As noted above, the relationship between document and index files has progressed through several evolutionary stages. The methodology of organizing or structuring the files themselves has also evolved as new knowledge and better equipment have become

available. In the early days of IS&R technology, punched cards were often used for the index file. The organization of the file was inverted so that a deck of cards contained all the document numbers associated with a single index term. The advantage of the inverted file was that only those term decks which pertained to the search criteria had to be processed in the search. Consequently, most of the large card installations used either punched cards or peek-a-boo cards in inverted order. The edge-notched McBee card systems could be ordered at random but lacked sufficient selectivity and flexibility in their coding schemes to accommodate a system using even a small number of index terms. The random, superimposed coding of the Zator edge-notched Zatocards increased the flexibility and selectivity a good deal, but these systems never received wide acceptance.

8.4.2 File Structures on Tape-Oriented Computers

When the first large scale IS&R computer installations were designed, the designers separated the index file from the document file (see Paragraph 8.3), and converted the index file to computer language in the same inverted form that was used for punched cards.

The inverted structure has proven to be an inefficient concept for tape-oriented computer searching. Unlike the simple hand pulling of punched card term decks, the entire index file tape must be passed through the computer in order to draw off the inverted term record sets. Only then can the search itself begin. More recent designs have incorporated a linear search into the initial pass of the index tape. The linear search requires that codes of all index terms associated with a document be placed in a single record which contains the address or identifying number of the document. Each document in the document file is represented by such a record in the index file. The records can be arranged at random since each record must be scanned to see if its index terms agree with the search criteria. Queries can be batched so as to render each pass of the tape more efficient. A maximum number of queries can be handled efficiently in a batch. The maximum number is reached when the time to compare all the query criteria against a single record equals the average time to read that record from the magnetic tape. Any further queries in the batch will tie up the computer for extra time on each record while the tape drive stands momentarily idle. Greater efficiency would be achieved by postponing the excess queries until the next batch.

8.4.3 Random Access File Structure

The increasing use of random access devices portends a new era in IS&R search system design because it permits the use of two file structures for searching which have heretofore been impractical for many applications. The new era is thought to be so significant that a separate appendix (Appendix B) has been set aside to describe in some detail these two new methods for structuring random access files, along with several older methods.

One of the two methods is called the List Structure or Threaded File method. This method is best suited to searching situations in which a large amount of file maintenance is required and in which the time constraints on query responses are not too severe. A record is retained in file for each document. The record contains the code for each index term attributed to it. Next to the index term code is the address of another record which contains the same term code. A search for all records with a certain term code may be performed by "threading" through the file from one random access address to another. Boolean questions may be asked as each record is perused. Thus, only those records which contain at least one pertinent term are read from the random access device, saving time when compared with a serial tape search.

The second method has the same purpose (i.e., reducing the number of accesses) as the List Structure method, but it uses a different approach. It is called the Inverted List method, and has been developed at the AUERBACH Corporation. The method is best suited to those search situations where file maintenance is not heavy and where an unusually rapid response to the queries is essential. The addresses of records which have been assigned (or contain) certain index terms are arranged in sequential lists according to term. The lists are retained in random access storage. Directories are maintained for the purpose of finding the proper lists at the proper time. The directories contain the address in storage where each inverted term list may be found. A search may be conducted taking steps similar to those of an inverted punched card system. The directory is consulted to read from random access those term lists which are involved in the query. The records in the lists are manipulated (as with punched cards) to determine which addresses satisfy the query criteria. Then, those addresses are used to read from random access memory only those records which answer the query, thereby reducing the number of accesses required by the List Structure method.

A further requirement of the Inverted List method has been developed at the AUERBACH Corporation. To reduce the amount of list manipulation, a technique is employed which selects the lists most liable to be critical in the search. Only these lists are manipulated, and the number of records ultimately retrieved from random access is greatly reduced. Thus, the amount of recall is diminished and a high degree of relevance remains.

SECTION IX. INFORMATION STORAGE AND RETRIEVAL SOFTWARE

9.1 SCOPE

The previous section described the various functions performed by special-purpose hardware devices in the field of information retrieval and announcement. In most cases, each device was restricted in its scope to a specific task within the framework of a much larger system, so that an information system was composed of many specialized devices. General-purpose computers, by means of their stored logic, are able to absorb many of the specialized functions into a single logical network contained within the computer in the form of a collection of programs, often called software.

The general IS&R functions best suited for computer software are those of announcement, index operation, and document management. Theoretically, general-purpose programs could be written in a packaged form which could be used by several information systems requiring essentially the same tasks to be performed. Pragmatically, the packaged IS&R system program is impractical because the requirements of each system have too many differences. However, a number of well-conceived programs have been written to meet specific conditions, and could well be modified to adjust to altered conditions.*

Discussed in this section are the many programs which have been written for general-purpose computers. The scope of IS&R in this context is subject to many interpretations. The following types of computer programs will be covered:

(1) Permuted Title Index Programs — whereby significant words in document titles are arranged alphabetically and are published in the form of an index.

(2) Search Programs — whereby the criteria of a user-initiated question are compared with information contained in a file in order to obtain information or references to information relevant to the question from the file's contents.

---

* This section is based on a study performed by AUERBACH Corporation for the National Science Foundation.

(3) Selective Dissemination Programs — whereby index information associated with incoming documents is compared with a series of profiles representing users' interests, and a reference or abstract is sent to the user automatically whenever a close comparison occurs.

(4) Automatic Indexing and Abstracting Programs — whereby the words of the input document are analyzed by statistical, syntactical or associative techniques in order to produce a consolidated list of meaningful words or sentences bearing on the document.

(5) File Maintenance Programs — whereby additions, deletions, and changes are effected in files.

The five types of programs listed above perform separate functions within the framework of IS&R and will be considered and analyzed separately. However, most of the types process common input data, and they often operate concurrently within the same system. Table 9-1 outlines the many functions of the various IS&R programs which may be integrated into a single system by means of an executive control program.

9.2 PERMUTED TITLE INDEX PROGRAMS

Permuted title programs for producing an automatic title index are the most popular IS&R programs in general use today. Their popularity stems from their simplicity and low operating cost. The concept was first advanced by IBM's Hans Peter Luhn, who called it the KWIC (Key Word in Context) index. The program selects each significant or key word in the title of a document and arranges the words alphabetically. The rest of the words in the title are permuted around the key word so as to reflect the general subject area of the document.

9.2.1 IBM 1620 KWIC Index Program

IBM has a variation of the original KWIC index program for almost every computer it manufactures. The simplest but most inflexible program package is for the IBM 1620. Each document title and its identifying label are formatted and punched into cards as input. The label is pre-assigned and is not created by the program. At the users' option, a list of non-significant words may be punched into subsequent cards to indicate to the program those words which should be deleted from the titles. Each

TABLE 9-1. IS&R PROGRAM FUNCTIONS

INPUT	PROCESSING	OUTPUT
FILE DATA (1) Title, author, journal (alone). (2) Title, author, journal with. (a) Abstract (indexed). (b) Abstract (unindexed). (c) Full text (indexed). (d) Full text (unindexed). (3) Document locator code with index terms. (4) Other.	INDEXING (1) Permutation of titles or abstracts from input. (2) Preparation of text for rapid searching. (3) Statistical analysis of words for automatic index. ABSTRACTING (1) Statistical analysis of words for automatic abstract. PROFILE SEARCHING (2) Matching of indexed input against interes' profiles.	AUTOMATIC (1) Permuted titles or abstracts: (a) Permuted index. (b) Index bibliography. (c) Author index. (d) Other indexes (by chemical structure, formula, etc.). (2) Auto index or auto abstract. (3) Selective dissemination: (a) Full text. (b) Abstract. (c) Citation. (d) Locator code. (4) Citation index.
SEARCH DATA (1) User profile for automatic selective dissemination. (2) Query statement for initiated dissemination.	MAIN FILE SEARCHING (1) Matching of query statements against terms in file. FILE MAINTENANCE (1) Posting and deleting user profile terms to a dissemination file. (2) Posting and deleting document and terms to a linear file. (3) Posting and deleting document number to each term of an inverted file.	INITIATED (1) Search product: (a) Full text. (b) Abstract. (c) Citation. (d) Locator code. (e) Count of "hits."

remaining significant word is simply formatted into a fixed field, surrounded by its adjacent words in context, and punched into a card in unsorted order. Three additional tabulating tasks must be performed after the program is completed: the cards must be sorted, the sorted cards must be listed, and, in a separate operation, a bibliographic or interpretive list must be created to give meaning to the labels affixed to each index entry.

9.2.2 IBM 1401-1410 KWIC Index Program

In contrast to the above, the IBM 1401-1410 KWIC system is flexible and complete (see Figure 9-1). The system not only produces a permuted title index but prints a bibliography and author index as well, if the user asks for either or both. In addition, a tape containing bibliographic references is written, which may be used as the source for preparing other indexes according to specially written programs. Printing options include a choice of three page formats for each of the three printouts. As many specific words as desired may be excluded from the final KWIC Index printout. Another option provides for frequency counts to be made on all words appearing in the KWIC Index, or all words excluded from the index.

9.2.3 GE-225 KWIC Index Program

Other computer manufacturers have developed permuted title index programs comparable to those of IBM. An example is the GE package which follows the same pattern. Document titles are punched into cards and fed into the computer. Non-significant words are purged. The remaining words are sorted alphabetically (one entry for each word) and are printed in the form of an index in such a way that each index word is highlighted (usually indented) by the print format and are surrounded by the rest of the title words in context. The GE package does not provide a supplementary bibliography for the index (nor an alphabetized author index). Therefore, the codes or labels identifying each index entry must refer to addresses or to coded entries in a list prepared by a process independent of the program package.

9.3 SEARCH PROGRAMS

9.3.1 General

Figure 9-2 is a graphic description of how various kinds of search programs relate to each other. The search programs placed above the dotted line are the main concern of this section. However, the programs placed below the line which locate documents or manipulate data will be discussed briefly if they are an integral part of the search process.

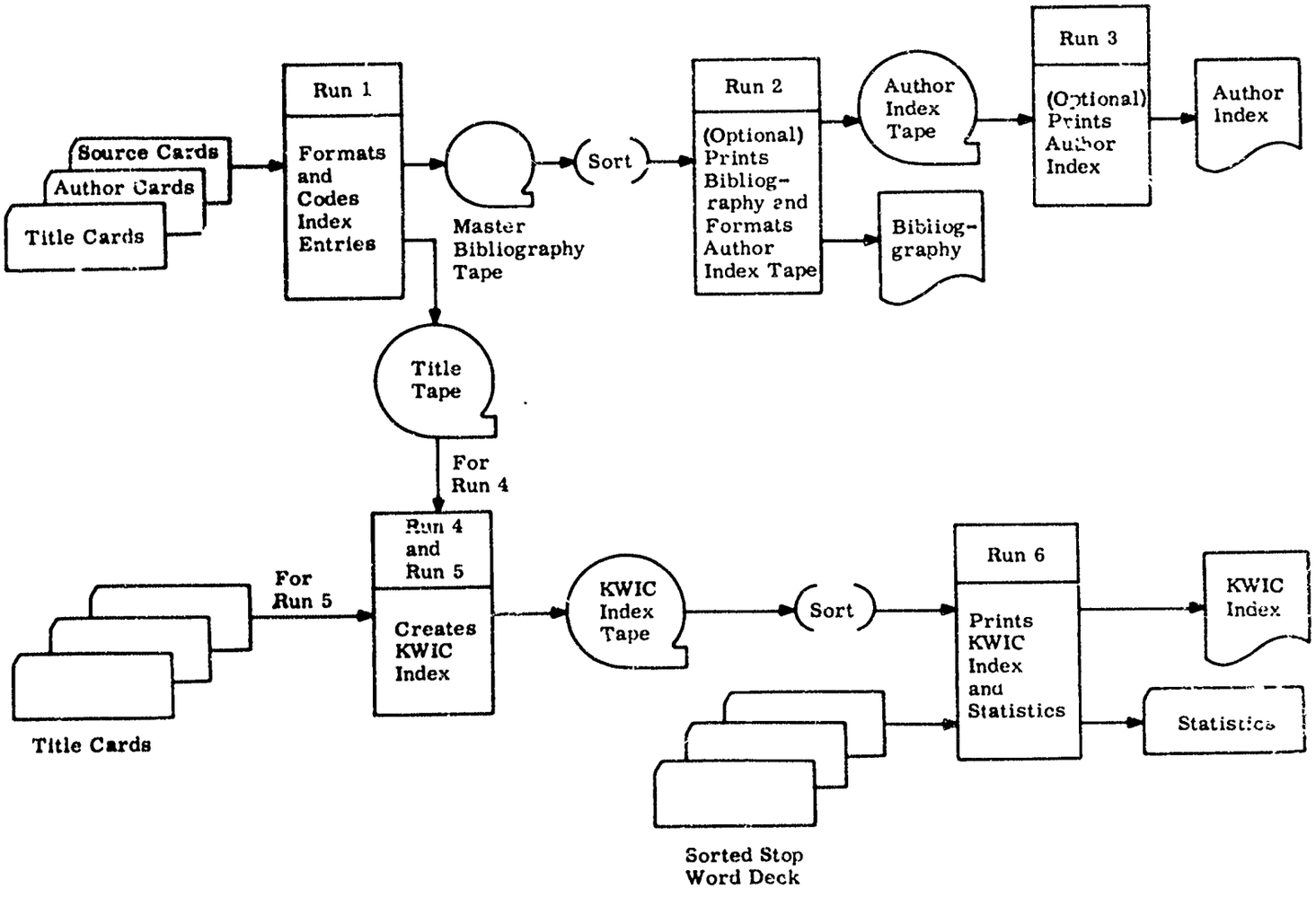
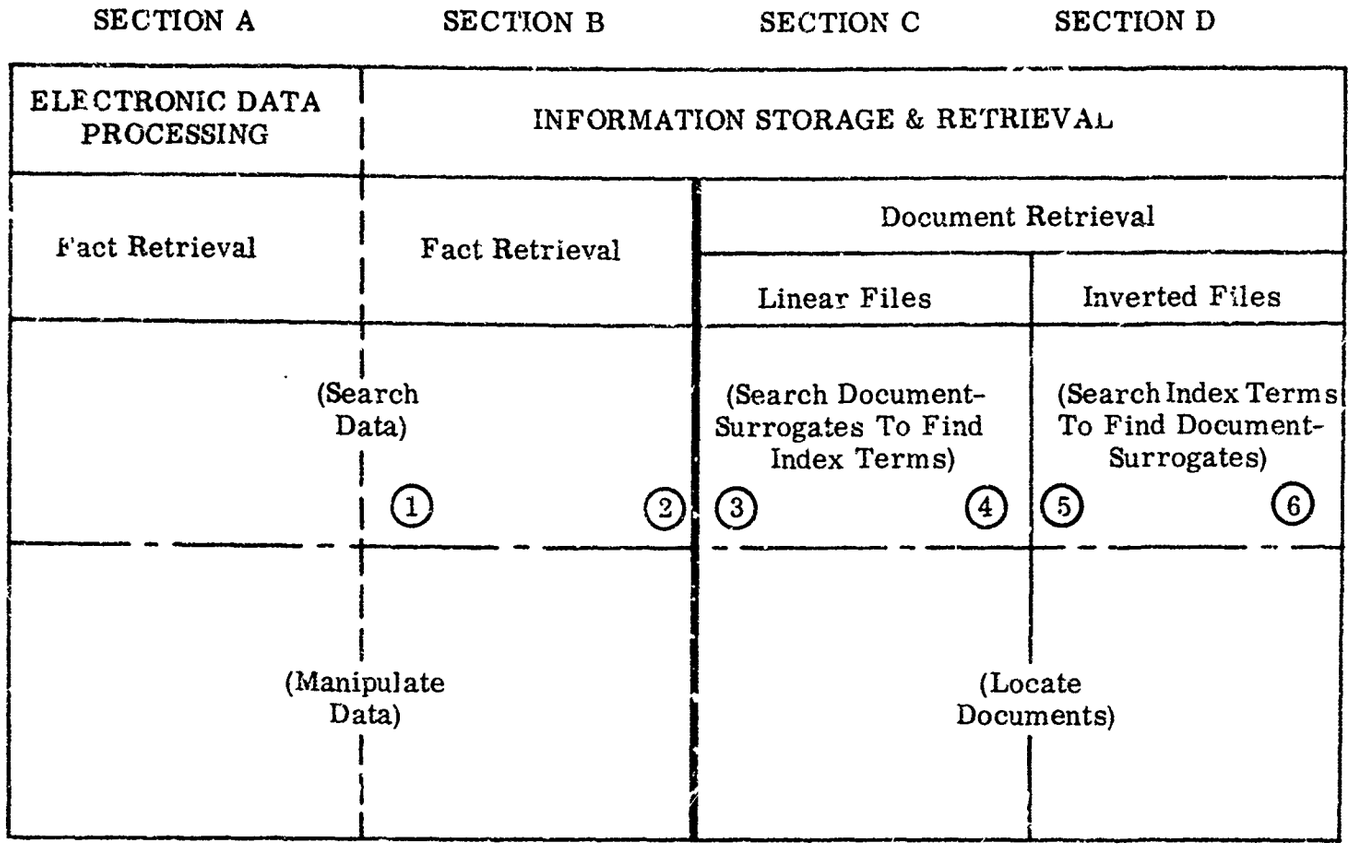


Figure 9-1. IBM 1401 KWIC Index System



KEY:

1. IBM ASK II (or RCA RECOL)
2. IBM IPS (NAVCOSSACT)
3. Documentation Inc. . Linear Search System
4. WRU-GE-225 Text Searching System
5. IBM 1401 Inverted Card File Searching (or IBM 650)
6. University of Pittsburgh — IBM 7070 Legal Searching System

Figure 9-2. Classification of IS&R Search Programs

Figure 9-2 is divided into four sections. Section A covers electronic data processing programs which have some search aspects but are not sufficiently flexible to be considered IS&R programs. Sections B, C and D describe IS&R programs both for fact retrieval and document retrieval as defined in Section II.

It is implied that, within each Section, programs tend to become more complex as they go from left to right. Six sample IS&R search programs are described in Paragraphs 9.3.6 through 9.3.11. There are two programs for each of the three IS&R classifications. An attempt has been made to choose a simple (left-most) and a complex (right-most) example for each classification.

9.3.2 Electronic Data Processing Search Programs

The various search programs which deal with files containing data or facts (as opposed to documents or references) may be represented by a spectrum as seen in Figure 9-2 between Sections A and B. As is true of all spectra, differences are those of degree, in this case, complexity and flexibility.

EDP search programs (in Section A) tend to be less flexible, with concentration on the scanning of specific data fields for specific purposes. Searching is usually an integral part of a much larger function, such as file maintenance. The product of the search is usually absorbed by the overall program for use within the system. The integrated aspect of the system leads to the consequence that if the EDP search program were removed, the system usually would collapse.

EDP search programs are typified by binary search techniques, list processing, and other table look-up functions, always with well-defined formats and predetermined search goals.

9.3.3 IS&R Programs Which Search Data

At the opposite end of the spectrum (in Section B of Figure 9-2) may be found a much more flexible kind of fact retrieval program. Whereas EDP programs may or may not have Boolean relationships or range comparisons within their search criteria, the IS&R search programs contain them almost by definition. Furthermore, the search strategy of IS&R programs, expressed in Boolean statements, tend to be more complex.

Many statements or questions may be batch-processed in one pass of the search file, each question involving a number of logical demands.

Another noticeable difference of degree between IS&R and EDP search programs is that the IS&R programs tend to stress the man-machine communication aspects of information retrieval. EDP programs will retrieve data and produce reports for human perusal on a regular basis, but IS&R programs will usually accept a tremendous variety of requests without reprogramming, and will deliver answers within a relatively short period of time. They process revised search questions. Human monitoring of the search is sometimes possible. With this feature, humans may observe the progress of a search and change the search criteria, if desired. For example, many IS&R search programs will produce a count of the number of documents retrieved, either at periodic times during the search or at the end of the search. If the number retrieved is too large, the user may increase the specificity of the search question to reduce its scope. If the number retrieved is too small, a more generic question may be asked.

In summary, the IS&R search programs are more an adjunct to, rather than an integral part of, other programs. The searches are performed to satisfy data requirements which are periodic and varied, rather than continuous and standard; but there is no absolute distinction between EDP and IS&R search programs.

9.3.4 IS&R Programs Which Search Linear Reference Files

Although the line between Sections A and B in Figure 9-2 is ill-defined, the line between Sections B and C can be defined quite clearly, since all programs to the right of the line deal specifically with files which contain either documents or surrogates to documents. Surrogate files fall into two categories:

(1) Linear Files — in which each record represents a surrogate. The index terms for each surrogate are contained within the surrogate record. Since for any search all records must be read, the file may be arranged either in random order or in any desired sequence.

(2) Inverted Files — in which the basic record represents an index term. Representations of the surrogates (usually document number) which are associated with

each index term are arrayed "behind" the term, either in the term record or in subordinate records. The index records are arranged in a precise order, and the surrogates are ordered numerically behind each term record.

This paragraph deals with IS&R programs which search linear files. In addition to file content, there is another distinction between Section B and Section C (in Figure 9-2). Search programs in Section B usually deal with search data which are contained in well-defined fields. Thus, specific Boolean search criteria may be compared with specific field definitions. Section C search programs, however, may find the index terms arranged in any order within the reference document. The program must compare the search criteria with each of the fields designated for index terms to cover all combinations of terms.

9.3.5 IS&R Programs Which Search Inverted Document Surrogate Files on Magnetic Tape

A typical search of an inverted tape file involves the scanning of the entire file, the selection and copying of all term records associated with the question criteria onto a separate tape, and the comparison of reference numbers behind each term record to see which references satisfy all the search criteria. The unique characteristic of an inverted search program is its file organization. Whereas linear searches require that each record in the file be compared with each set of question criteria, the inverted searches can wait for such detailed processing until after the pertinent term records have been extracted from the main file. Inverted file searches can be more efficiently carried out on mass random access storage devices as described in Paragraph 8.4.

9.3.6 RCA RECOL (REtrieval COmmand Language)

RECOL may be classified as a fact retrieval query language with many of the aspects of an EDP search program (see Figure 9-2). Being a language, the program must be compiled each time a search is desired, but up to five questions may be asked in a single pass through the file. The language is designed as a general interrogation scheme for small- or medium-sized files requiring a minimum programming effort.

AUERBACH®

Specification of the data format of the file is accomplished by the insertion of a table of record contents within the executive routine at the time of assembly. The table contains the names of the record items or fields, their location within the record, whether the item is numeric, and so on. The arrangement of fields within records must remain constant throughout the file, but the record items may be variable length.

RECOL processing is activated by five basic order types:

(1) A SELECT Order — whereby a file may be searched for records containing desired information by specifying the logical connectives (AND, OR, and NOT) to be associated with each record item desired. Values may be expressed as a single value, a range of values, or a list of several equally acceptable values.

(2) A NAME Order — whereby new records being entered into a file may be classified according to criteria stated in much the same manner as for the SELECT order above. The NAME order states a series of logical conditions to be met, and a list of corresponding values to be assigned to the new items if the conditions are true.

(3) An ASSOC Order — whereby record fields in a file may be compared to determine if the record contents are sufficiently related to warrant regrouping of records.

(4) An EDIT Order — whereby records may be arranged for appearance in printouts or displays.

(5) A SUM Order — whereby various levels of record counting (totals, subtotals, etc.) may be specified to determine which should be grouped together for computing average values.

The five basic orders listed above may be combined in various ways to form an interrogation statement. For example, it is possible to specify the records to be SELECTed from the file; using this subfile, the ASSOCiate order could be applied forming another subfile which could be EDITed.

The many combinations of the five search orders assure considerable flexibility in the output format. Specifically, the use of the EDIT command allows re-arrangement of fields within records, and the ASSOC command allows various groupings of output records.

9.3.7 Information Processing System (IBM IPS)

The IBM Information Processing System (IPS) was designed for the Navy Command Systems Support Activity (NAVCOSSACT) to perform a multitude of data processing tasks. IPS is a large and flexible system, with slow response speed, which is a normal characteristic of a system that permits a wide variety of queries. The IBM IPS is an example of the more complex fact retrieval systems classified to the right of Section B in Figure 9-2.

A File Maintenance Program and Information Retrieval Program are the two basic components of the system, and are used to update and query all files contained within the system. An executive routine controls the processing functions being performed by the system, which is entirely tape-oriented and which is designed for the CDC 1604A.

Retrieval queries are stated in the language of IBM IPS, which is called the Query Statement Language. The system permits many logical variations in the queries, including:

(1) The Boolean connectives of AND, OR and NOT in any combination.

(2) Comparisons of greater or less than, as well as ranges.

(3) North (N), South (S), East (E) and West (W), for use in comparing latitude and longitude.

(4) Summations and frequency counts.

(5) Arithmetic calculations, which may be performed on various fields before comparisons are made.

(6) A special capability for making such queries as "Name all ships presently positioned within 100 miles of Guam," whereby from latitude and longitude the distance function is calculated and compared.

The queries are introduced to the system by means of punched cards which are converted to magnetic tape records. Queries may be batched in groups of 100, but only one file may be referred to in each batch, because each file has its own sub-program for answering queries and only one query program may be run at one time.

The system maintains a library of query programs, one for each file, as well as a format table to describe the format of each file. A record format table is placed at the head of each record in the file to describe the arrangement of fields within each record. The query cards prompt the IPS executive routine to call in the proper library subroutine (which runs in the interpretive mode) to translate the query. Further subroutines are called in to execute the various aspects of the search logic as each aspect occurs. Consequently, the program is flexible but rather slow and quite large. Only one file may be searched at a time, because separate query subroutines are maintained for each file.

Consistent with the flexibility of the rest of the system, the output may be formed in many ways. Several sort keys may be specified, such as the ordering of output data according to the query which asked for it. The output is printed in the format which is specified by the user.

9.3.8 Documentation Inc. Linear File Search System

The Documentation Inc. System is a relatively simple and straightforward search of a linear file containing document surrogates (citations), hence, it may be classified at the left-hand side of Section C in Figure 9-2. The system employs several techniques to reduce search time, and allows considerable flexibility to the user in the framing of questions and the formatting of answers in the format of computer output.

Search questions are punched into cards in "raw" form and are converted from cards to magnetic tape. A 1401 Input Format Program arranges the questions in a validated format, assigns core memory for the processing run, internally sorts the search terms into the order best suited for machine processing, and writes excess questions (over and above core capacity) on a separate tape to be used for a second search pass.

The search questions are processed by the IBM 1410 at two levels:

(1) Limit Searching — The user may employ any or several of fourteen fields in the record as search limitations. If any of these limit fields are used in a search question, they become required terms. If they are not present in the form that the searcher has specified, no further investigation of the document will be made for that question, thereby speeding up the search considerably.

(2) Depth Searching — Having passed the limit search, the search program will process a second level of interrogations with the following optional characteristics:

(a) All Boolean expressions; this includes the binding of one term to another.

(b) Weighting of terms, whereby the user may assign relevant weights to any given term.

(c) Weighting of questions, whereby the user may assign a relevant threshold weight to the whole question, below which the answer is ignored.

An IBM 1410 sort program arranges acceptable answers by queries and relevance weights.

The product of the search query is in the form of citations to documents which appear relevant to the query. The citations are sorted by query, or by any other sort key specified by the user. The final output is produced by an IBM 1401 print program which prints the citations in the form of a bibliography.

9.3.9 GE-225 "Text Searching" System

Western Reserve University Center for Documentation and Communication Research has collaborated with General Electric Company in the development of the GE-225 Text Searching System. The system involves manual coding of abstracts such that the meaning of the words may be understood by the computer. The computer then converts the words into a format which allows rapid search of the text by computer at a later time. The actual words within the abstract may be considered the index terms. The file is linear because the abstracts are arranged in a serial fashion, and the system classification falls on the right-hand side of Section C in Figure 9-2.

There are two types of input to the system:

(1) The coded words of the abstracts are punched into cards. The Text Conversion Program converts the coded words of the text to a series of symbols which follow a precise classification system and which speed up machine searching.

(2) The Search Compiler accepts logical expressions indicating which symbols are pertinent and in what way they are pertinent. The logical expressions which may be combined are enumerated below:

(a)	Positive	A
(b)	Negative	Not A
(c)	Conjunctive	A and B
(d)	Disjunctive	A or B
(e)	Sequential	A before B
(f)	Combinational	m of $(A_1, A_2, \ldots A_n)$
(g)	Multiple Occurrence	a before A before ... before A
(h)	Nesting	B or (C and (D or (E and F)))
(i)	Magnitude (for values)	A equals B
		A not greater than B
		A not less than B

The number of symbols mentioned in an expression is limited to sixty. If core storage is not exceeded, up to one hundred expressions may be searched at one time. The Search Compiler creates the program which is used to scan the prepared text.

Search programs are created each time a new batch of questions is to be asked. Search time may be estimated by the following equation:

$$\text{search time (minutes)} = \frac{\text{symbols in text x symbols in expressions}}{150,000 \times 60}$$

For example, if texts contain an average of 150 symbols each, expressions mention an average of 10 symbols each and 60 such expressions are to be searched for at one time, then about 100 texts can be searched per minute.

One of two types of output may be requested: the texts that satisfy the expression may be named and counted, or only counted. Provisions for presenting any part of the text other than its name have not been made a part of the programs.

9.3.10 IBM 1491 Inverted Card File Retrieval System

Inverted card files may be searched with the IBM 1401 allowing a maximum of 9,999 documents, 10,000 terms (or keywords) with as many as 750 document numbers per term. The program is a fairly elementary adaptation of the IBM 650 program, and is best classified to the left of Section D in Figure 9-2. Records are maintained on tabulating cards.

The files represent library information or documentation collections which have been encoded by coordinate indexing techniques where each term record contains a series of document reference numbers.

There are two basic functions requiring input:

(1) File Creation and Maintenance — A batch of indexed documents is used to create the original or additional entries to the system. A document card is punched for each document, with a four-digit document number and four-digit term number punched in each card. The document cards are sorted into document number sequence within term number, and are placed behind the appropriate term cards, which have been pulled from the main term file. The program punches updated term cards (containing 18 document numbers per card) which are placed back in the term file, replacing the ones which had been pulled manually.

(2) Retrieval — The term file is the basis of retrieval inquiries. Inquiries are programmed by inserting AND, AND NOT, and OR program decks between term decks and feeding the whole batch into the computer.

The document numbers are compared on the series of term cards looking for matches signifying the fulfillment of the query logic. When all comparisons have been made, the resulting matches form the search product. The system is not process bound; it operates as fast as the reader, punch and printer permit.

Two print programs are available: one will print a document number per line, and the other will print selected information from a separate bibliography card file.

9.3.11 University of Pittsburgh Legal Retrieval System

The University of Pittsburgh Computation and Data Processing Center has developed a system for retrieving legal data pertaining to statutes (see Figure 9-3).* The complete statutes of the State of Pennsylvania have been converted to machine language and fed into an IBM 7070 computer. A concordance has been compiled which not only shows the frequency of each word used in the statutes, but also refers back to each word's position in the text. The frequency concordance, in printed form, may be used much like an authority list to help structure queries for the system. When recorded on magnetic tape, the concordance, with its references, becomes an inverted file and may be searched in basically the same manner as described in the previous paragraph. The files are maintained on magnetic tape. A programming language designed by the University of Pittsburgh allows the user considerable flexibility in framing his search question.

Since the Pittsburgh system provides a complex search of a large inverted file, it is best classified to the right of Section D in Figure 9-2.

The search question is prepared in terms of Boolean statements of AND, OR and NOT. A major semantic burden is placed on the user because of the wide and uncontrolled use of words in the system. Consequently, a large number of OR connectives are required to pull the synonyms together.

Short sections of the statutes are defined as "documents." Desired AND relationships of terms (or words) within documents may be stated in several ways:

(1) The words must simply exist within the document.

(2) The words must be contained within the same sentence.

(3) The words must be in the same sentence and be within a specified range (number of words) of each other.

(4) The words must satisfy any one of the above conditions, but also have one of the words prior to the other.

Search questions are punched on cards and given to the search program. The magnetic tape concordance file is scanned and all word records matching the search question

---

* Kehl, William B., John F. Horty, Charles R. T. Bacon, and David S. Mitchell, "An Information Retrieval Language for Legal Studies," Communications of the ACM Vol. 4, 1961, 380-389.

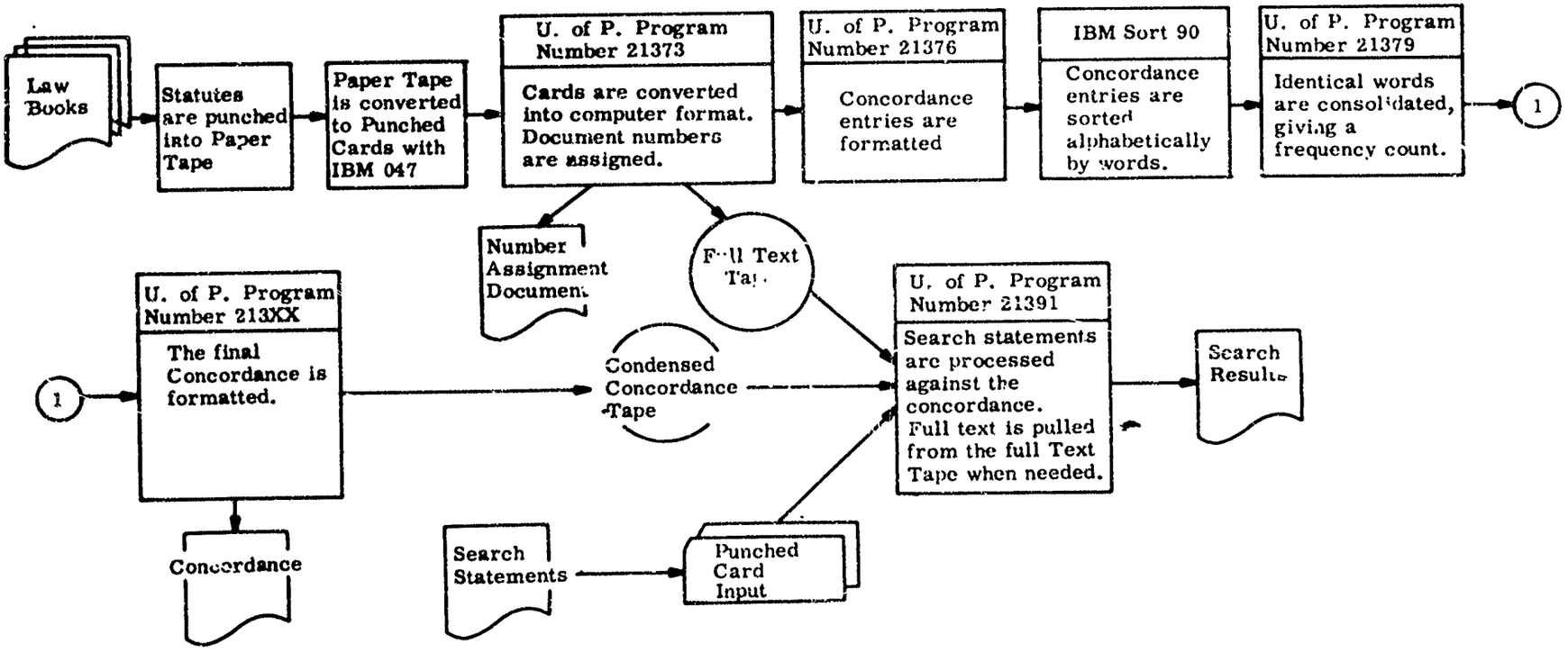


Figure 9-3. Flow Chart of the University of Pittsburgh Legal Search System

are written on a separate tape. The selected word records are then processed against the logic of the search question to determine which words satisfy the search criteria.

The search product may be sorted in any desired order. Three output options exist:

(1) The document numbers and line numbers of relevant documents are printed, five to a print line.

(2) The title lines of relevant documents are pulled from the original full-text statute tape and printed.

(3) The relevant "documents" are printed in their entirety.

9.4 SELECTIVE DISSEMINATION PROGRAMS

9.4.1 General

A description of the objectives and methodology of Selective Dissemination of Information (SDI) has been presented. IBM has devoted considerable effort to preparing SDI program packages for almost all of its business oriented computers. An example of a package for the IBM 1401 is presented below.

9.4.2 IBM 1401 Selective Dissemination of Information (SDI-3)

The IBM 1401 SDI System establishes and maintains a current list of people who join the SDI system. Each person is represented on magnetic tape by a list of keywords, originated by him, that describes his work interests. As new documents are published, a list of keywords from each document is entered into the system. The computer matches the keywords of each new document against the records of all users punching out a notification card only for users selected by the system as potentially interested in a particular document.

There are five types of input to the system:

(1) User Information — The user submits his interest profile by compiling a list of keywords which reflect his areas of interest. The vocabulary may or may not be controlled by an authority list of acceptable words.

Along with certain identifying and address information, the keywords are punched into cards and given to the computer to either create the user file or to effect changes in the file. Keywords are retained in alphabetical order.

(2) Document Information — Incoming documents are abstracted and are assigned keywords, which are punched into cards with six keywords in each card. The abstracts are typed on offset mats and held in anticipation of later dissemination.

(3) User Response Cards — After being notified by the system of a potentially interesting document, the user may respond by requesting the document or by stating that he is not interested in it, by means of a PORT-A-PUNCH produced card.

(4) Document Rating Card — After receiving a document, the user may rate it as relevant or not and send his response back to the SDI system by means of a PORT-A-PUNCH card.

(5) User Profile File — The user profile file is maintained on magnetic tape.

Several processing functions are performed by the SDI system:

(1) User Selections — The computer alphabetizes the document's keywords and compares them sequentially with those in the user profile file. If any user has a sufficient number of keywords in his profile to meet or surpass the matching criterion set for any particular document, the document is selected as being of probable interest to the user, and a notification card is punched.

(2) Notification Response Analysis — Punched card equipment is used to sort and segregate the various types of notification responses after the document has been forwarded to the user. Human analysis is performed to determine if interest profiles have changed.

(3) Document Relevance Rating Analysis — Punched card equipment is used to sort and segregate the various types of document relevance responses made by the user after he has read the document. Human analysis is performed as in item (2).

The SDI system has the following two outputs:

(1) Notification Procedure — A notification card is punched and sent to the user whenever the computer selects a document of potential interest. Before being sent, the notification card is associated with, and stapled to, a card with an abstract of the document.

(2) Document Forwarding Procedure — Positive responses from the notification card to the user initiate the sending of the actual document to the user.

9.5 AUTOMATIC INDEXING AND ABSTRACTING

The subject of automatic indexing and abstracting from full text is covered in Paragraph 3.3. The methodology is discussed, and an evaluation of the products from various techniques is presented.

From the viewpoint of computer software, the state of the art in automatic indexing and abstracting has not advanced to the point where a significant number of operational programs have been developed. A large number of theoretical and experimental papers have been written on the subject of automatic indexing. Several programs have been developed to provide empirical data in support of theories advanced. Other programs have been proposed in outline form as potential solutions requiring further refinement. But no program has yet been devised to provide program directed automatic indexes or abstracts to an operating information system directly from full text documents.

9.6 FILE MAINTENANCE PROGRAMS

The programs used in the area of IS&R to update files are much akin to the file maintenance programs found in the commercial electronic data processing (EDP) applications. The principles of addition, deletion and alteration are the same in both areas. Designers of IS&R systems have simply adopted the EDP file maintenance techniques to the specific problems of maintaining a search file. The variety of maintenance techniques is quite large, and the decision to use any particular technique depends largely on the structure of the file being maintained. File structure in an IS&R system is predicated almost entirely on the search strategy adopted for the system. A detailed discussion of file structure, file maintenance and search strategy may be found in Paragraph 8.4.

9.7 EXECUTIVE PROGRAMS

As illustrated in Table 9-1, most of the program functions previously discussed may be integrated into one large IS&R system. (At present the citation index program is a separate system since it requires inputs - citations - not normally included in IS&R systems.) The indexing and abstracting programs, the SDI program and the file maintenan

program would all use portions of the same input. The search programs and the maintenance programs would refer to the same search file. In a large IS&R system, the most effective manner in which control may be achieved is through a centralized executive system. Such a routine serves several purposes. An executive routine is essential both for the on-line communications problem and for the retrieval problem. For the latter, the speed of retrieval can be enhanced if the executive batches questions for tape files and provides features to enhance disc or drum retrieval.

The precise scope of the executive routine depends upon the problem parameters. For example, on-line communication may or may not be a system requirement. If it is, then provisions must be made in the executive routine to provide for such a feature.

Some of the functions of an executive routine are:

(1) To service hardware interrupts.

(2) To interpret inputs to determine their origin, the destination of the output, and security requirements.

(3) To maintain a task list queue of processing requests and on-line inputs that have been brought into the system.

(4) To initiate indexing, abstracting, storage, retrieval and dissemination programs in their proper order.

(5) To maintain a queue of input-output calls; calls for peripheral equipment and to provide centralized input-output editing functions.

(6) To maintain directories for all files, to permit translation of terms to access addresses and to execute accessing the data in the files.

(7) To maintain a directory to descriptions of all records in a file.

(8) To compile search programs from inquiries stated in a query language.

There are likely to be many different machine inputs. Inputs may enter from punched cards, paper tape, and on-line interconnections. The format of the input will

vary according to the nature of the data and the operations to be performed. A scanning routine can be developed to operate upon the input to determine the type of input and to transfer to an appropriate routine which operates upon the data.

Within the system complex, there are likely to be many files. Each file can have a different record format. Indeed, record formats may change. Such changes must be made in a manner such that havoc is not wreaked within the entire system. This problem can be solved if data requests are presented in terms of the data name rather than the data location. The executive program using a directory which specifies the data location provides proper access to the data location. Thus, changes to records in one part of the system need not cause changes to other parts of the system.

A major portion of a retrieval program will be the process involved in input-output requests for data. It is possible to develop input-output routines which will facilitate the use of the peripheral equipment and thereby speed retrieval. If a disc file is part of the system, requests can be queued and sequenced by an executive routine in such a fashion as to minimize disc positioning time.

An excellent example of an executive routine which controls the programs of a large IS&R system is a part of the IBM Information Processing System (IPS) described in Paragraph 9.3.7 above. This system involves mainly file maintenance and information retrieval but the system is sophisticated and the executive routine performs all of the eight functions outlined above.

SECTION X. EVALUATION OF THE STATE OF THE ART AND PREDICTION OF TRENDS

10.1 SCOPE

This section presents a brief evaluation of the present state of the art and a prediction of trends for each IS&R system function.

10.2 ORIGINATION

The origination function, which involves the origination and initial publication of information encompasses the technologies of technical writing, typesetting, platemaking, printing, and distribution (e.g., subscription fulfillment).

10.2.1 Evaluation

The present state of the art involved in the publication of the serial (journal) literature is considerably behind existing technology. Most journal publications are still produced by Monotype or Linotype composition and letter-press printing. The costs of journal production are increasing due to increases in the manual labor costs inherent in these methods. Subscription incomes are remaining relatively stable due to a trend toward subscriptions by corporations and libraries rather than by individual subscribers. To meet the problem of the decaying financial situation, many journals are relying upon government financial support, page charges, and other means of additional financing.

The technical report literature (separates) which is continuing to grow at an accelerating pace, has typically been characterized by a lesser emphasis on typographic quality. The usual method for producing a technical report is by typing on offset paper masters and printing on offset duplicators. The technical quality of the report literature is not consistent since there is no formal refereeing process as in the journal literature published by professional societies. Finally, the distribution process for reports is woefully ineffective since it is in the hands of the originators rather than in the hands of the potential readers (e.g., the freedom to subscribe to a journal).

The state of the art in typesetting and printing has advanced considerably beyond that which is presently being applied by either the publishers of report literature or journal literature. This is illustrated by the advanced technology being employed by some of the nation's newspapers, by developments within the Government Printing Office, and by the MEDLARS system of the National Library of Medicine.[7] [14] [40]

10.2.2 Predicted Trends

10.2.2.1 Publication of Separates. The publication of separates (reports, reprints, conference proceedings, etc.) will grow faster than the serial literature. There will be a gradual increase in both the intellectual and typographic quality of the separate literature, particularly reports. Methods for refereeing of reports and for standardization of report formats will be introduced. It is also predicted that a number of Government agencies will begin to re-originate technical reports received from their contractors in two new forms. One of these will be in printed form utilizing graphic arts quality composition. This will be accomplished by the utilization of central automatic typesetting facilities such as the facility established for the MEDLARS system at the National Library of Medicine and the facility being established by the Government Printing Office. Either by-product paper tapes will be required from the contractor, or automatic page readers will be employed to transform the original text into machine language. The second form will be microform copies, which will be disseminated to a large number of information centers throughout the country. These information centers will maintain larger collections in microform than would otherwise be possible. In particular, it is predicted that if microform copies are available from the Government Printing Office, the so-called "Depository Libraries" designated by Congressmen in each state will elect to receive much greater quantities of free copies of Government reports than at present because of space limitations. If this prediction materializes, there will be a significant increase in the use of photocopy devices in hundreds of depository libraries throughout the country.

10.2.2.2 Cooperative Typesetting of Serials. The cost and time lag in the production of serial literature will gradually be reduced by the establishment of automatic typesetting facilities for cooperative publication of serials. By combining the production load, and thus the buying power of a number of journals, the resulting cooperative group will be able to afford the high investment and fixed charges involved in the operation of automatic typesetting equipment. The result will be considerably reduced unit costs and faster response time in journals.

10.2.2.3 Secondary Distribution of Separate Literature. Improved availability of the report literature is already being accomplished by hundreds of information centers, which acquire and announce the existence and availability of the separate literature in abstract journals or the equivalent. These announcement media are scanned by technical personnel as a means of maintaining technological current awareness; items of interest are then obtained by ordering them from the information center. It is predicted that the secondary distribution of separate literature in this fashion will grow rapidly.

10.2.2.4 Information "Packages". It is predicted that there will be a trend toward the standardization of information "packages." These packages will be of two types. The first type will be data packages, which will be utilized in such areas as engineering project reports, wherein the information can be converted to a convenient "package," i.e., to a data sheet. Such compact "packages" are easy to index, store, retrieve, and disseminate. (This is a trend from document to data storage.) It is predicted that a considerable amount of this "packaging" operation will be forced upon the originator himself by providing him with highly formatted data collection forms. The second type of information "package" encompasses the abstracting, indexing, and cataloging functions. It is predicted that standard or "recommended" formats will be provided to authors and publishers who will be required to provide an abstract, catalog cards, and suggested index terms along with the article itself.

10.3 ACQUISITION

The acquisition function includes the acquiring of documents, evaluation, selection, duplicate checking, and accessioning.

10.3.1 Evaluation

The fact that the acquisition function is one of the most important in an IS&R system is often overlooked by newcomers to the field. It is primarily an intellectual function and, consequently, does not generally involve any hardware. The acquisition function is performed reasonably well by most well-established information center operations; however, the procedures are not adequately formalized. In intelligence systems, in particular, the acquisition function is extremely critical, and formal procedures for acquiring documents and for evaluating their validity are essential. Improved methods

are conceivable for checking duplicates and for selection or evaluation. Most of these improved methods would not be economically justified, however, where only a small percentage of duplicates is experienced or where the cost of processing non-relevant material is less than the cost of preventing the selection of such non-relevant material.

10.3.2 Predicted Trends

It is predicted that there will be a significant increase in formal efforts to determine individual user requirements for information systems. As a result, these user requirements will directly affect acquisition policy. This will be particularly true in the field of management information systems. In the area of scientific and technical information systems, a clearer understanding of needs should result in a greater amount of interaction between information center operations and a clarification of the center's acquisition assignments. This, in turn, will result in a tendency to develop compatible thesauri or compatible vocabularies, as well as a tendency to develop compatible forms of information interchange, i.e., microfiche, aperture cards, EAM cards, magnetic tapes, and announcement media.

10.4 SURROGATION

The surrogation function, which includes cataloging, abstracting, and indexing, often accounts for as much as 50 percent of the total cost of an information system. This was not the case, however, in the cost analysis of the two government information centers referred to as Center A and Center B in Figure 7-1 because these centers service the entire nation. Consequently, the initial analysis and input costs were amortized over a considerable number of requests. Such is not the case, however, in a local or an internal company information system, either technical or administrative. The surrogation operations are inherently intellectual and, for the most part, have not been successfully mechanized, although considerable research and development is underway in the fields of automatic indexing, automatic abstracting and mechanical translation.

10.4.1 Evaluation

10.4.1.1 Cataloging. The descriptive cataloging of an item is a surprisingly expensive task as illustrated by Figure 7-1. A number of approaches have been investigated for ameliorating this situation. One of these is cooperative cataloging which is currently

being investigated by the Library of Congress and by several of the university libraries in the Massachusetts area. One of the key problems is getting the cataloging information into machine interpretable form.[18] This involves producing a machine-readable by-product tape on devices such as a Flexowriter as well as the standardization of code representations for "tagging" the various information elements contained on a library catalog card.

10.4.1.2 Abstracting. Various types of abstracts can be prepared manually, e.g., indicative abstracts, informative abstracts, extracts, author abstracts, and the like. It is difficult to evaluate the advantages of each of these types as there is considerable difference of opinion on the subject. Since the abstracting function is costly, it is reasonable to assume that less costly methods of abstracting, e.g., author abstracts and indicative abstracts will be acceptable for most purposes and they are, in fact, becoming more prevalent. The indicative abstract also has the advantage of not being biased to a given field of interest and hence is more readily usable in the re-acquisition mode by other information center operations.

To date, auto-abstracting still remains in the research stage. It is considered that the problems associated with automatic abstracting will be more difficult to solve than those associated with automatic indexing because relationships as well as terms must be manipulated. This is due in part to the syntax problem associated with creating sentences.

10.4.1.3 Indexing. Operating information systems almost exclusively employ manual intellectual subject indexing as opposed to automatic indexing. The pure Uniterm or "free indexing" technique has generally been abandoned in favor of some form of controlled vocabulary. It has also been found that index terms may properly consist of more than one word. This is, in part, due to the fact that a single word is frequently insufficient to describe a concept. For example, in the field of law, the concept "last clear chance" would not be well represented by the words last, clear, and chance. The state of the art in the control of indexing vocabularies is sufficiently advanced to provide adequate solutions to the semantic problems. Thesauri, which provide cross references among vocabulary terms to set forth explicit, synonymous, hierarchical, and various

other relationships, are becoming common. Such thesauri, in effect, provide a sort of "road map" to the vocabulary enabling the indexer or searcher to employ the most desirable number of terms of correct connotation.

In addition to such semantic control of indexing (and of retrieval), syntactical controls are being utilized to minimize "false drops" – the retrospective retrieval of non-relevant documents from large collections. In general, a small set of common term modifiers is employed; these are known as role indicators. The well-established technique of document subdivision has been rediscovered by the documentalists and is utilized in a new form known as associative linking. Complex documents are indexed as though they are two (or more) individual documents; each separate set of index terms which results is known as a "link."

A great amount of effort has been expended during the last decade in attempting to develop automatic indexing techniques. There are three basic types of automatic indexes; the permuted title index, the citation index, and automatic subject indexing from full text.

The permuted title (or Key Word In Context) method of indexing has proven to be moderately useful both for current-awareness purposes and for retrospective search. Its utility is limited for retrospective purposes since words and titles are not necessarily the best index entry, and in any event, indexing is shallow. Attempts to correct deficiencies of the permuted title indexes by augmenting "cerebrally" the titles with added index terms eliminates completely the only real advantage of such indexes - timeliness, and low cost.

Another type of automatic index which has been relatively successful is the citation index. A citation index is particularly useful for bringing a bibliography up to date. This is accomplished by tracing the history of an article by observing what other articles have cited it. The raw input to a citation index system is the citations contained in an article. These are keypunched, organized by a computer, and printed out in book form. In order to create a citation index for a narrow specialty, the entire literature relevant to that specialty must be processed. Therefore, citation indexes will probably only be produced in broad fields such as medicine, law, science, chemistry, etc., and these will be produced centrally.

The third type of automatic indexing involves the attempt to develop deep subject indexes based on the analysis of full text input by a computer. This type of automatic indexing is still in the research stage and is not considered to be within the present state of the art. The techniques are based on a statistical analysis of the full text input, including frequency counts of words and associations between words such as co-occurrence of word pairs. These techniques are described more fully in Sections IV and VIII. The basic technical problems to be overcome are semantic and syntactic. These same problems exist with human intellectual indexing; however, some practical manual solutions exist as described above. The primary economic problem associated with automatic indexing is the necessity of putting the full text into machine-readable form before it can be computer processed. This problem may be solved within the next five years by the use of by-product tapes from automatic typesetting and by automatic page reading devices.

10.4.2 Predicted Trends

10.4.2.1 Cataloging

(1) Standards — It is predicted that standards will be adopted for descriptive cataloging, covering format and coding for machine interpretability.

(2) Cooperative Cataloging — It is predicted that cooperative cataloging efforts, such as those presently being experimented with by the universities in the Massachusetts area, will expand.

(3) Substitutes for Cataloging — It is likely that inexpensive substitutes for the cataloging function will slowly be adopted by small information center operations and as a temporary measure where catalog cards are later obtained, from a central bibliographic processing source. Such techniques include utilizing a copy of the title page and perhaps the first and last page of a document. This concept is being utilized by the Central Intelligence Agency in its so-called DARE system.

10.4.2.2 Abstracting

(1) Indicative Abstracts — It is predicted that there will be a trend toward indicative rather than informative abstracts in the centers covering broad fields since these are less expensive to produce. There will nevertheless continue to be a considerable amount of informative abstracting, probably in the form of digest journals for narrow fields.

(2) Author Abstracts -- Author abstracts will be more widely utilized than they are currently.

(3) Cooperative Abstracting — It is anticipated that the major discipline-oriented abstracting and indexing services, such as Chemical Abstract Service and Biological Abstracts, will begin to provide indicative abstracts which can be re-used by the specialized interdisciplinary abstracting services. The 18 major discipline-oriented services produce the bulk of the raw material which could be utilized by the several hundred interdisciplinary abstracting services. Consequently, it is hypothesized that some form of cooperative venture will result wherein the present level of duplication of effort is reduced. The technique described under Paragraph 10.4.2.1, item (3) may also grow as a substitute for abstracting.

10.4.2.3 Indexing

(1) Automatic Indexing — The permuted title index, which is relatively simple and inexpensive to produce, will be adopted slowly by under-financed information center operations for current-awareness purposes. The citation index is also likely to grow; however, these will be produced centrally by such organizations as the Institute for Scientific Information and by Shephard's Citations. Automatic deep subject indexing from full text will not be utilized operationally for the next five years.

(2) Manual Indexing — The use of thesauri will become more widespread and it is predicted that standard thesauri will be developed for broad fields. An example of this trend is the thesaurus of Engineering Terminology produced by the Engineers' Joint Council. More specific and more detailed thesauri will be produced by local organizations for collections in narrower fields of interests. The use of syntactical control devices, such as role indicators and associative links, will increase somewhat over the next five years; however, their use will be limited to large collections.

10.5 ANNOUNCEMENT

The announcement function helps to serve current-awareness requirements by announcing newly obtained documents through the medium of announcement journals, abstract journals, and book-form indexes. Another form of announcement, which has gained in popularity in the last few years, is selective dissemination of information, which is based on a matching of user profiles against the index terms assigned to newly accessioned documents.

10.5.1 Evaluation

10.5.1.1 Announcement Journals. The state of the art has advanced rapidly in the last few years in the area of mechanized techniques for producing announcement media. The NASA system described in Section VI and the MEDLARS system of the National Library of Medicine, which is briefly described in Paragraph 8.1.4, illustrate the advancements which have been made.[7] [14] [40] It is expected that more and more announcement journals will be prepared by mechanized systems involving tape typewriters, computer processing and high-speed tape-operated photocomposing machines.

10.5.1.2 Selective Dissemination of Information. SDI has some inherent technical problems. The index terms assigned to a newly accessioned document are matched against an "interest profile" for each user. When a match occurs, the user is usually notified by sending him an abstract of the document which he can look at to decide whether he has any real interest in seeing the actual document. Interest or no interest is recorded on the SDI card which, when returned, is used as feedback to modify the "interest profiles." A major difficulty is that such feedback seems to be unable to modify the logical structure of the interest profiles. Accordingly, most SDI programs to date have degenerated into using, as interest profiles, merely logical unions of all terms of interest to a potential user. The result is that users tend to be deluged with too much non-pertinent information. Another major difficult of SDI systems is the language problem. Until recently, most SDI systems utilized automatic key-word-in-title indexing of documents for document profiles and the list of words supplied by the users for "interest profiles." The chances of obtaining an accurate match between these two profiles are considerably reduced causing the SDI system to ignore documents of potential interest to the user. SDI can probably work in situations in which feedback can be provided from the potential customer direct to a skilled human intermediary and situations in which the system vocabulary can be controlled. It is by no means certain that a broadbased, centralized, SDI system employing automatic feedback will ever be highly effective. A significant consideration is that SDI systems tend to be very expensive.

10.5.2 Predicted Trends

10.5.2.1 Contents Journals. It is predicted that simple title listings, such as the publication Current Contents, will gain in popularity as an announcement medium because of their ease of production and ease of use.

10.5.2.2 KWIC Indexes. Key-word-in-context (KWIC) indexes will also increase somewhat in use as an announcement medium.

10.5.2.3 Selective Dissemination of Information. SDI systems will be employed more widely primarily because of the attractiveness of the concept, in spite of the rather low level of perfection which has been achieved thus far. Many of the new SDI systems may simply involve a slight upgrading of manual procedures which have been employed by libraries for years. The result may be that SDI systems will become more dependent upon the man in the system than upon the machine.

10.5.2.4 Automatic Composing. It is predicted that more reliance will be placed upon computers, photocomposition equipment, and direct image platemaking materials in the production of announcement journals.

10.6 INDEX OPERATION

Index operation involves the functions of storing and searching index data to retrieve the address or other surrogate of a record or document. The index operation function is the one which has most often been mechanized, although mechanization is not essential. In "fact retrieval" systems (see Paragraph 10.8), the index to the record as well as the record itself may be stored on the same machine, although usually in separate files. In "document retrieval" systems, the index and the record (document) are usually stored and retrieved in completely separate operations.

10.6.1 Evaluation

10.6.1.1 Separation of Index and Document Files. Search system designers have, for the last 15 years, been developing hardware systems which combine the index with the document so that no extra steps would be necessary once the proper combination of index terms has been found. Many of these systems employ microfilm for the document images with index coding in binary form adjacent to the image. None of these systems have been very successful due to the following inherent problems:

(1) Generally, the entire file must be passed. The file is expanded because the document is carried along during each search.

(2) None of these machines contain adequate logic to permit the efficient batching of queries. As a result, the entire file or section of a file is passed for each question.

(3) The index file is usually rigidly structured in serial form and cannot be manipulated to ease the searching problem under varying criteria.

(4) These systems are all specialized around the searching function and offer no by-products from the large store of information maintained. Such by-products as accession lists, book-form indexes, and other products are available when index data is stored in a digital computer.

It is therefore believed that the index file should be separated from the document file. The document file can be arranged in numeric order by accession number, in which case it is usually more efficient to have a human retrieve a document from a manual file (especially with unit records) than to attempt to perform this function automatically.

10.6.1.2 Large Index Files. General-purpose computers with special-purpose programs have been found to be relatively successful for file maintenance and searching of large index files. It is usually necessary, however, that the computer also be utilized for some other application or system function so that the entire cost or rental of the computer may not be directly chargeable to the index operation function. Computer systems for fact retrieval are similar in many respects to computer index files which provide addresses of documents or document surrogates. Fact retrieval systems which provide answers to specific questions are discussed under the correlation function (Paragraph 10 8).

10.6.1.3 Small Index Files. For small files, manual indexes in book and card form have been reasonably successful. Traditional card catalogs and book form indexes are widely utilized. Of the more recent developments, the peek-a-boo concept embodied in the Termatrex system and the tabulating card have been most widely applied. One of the primary difficulties in the Termatrex system is file maintenance — the recording of new entries into the file. With Termatrex, you must pull the term card for each index entry for a given document and drill a hole in the appropriate space dedicated in each card to the particular document number. This operation is quite time-consuming and prone to error.

10.6.1.4 The Intellectual Aspect of Searching. Searching is basically an intellectual process. The user must properly formulate the logic and terminology of his query. This is a complex task which often requires either an expert human intermediary (a reference librarian) or a dialogue between the index and the searcher, thus permitting him to modify his query (by choosing additional or different terms and/or logic) prior to the actual search. Today the intermediary approach is common; the "dialogue" approach may become practical as real-time, remote, computer interrogation systems become economical for this purpose.

10.6.1.5 Performance Measurement. One of the problems associated with the operation of an index has been the inability to adequately measure performance. The parameters of performance which should be measured are not known. In the last few years, however, it has been generally agreed that the principal performance parameters are timeliness, cost, completeness, relevance, and specificity of the information retrieved. The first two parameters are fairly easy to measure; the last three are difficult to measure. Furthermore, even if performance can be measured, user requirements must first be determined.

Despite a lack of progress, attempts at measuring performance are certain to (and must) continue. Unless performance can be measured, IS&R will progress slowly.

10.6.2 Predicted Trends

10.6.2.1 Automatic Document Retrieval. Automatic document retrieval systems, which combine the graphic and index information and search on index terms, will not gain acceptance. Relatively inexpensive document retrieval devices which search on document number or address may meet with limited success.

10.6.2.2 Computer Index Searching Systems. The trend will be toward computer search systems which provide at least a title if not a complete bibliographic citation. More systems will begin to utilize random access files as their cost becomes more attractive. Considerable research will be conducted in the area of file structuring and search strategies for random access type searching. Natural language query languages will be developed which will provide much greater flexibility to the user in formulating his query and will eliminate the necessity for reprogramming for special type questions. Manufacturers will

provide search programs and programs developed by the major Government information centers will also be made available. These will not be sufficiently flexible, however, to satisfy the needs of most internal information systems. Tailor-made designs will remain the general rule.

10.7 DOCUMENT MANAGEMENT

The document management function includes document dissemination, document storage, retrieval, and document replication. Document dissemination may be made either in full-size printed copies or in microform, such as microfiche and aperture cards.

The document storage function may serve several purposes: storage for the purpose of automatic retrieval, storage of an inventory of copies for supplying on-demand requests, storage of a master copy for furnishing on-demand requests by photoduplication, and storage of a copy for occasional reference.

10.7.1 Evaluation

10.7.1.1 Document Dissemination. The processing of documents for dissemination in hard copy printed form or in microform is a reasonably straightforward problem. For hard copy distribution, the techniques generally involve inexpensive photographic plate-making processes and offset duplication. For microform dissemination the aperture card has become popular for single page documents, and the microfiche is also becoming popular for multiple page documents. Both of these have the advantage of being unit records which are easy to file and retrieve manually directly from a file by number. In both cases, an eye legible title is provided. The technique for duplicating aperture cards and microfiche normally involves card-to-card contact printing, using silver, diazo or Kalvar film. More equipment is available for duplicating aperture cards because standards for aperture cards were promulgated by the Department of Defense several years ago, whereas standards dealing with format and size of microfiche are just now being promulgated by the National Microfilm Association.

10.7.1.2 Document Retrieval. The mechanization of document retrieval by searching a combined index and document file was evaluated in Paragraph 10.6.1.1. Mechanization of the retrieval of a document by address or document number is competing with the simplest

and most effective of all clerical operations, i.e., filing. Where there is considerable activity in a file, there may be justification for an automatic document retrieval device. An example of a highly active file is the U.S. Patent Office Patent Copy Sales function which retrieves and supplies 25,000 copies per day by patent number, from a file of 3,000,000 different items. While the Patent Office has been planning the mechanization of this function for the past five years, it has thus far been unsuccessful in obtaining the necessary funds from Congress.

The argument in favor of automatic devices for retrieving documents by address is that the documents can be filed in random order, thereby preventing misfiling. Thus far, such devices have not met with commercial success. An example of probably the most active file in existence is the Social Security Administration National Employee Index. It is interesting to note that while this file is created and updated on 750 reels of magnetic tape, the search function is performed by humans on over 1,000 microfilm reader devices, in particular the Recordak Lodestar. The computer output is automatically recorded on microfilm by a magnetic tape to microfilm printer.

While it is generally desirable to separate the document from the index file, it may still be desirable to store the document digitally in some form of computer memory although separate from the index. The criteria for deciding whether a record or document should be stored digitally or graphically generally include the size of the document and the activity or number of accesses to it. As a general rule, documents under 200 characters will be stored digitally and documents over 1,000 characters will be stored graphically, expecially if the information is best expressed in pictures or drawings. The choice of storage in the range between 200 and 1,000 characters will depend upon the particular requirements of the system and on economics.

10.7 1.3 Pre-Stock vs. On-Demand. The problem of deciding whether to pre-stock copies or produce replica copies on demand is common to all information centers which supply document copies on request. The cost per copy printed is almost always cheaper as the number of copies being printed increases. The cost per copy printed even with a small print run of 25 or 50 copies is still considerably cheaper than the cost of making a single copy on demand. This assumes, however, that all copies will eventually be requested or sold. Therefore, it is necessary to take into consideration the expected demand for a

particular item, the storage space required for the inventory, the length of time in storage, and the cost of financing the inventory

10.7.2 Predicted Trends

10.7.2.1 Microfiche. It is predicted that microfiche will rapidly gain in popularity as users begin to realize that major information centers are disseminating microfiche in a standard form and format. It is expected that microfiche will become competitive with microfilm in acetate jackets even where there is little, if any, multiple copy requirement.

10.7.2.2 Pre-Stocking. It is predicted that pre-stocking of copies by a central information or document center will continue to be practicable except in high activity collections in which case mechanized on-demand copy systems may be feasible.

10.7.2.3 Satellite Collections. It is predicted that there will be a trend toward decentralization of physical access to document collections by the widespread dissemination of microform copies. The bibliographic control function and the index searching function will continue to become centralized. As a result, the use of facsimile devices in document storage and retrieval will be limited to those areas in which immediate access to the document is required and the communication link is readily available.

10.7.2.4 On-Demand Copying. The volume of on-demand copying will greatly increase because of the decentralization of reproducible microform collections. Copy costs will continue to decrease while quality will improve. Low cost reader-printers will be developed which will be simpler to operate and more convenient to use.

10.7.2.5 Microfilm Standards. Standards for the various microforms will slowly be adopted.

10.7.2.6 Microfilm Acceptance. Users will gradually begin to accept microfilm as a substitute for hard-copy. For the next few years, however, microfilm will be utilized more for browsing than as a substitute for take-away copy. Consequently, "take-away" facilities in the form of reader-printers or separate printers (e.g., Xerox 914, 1824), will be required at user's stations.

10.8 CORRELATIONS

Correlations provide users with the exact information required at the moment. For current-awareness purposes, the usual example of a correlation is the state-of-the-art report which synthesizes information a user may require at some future time. While such state-of-the-art reports can also be prepared retrospectively upon request.they are generally time-consuming and expensive because of the extensive intellectual effort involved. They are justified if the decision to be made on the basis of the information is a big one. This report is an example of a retrospective correlation of largely non-quantitative information.

Correlations of quantitative (or near quantitative) information fall into the area we have defined as fact retrieval, which has become an active field particularly since the advent of computers.

Fact retrieval is characterized by the finding of specific answers to specific questions, e.g., to the question — "What was the dividend paid to stockholders of the ABC Corporation during the second quarter of 1964?" The response to this question would be a specific piece of data - the dollar amount of the dividend per share.

10.8.1 Evaluation

10.8.1.1 Small Fact Retrieval Systems. For small fact retrieval systems, the data is usually highly formatted. Punched card tabulating equipment is often utilized because of the ability to sort and perform simple operations such as addition or listing without too great an expense. Manufacturers of computers are beginning to build general-purpose retrieval programs which may operate on many types of files. An enterprising data-processing center may build such a routine to sell computer time, or an organization such as a professional or trade association serving many small organizations may sponsor such an effort. The file structure utilized is generally quite simple and the information is usually organized under one particular key or index term. This is practical only with small files and permits only simple type questions.

The state of the art in retrieval languages is at an early stage of development. An example of an elementary type of query language is RECOL, which is discussed in Section VIII.

10.8.1.2 Large Fact Retrieval Systems. Where large files are involved, sequential searching of magnetic tape becomes both time-consuming and expensive Random access files which are on-line to a computer provide a practical solution when accompanied with efficiently designed file structures. Until very recently, large-volume, random access files have been exceedingly expensive. Indications are that new developments in random access devices are likely to overcome the cost problem. An example of such a development is the magnetic card concept employed in the RCA Model 3488 system and the IBM Model 2321 Data Cell Drive.

The choice of file structure within a random access device is dependent upon the problem. Multi-list file structures, chaining, address generation techniques, and dictionaries have all been found valuable. A more complete discussion of random access file structures and devices is contained in Appendix B.

For large files, the present trend is toward the development of a "free input" form, rather than the restriction of the input to fixed fields on a card. This is due to the fact that gaps in information exist and data sometimes exceeds the capacity of an 80-column card. While formatted rather than unformatted records are apparently more prevalent even in large data systems, it is believed that natural language inputs will begin to become more prevalent. Character recognition devices will be useful to solve the conversion problem. Editing programs will be utilized to generate the machine record; storage and cross indexing of the record will be accomplished automatically. The use of natural language input is still in a more or less research stage. Considerable attention is being given to syntactic and transformational analysis of natural language text input prior to further operations on the input.

For large systems, a more sophisticated query language is required which can retrieve data based upon algebraic, logical, and hierarchical conditions. The query language will generally utilize a natural, language-like form; however, because of the ambiguities of English, many problems exist. The system, and to some extent the

language, will be designed to take advantage of the particular type of random access storage medium utilized in order to minimize accesses to the device.

10.8.2 Predicted Trends

10.8.2.1 Fact Retrieval Systems. It is predicted that there will be a major emphasis on fact retrieval systems over the next several years. This emphasis will include implementation of both large and small-scale systems, as well as research and development on the technology and equipment involved in fact retrieval. As evidence of this, we may cite the Army Chemical Information and Data System, the IBM Information Processing System (IPS) being developed for the Navy Command System Support Activity (NAVCOSSACT), the Reliability Central data system being developed by the AUERBACH Corporation, the fact retrieval systems in the planning stage by Interstate Commerce Commission and the Agency for International Development, the inventory control systems which are in common use. command and control systems, and management information systems.

10.8.2.2 Real-Time Systems. It is predicted that the trend will be toward on-line real-time computer systems for fact retrieval and, in limited cases, for document retrieval (or at least the indexing functions). These systems will involve random access files, communications links, and sophisticated computer software, including executive programs and query or retrieval languages.

BIBLIOGRAPHY

(1) Alexander, S.N., "The Current status of graphic storage techniques: their potential application to library mechanization," Libraries and Automation Washington, D.C., Library of Congress, 1964.

(2) "Automating a California Land Rush," Business Automation 8(9): 27-30 (Sept 1962).

(3) Bagg, T.C.; and Stevens, Mary E., Information selection systems retrieving replica copies: a state-of-the-art report, Washington, U.S. Dept. of Commerce, 1961.

(4) Ballou, H.W., "Guide to micro-reproduction equipment." Annapolis Md., National Microfilm Association, 1959.

(5) Batten, W.E., "Specialized files for patent searching," a chapter in the book entitled: Punched Cards, ed. by R.S. Casey et al., New York, Reinhold, 1958.

(6) Berelson, B., Review of studies in the flow of information among scientists, New York, Bur. of Applied Social Research, Columbia U., 1960.

(7) Berul, L.H. Selecting a system for producing higher quality announcement journals, Wakefield, Mass., Information Dynamics Corp., 1962.

(8) Bourne, C.P., Methods of information handling, New York, Wiley & Sons, 1963.

(9) Chasin, Lawrence I., "Planning, organizing and implementing mechanical systems in a space technology library," Automation and Scientific Communication, American Documentation Institute, 26th Annual Meeting, Chicago, Ill., (p. 303-305) October 1963.

(10) Chasin, Lawrence I.; and Kodroff, Bernard, "The Automation of a document library," Bulletin of the Special Libraries Council of Philadelphia and vicinity, Vol. 28, No. 3; February 1962.

(11) Cleverdon, C.W., Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems, London, England, ASLIB, 1962.

(12) Cutter, C.A., Rules for a dictionary catalogue, 4th ed. Chicago, Ill., Amer. Library Assoc., 1935.

(13) Day, M., "Scientific and technical information program of the National Aeronautics and Space Administration," Proc. Literature of Nuclear Science Conference, spons. by the U.S. Atomic Energy Comm., Div. of Technical Information Extension, Oak Ridge, Tennessee, Sept. 11-13, 1962 (pp. 361-77).

(14) Department of Health Education and Welfare, "The MEDLARS Story at the National Library of Medicine"; Washington, D.C., 1963.

(15) de Solla Price, D.J., Little science, big science, New York, Columbia University Press, 1963.

(16) Dewey, M., Dewey decimal classification and relative index. 16 ed. (2 vols). New York, Forest, 1958.

(17) Eldridge, W. B. ; and Dennis, S. F., "The Computer as a tool for legal research," Law and Contemporary Problems 28:1 (1963).

(18) Fasana, P. J., "Bibliographic encoding: a machine-interpretable natural format for highly structured data," presented at the 26th Annual Meeting of ADI, Chicago, Illinois, Oct. 1963. Published in proceedings entitled, Automation and Scientific Communication, Part 2, p. 108.

(19) Hattery, L. H. ; and McCormick, E. M., eds. Information retrieval management, Detroit, Mich., American Data Processing, 1962.

(20) Houston, N. ; and Wall, E., "The Distribution of term usage in manipulative indexes," Amer. Doc., 15(2):105-14 (April 1964).

(21) Howerton, P. W., Information handling: first principles, Washington, D. C., Spartan Books, 1963.

(22) Information retrieval in action, proceedings of a conference held on April 16-18, 1962, Western Reserve University, Ohio. Cleveland, Ohio, Western Reserve Univ., 1963.

(23) Information systems workshop: the designer's responsibility and his methodology, based on a conference sponsored by ADI and U. of California, Los Angeles, May 29-June 1, 1962. Washington, D. C., Spartan Books, 1962.

(24) Jonker, F., Indexing theory, indexing methods and search devices, New York, Scarecrow Press, 1964.

(25) Kurth, W. H., Survey of the interlibrary loan operation of the National Library of Medicine Washington, D. C., Dept. of HEW, 1962.

(26) Kyle, Barbara., "The UDC--a study of the present position and future developments, with particular attention to those schedules which deal with the humanities, arts and social sciences," UNESCO Bull. for Libraries 15(2):53-70 (1961).

(27) Little, A. D., Inc., Centralization and documentation, (PB 181548) final report to the National Science Foundation, July 1963, Washington, D. C. Government Printing Ofc., 1963.

(28) Luhn, H. P., "The Automatic creation of literature abstracts," IBM Jl of Res & Development 2(2):159-65, 317 (April 1958).

(29) Luhn, H. P., The Automatic derivation of information retrieval encodements from machine-readable texts, New York, IBM, ASDD, 1959.

(30) Luhn, H. P., "Keyword-in-context index for technical literature (KWIC Index)," Amer. Doc. 11(4):288-95 (Oct 1960).

(31) Montague, Barbara A., "Testing, comparison, and evaluation of recall, relevance, and cost of coordinate indexing with links and roles," Proceedings of the American Documentation Institute, 1964 Annual Meeting, Vol. 1: 357-367 Philadelphia, Pa., October, 1964.

(32) Mooers, C.N., "Mooers' Law, or why some retrieval systems are used and others are not," Zator Co., Technical Bull. No. 136 (Dec 1959).

(33) Mooers, C.N., Information retrieval, New York, Scarecrow Press, 1959.

(34) Overmeyer, L., "The Dollars and cents of basic operations in information retrieval," presented at conf. held at Western Reserve U., Cleveland Ohio, April 16-18, 1962. Published in the book entitled: Information Retrieval in Action, Cleveland, Ohio, Western Reserve U., 1963, pp. 199-211.

(35) Raganathan, S.R., "Depth classification, tools for retrieval, and organization for research," Chap. P2 in Documentation and Its Facets, New York, Asia Publishing House, (1963) pp. 604-20.

(36) Rules for descriptive cataloging, Washington, D.C., Library of Congress, 1949.

(37) Salton, G.; and Thorpe, R.W., "An approach to the segmentation problem in speech analysis and language translation," Proc. 1961 Intl Conf. on Machine Translation of Languages & Applied Language Analysis, held in Teddington, England, Sept. 5-8, 1961. London, Her Majesty's Stationery Ofc., 1962.

(38) Shine, T.L., "Savings from automated indexing of deeds," Data Proc. for Management 5(2):38-9 (Feb 1963).

(39) Sinnett, Jefferson D. "An Evaluation of links and roles used in information retrieval" - Master's Thesis - Air Force Institute of Technology, Air University, U.S. Air Force, Wright-Patterson Air Force Base, Ohio, AD 432198, December 1963.

(40) Sparks, D.E.; Berul, L.H.; and Waite, D.P., "Output printing for library automation," Libraries and Automation, Washington, D.C., Library of Congress, 1964.

(41) Strauss, L.J.; Strieby, I.M.; and Brown, A.L., Scientific and technical libraries: their organization and administration, New York, Wiley (Interscience), 1964.

(42) Taube, M.; Gull, C.D.; and Wachtel, Irma. "Unit terms in coordinate indexing," Amer. Documentation 3(4): (1952).

(43) Thesaurus of engineering terms, New York, Engineers' Joint Council, 1964.

(44) US Govt-US Senate 87:1, Interagency coordination of information, hearing before the Subcommittee on Reorganization and International Organizations. Comm. on Govt. Operations. Part 1 (of two parts). Sept 21, 1962. Washington, D.C., Govt Printing Office, 1963.

(45) US Govt-US Senate 86:2, Documentation, indexing and retrieval of scientific information, (Senate Document No. 113), A study of federal and non-federal science information processing and retrieval programs. Washington, D.C., Govt Printing Office, 1961.

(46) Vickery, B.C., Faceted classification: a guide to construction and use of special schemes, London, England, ASLIB, 1960.

(47) Wall, E., Information retrieval thesauri, New York, Engineers' Joint Council, 1962.

(48) Wall, E., The Mechanization of information dissemination, New York, Engineers' Joint Council, 1962.

(49) Wall, E., "The Productive use of engineering information -- the EJC action plan for improving dissemination of engineering information," Annual Mtg. American Soc. for Engineering Education, held in Philadelphia, Pa., June 1963.

APPENDICES

APPENDIX A. SUPPLEMENTAL BIBLIOGRAPHY

(A1) Artandi, S. "A Selective bibliographic survey of automatic indexing methods," Special Libraries 54(10) :630-34 (Dec 1963)

(A2) Baxendale, P. B. "An Empirical model for machine indexing," presented at the Third Inst. of Info. Storage and Retrieval, Feb 13-17, 1961 Published in Machine Indexing: Progress and Problems, Washington, D.C., American Univ., 1961; pp. 207-18.

(A3) Baxendale, P.B. "Machine-made index for technical literature - an experiment," IBM Jl of Res. & Development 2(4):354-60 (Oct 1958).

(A4) Becker J.; and Hayes, R. M. Information storage and retrieval, New York, Wiley, 1963, pp. 130-37.

(A5) Bohnert, L. M. "New role of machines in document retrieval; definition of scope," presented at the Third Inst. of Info., Storage and Retrieval, Feb 13-17, 1961. Published in Machine Indexing: Progress and Problems, Washington, D.C., American Univ., 1961; pp. 8-21.

(A6) Borko, H.; and Bernick, M. "Automatic document classification," Jl of the ACM 10(2):151-62 (Apr 1963).

(A7) Bourne, C.P. Methods of information handling, New York, Wiley, 1963; pp. 151-60.

(A8) Climenson, W. D.; Hardwick, N. H.; and Jacobson, S. N. "Automatic syntax analysis in machine indexing and abstracting," Amer. Doc. 12(3):178-83 (July 1961).

(A9) Cornelius, M. E. "Machine input problems for machine indexing: alternatives and practicalities," presented at the Third Inst. of Info., Storage and Retrieval, Feb 13-17, 1961. Published in Machine Indexing: Progress and Problems, Washington, D.C., American Univ., 1961; pp. 41-9.

(A10) Doyle, L. B. "Indexing and abstracting by association," Amer. Doc. 13(4):378-90 (Oct 1962).

(A11) Doyle, L. B. "Semantic road maps for literature searcher," Jl. ACM 8(4):553-578 (Oct 1961). This theoretical paper describes the associative principles upon which much of the author's subsequent work depended.

(A12) Edmundson, H. P.; and Wyllys, R. E. "Automatic abstracting and indexing: a survey and recommendations," Communs. ACM 4(5):226-34 (May 1961).

(A13) Garfield, E. "Citation indexes for science," Science 122(3159): 108-11 (July 15, 1955)

(A14) Luhn, H. P. "A Statistical approach to mechanized encoding and searching of literary information," IBM Jl of Res. & Development 14(4):309-17 (Oct 1957).

(A15) Luhn, H. P. "Automatic creation of literature abstracts," IBM Jl of Res. & Development 2(2):159-65 (Apr 1958).

(A16) Luhn, H. P. "Auto-encoding of documents for information retrieval systems," presented at the Symposium on Documentation, Univ. of Southern Calif. - School of Library Science, Apr 9-11, 1958. IBM Res. Center, Yorktown Heights, N. Y., 6pp.

(A17) Luhn, H. P. "Keyword-in-context index for technical literature (KWIC Index)," presented at The Sixteenth Meeting of the Amer. Chemical Soc., Atlantic City, N. J., Sept 14, 1959; 4pp.

(A18) Luhn, H. P. "Machinable bibliographic records as a tool for improving communication of scientific information," IBM Report No. 225-1478, presented at the Tenth Pacific Scientific Congress, Honolulu, Aug 21 - Sept 6, 1961, White Plains, N. Y., pp. 1-14.

(A19) Luhn, H. P. "Potentialities of auto-encoding of scientific literature," IBM Res. Report RC-101, Yorktown Heights, N. Y., May 15, 1959; 22pp.

(A20) Maizelli, R. E. "Value of titles for indexing purposes," Revue de la Doc. 27:126-27 (Aug 1960)

(A21) Maron, M. E. "Automatic indexing: an experimental inquiry," Jl of ACM 8(3): 404-17 (July 1961).

(A22) Maron, M. E.; and Kuhns, J. L. "On Relevance, probabilistic indexing and information and retrieval," Jl of ACM 7(3):216-44 (July 1960).

(A23) Montgomery, C.; and Swanson, D. R. "Machinelike indexing by people," Amer. Doc. 13(4):359-66 (Oct 1962).

(A24) O'Connor, J. "Some remarks on mechanized indexing and some small-scale empirical results," presented at the Third Inst. of Info., Storage and Retrieval. Feb 13-17, 1961. Published in Machine Indexing: Progress and Problems, Washington, D. C., American Univ., 1961; pp. 266-79.

(A25) O'Connor, J. "Some suggested mechanized indexing investigations which require no machines," Amer. Doc. 12(3):198-203 (July 1961).

(A26) Oswald, V. A. "Automatic indexing and abstracting of the contents of documents," Report No. RADC-TR-59-208, prepared for US Air Force-ARDC, pp. 5-34, 59-113.

(A27) Oswald, V. A. "Interim report, the automatic extraction and display of the content of documents," Report No. PRC-R-91, Planning Res. Corp., Los Angeles, Cal., Mar 15, 1959.

(A28) Rath, G. J.; Resnick, A.; and Savage, T. R. "Comparison of four types of textual indicators of contents," Amer. Doc. 12(2):126-30 (Apr 1961).

(A29) Rath, G. J.; Resnick, A.; and Savage, T. R. "The Formation of abstracts by the selection of sentences. Part I - Sentence selection by men and machines," Amer. Doc. 12(2): 141-43 (Apr 1961).

(A30) Resnick, A. "The Formation of abstracts by the selection of sentences. Part II - The reliability of people in selecting sentences," Amer. Doc. 12(2): 141-43 (Apr 1961).

(A31) Resnick, A.; and Savage, T. R. "A Re-evaluation of machine generated abstracts," IBM Res. Report No. RC-230, Yorktown Heights, N. Y., March 1, 1960.

(A32) Salton, G. "Associative document retrieval techniques using bibliographic information," Jl of ACM 10(4):440-57 (Oct 1963).

(A33) Savage, T.R. "The preparation of auto-abstracts on the IBM 704 data processing system," IBM Res. Center, Yorktown Heights, N.Y., Nov 1958.

(A34) Schultz, C.K.; and Schwartz, P.A. "A Generalized computer method for index production, Amer. Doc. 13(4):420-32 (Oct 1962).

(A35) Swanson, D.R. "Searching natural language text by computer," Science 132(3434): 1099-1104 (Oct 21, 1960).

APPENDIX B. RANDOM ACCESS FILE STRUCTURES

B.1 INTRODUCTION

The advent of relatively low-cost digital random access devices has increased the economic feasibility of retaining large stores of information in machine-readable form, ready for rapid use by the central processor Such cost advantages are causing increasing attention to be paid to random access files, not only for data processing systems, but also for information systems involving short factual records and even documents. The trend toward random access for large files is fraught with difficulty because the larger the file becomes, the more cumbersome it is to handle efficiently. More sophisticated programs are required to maintain and address questions to it. More complex techniques are required to structure the file and establish the proper relationship between its index and its content.

Appendix B describes some of the problems involved in handling random access files, and outlines a number of methods available for structuring the files. The method chosen for any particular system problem would depend entirely on the requirements and characteristics of the system.

B.2 THE PROBLEM

There are three data processing functions which are of interest when using a large scale data storage/retrieval device: the loading of, the retrieval from, and the maintenance of a large data base. By "loading " the data base is meant the initial insertion of the bulk of the data into the file. "Retrieval" refers to the locating and withdrawing of a selected segment of the data from the loaded file. "Maintenance" involves the addition and deletion of data and the local reorganization of the data to enhance retrieval, if such a reorganization is required. When the storage/retrieval equipment is random access, i e . has the capability that data can be stored anywhere and retrieved directly without search, these three functions are so strongly interdependent that it is impossible to consider any one of them without taking into account the other two. For example, the

easiest way to load a file into a storage/retrieval device is to start at the "beginning" of the memory and continue in a serial fashion until the loading is completed, without regard to the specific allocation of storage. In the case of random access equipment, this technique has obvious defects with respect to retrieval on the file and, in particular, fails to take advantage of the fact that the file can be "randomly" accessed. One could say, therefore, load the file in pieces and keep a record of which pieces contain what. Carried to its limit, this indexing technique provides by far the most accurate retrieval capability, but even ignoring the difficulty of creating and loading a complex index, maintenance of a large indexed file even with a random access device can be far too difficult for practical applications. With respect to the problems of maintenance, perhaps the simplest file to maintain is one in which the file entries are stored serially and no entry appears more than once, i.e., there is no cross referencing. However, we recognize that we once more have the situation that retrieval on such a file is far too expensive involving serial, i.e., non-random search.

This then is the problem: one wants the same system to provide ease of loading, efficient retrieval, and a workable maintenance capability. Once satisfied that these criteria are being met, one can consider other dimensions of interest, such as taking advantage of the usage statistics of the items in the file and so forth.

The following paragraphs are concerned with two problem areas important to the topics previously discussed, namely, the problems of address generation ("randomizing") and file structuring.

B.3 ADDRESS GENERATION

When using a random access auxiliary memory coupled to a data processing system, the first technique which comes to mind for accessing the data stored in the auxiliary memory is to construct some sort of index to that data during the memory loading phase. Indeed, if it is not known at retrieval time exactly what data in the file is specifically required, some sort of index or equivalent device might well be mandatory. A classic example of this kind of retrieval system is the typical book library, where the stacks are the random access memory, the staff of the library constitute the data processing system, and the various card catalogues, etc., are the indexes generated by the data processor. For the kind of operation

of interest here, however, where something about a particular piece of data is specifically known, there is little utility in buffering the data from the user of the data by so elaborate a device as a highly structured index. Rather we would want what we already know about the data to lead us by as direct a route as possible to the actual data itself. That is, we prefer to give the librarian a specific request and have him provide to us the specific address in the stacks of the data we want. Although perhaps obvious to some, it is important to recognize the distinctions between these two types of retrieval, for the latter suggests that the well-known idea of a direct data to address conversion and the problems associated with this technique must be considered.

In a computer oriented system, the usual notion is that the data to be retrieved has been structured into highly formatted "records" to each of which has been assigned a unique "tag" or ID number. In general, it is not possible to control the assignment of these tags for the purposes of subsequent retrieval from a specific random access device. Hence, conversion algorithms must be invented which generate addresses from the tags. Since such generated addresses usually are intended to be quite specific, a detailed analysis of the source tags is required in order to determine the most effective conversion algorithm for any set of tags. Typically, such algorithms convert alphanumeric tags to pure numeric and then operate on the numeric to reduce it to a string of digits acceptable as an address by the storage/retrieval device being used. Such operations as the reduction of selected portions of the numeric by modulo arithmetic, perhaps after "randomizing" them by various well-known procedures, are completely employed.

However, it can be stated that, in general, no universal method for generating random addresses from any given set of source numbers is known. Furthermore, it can be demonstrated that, given a uniform distribution of source numbers, no algorithm which produces random addresses can produce uniformly distributed addresses. All of this is further complicated by the fact that any address generation algorithm will not produce unique addresses, a situation which worsens as the number of addresses to be computed increases.

B.4 FILE STRUCTURING

Because address computation produces neither uniformly distributed nor unique addresses, the manner in which a data file is assembled is of the utmost importance with regard to efficient utilization of available space, retrieval, and maintenance of the file. Actually it is the extent to which the systems designer is able to mold the structure of a file, given the storage/retrieval device and the means for inserting the data into it, which ultimately determines the efficiency of the final system. Several techniques for organizing the files have been discussed in the literature for the special case of storage and retrieval using random access memories. A representative sample of those techniques is presented. The mode of presentation will be algorithmic, that is, a step-by-step procedure for loading, retrieval, and maintenance for each technique will be given. A comment on each technique precedes its algorithm.

B.4.1 The Directory Method

This technique is effective primarily when a sequentially ordered file is loaded onto a random access device, and no further file maintenance is required (see Table B-1). In these rare instances, a directory is extremely efficient because it requires almost no file maintenance of its own, and it can be made as general or as specific as desired. If it is general, its size is reduced and it refers to a range of addresses for access, thereby increasing the response time. If the directory is specific, its storage requirement is large, but it refers to a specific address in answer to a query, thereby reducing response time.

(1) Loading

(a) Load the input file directly into the random access device in sequential order.

(b) Create in the directory an address entry for each addressable segment of the random access device (usually a track) which has been loaded.

(c) As the primary key of each entry in the directory, post the "tag" or sequence identifier which begins that addressable segment.

TABLE B-1. COMPARISON OF FILE ORGANIZATION METHODS

Method		File Order	Additional Storage Requirements	File Maintenance Requirements	Response Time	Retrieval Capability	Program Costs
(1) Directory		File must be ordered in strict sequence, with no updating allowed.	Low (The directory itself takes up little room)	Low (New records are simply added to the end of the old file)	Very Fast	As many keys as form the sequence.	Low
(2) Open		File must be ordered by a single key.	None	Moderate	Slow	One key	Low
(3) Minimal Search Open		File must be ordered by a single key.	None	High (Because of record movement)	Moderately Fast	One key	Low
(4) Direct Chaining		File is scattered in a controlled manner.	Moderate (1 address field x number of records)	Moderate	Unpredictable (Depends on chain length)	One key	Moderate
(5) Tag-address List		File is scattered in a controlled manner.	High (1 tag field and 1 address field x number of records)	Moderate (The Tag-addresses must be maintained)	Fast (Assures one access in most instances)	One key	Moderate
Multi-List:	(6) Indirect Chaining (List Structure)	File may be ordered in any manner.	Very High (2 address fields x number of records x number of list-structures)	Moderate (The links must be preserved).	Slow (But able to answer complex queries)	As many keys as list-structures.	High
	(7) Inverted Lists	File may be ordered in any manner.	High (1 address field x number of records x number of lists + directories for entry to lists)	Very High (The lists are maintained separately)	Slow (But able to answer complex queries)	Boolean logic, as many keys as lists.	High

(2) Retrieval

(a) Search the directory for the "tag" which is equal to, or most closely less than, the "tag" in the query.

(b) Read the addressable segment pointed to by the directory address.

(c) Compare tags sequentially until the desired record is found.

(d) If a higher tag is encountered, the record is not in the file.

(3) Maintenance

(not permitted)

B.4.2. The "Open" Method

This technique (see Table B-1), while perhaps the simplest and most straightforward, has obvious defects with regard to the retrieval function (and in this case, as a consequence, the maintenance function as well). Conceivably, if a transaction specified the tag of a record which was not in the file, almost the entire device might be serially searched before the situation was recognized. However, because of its simplicity, it could be reasonably employed in a low volume situation or perhaps with small subfiles in conjunction with an index to the subfiles.

(1) Loading

(a) Clear device to "unoccupied" symbols.

(b) Generate all unique addresses from record "tags"

(c) Load one record at each computed address.

(d) For each remaining record:

(i) Compute its address.

(ii) Go to that address and search serially, preferably in a single direction for the next "unoccupied" location.

(iii) Load the record into the location.

(e) Repeat step (d) until all records are loaded.

(2) Retrieval

(a) Compute address from tag given in transaction request.

(b) Read record contained at the address and compare tags.

(c) If tags are not the same, search serially from that address or until desired record is retrieved.

(3) Maintenance

(a) To add a record:

(i) Execute step (d) of loading procedure.

(b) To delete a record:

(i) Execute steps (a), (b), and (c) of retrieval procedure.

(ii) Clear location where record found "unoccupied" symbols.

(c) Undating for changing priorities:

(i) File must be reloaded.

B.4.3 Minimal Search Open Method

This technique improves the retrieval efficiency of the simpler Open Method in that it tends to minimize the serial searching inherent in that method (see Table B-1). A price is paid for this increase in efficiency in the more cumbersome loading and maintenance procedures where large segments of stored data have to be searched and shifted around. This technique, in general, is applicable in the same situations as the Open Method.

(1) Loading

(a) Clear device to "unoccupied" symbols.

(b) Generate all unique addresses from record "tags."

(c) Load one record at each of the computed addresses.

(d) For each remaining record:

(i) Compute its address.

(ii) Go to the next higher address in sequence. If this location is unoccupied, load the record into it.

(iii) If this location is occupied, compare the computed address of the record stored there with the computed address of the record to be loaded. If the former is less than or equal to the latter, repeat steps (ii) and (iii) under (d).

(iv) If the former is greater than the latter, shift it and all records between it and the next empty location up in the memory. Load the new record into the location vacated by the old record.

(e) Repeat step (d)until all records are loaded.

(2) Retrieval

(a) Retrieval procedure is the same as with the Open Method.

(3) Maintenance

(a) To add a record:

(i) Execute step (d) of loading procedure.

(b) To delete a record:

(i) Execute retrieval procedure.

(ii) Search serially up in the memory until the first record with a higher computed address than the computed address of the record to be deleted is found.

(iii) Shift all records between the former and the latter down in the memory.

(c) Updating for changing priorities:

(i) File must be reloaded.

B 4.4 The Direct Chaining Method

The chaining technique is probably the most popular and widely used file structuring method for random access devices (see Table B-1). Part of the reason for this is historical. Its chief disadvantage is its sensitivity to the address generation scheme with regard to chain lengths and the consequently unpredictable access times which can cause trouble in a critically timed program. The chaining technique, however, is vastly superior to the Open Method, in particular because only material of potential interest is retrieved during operations on the file. Furthermore, a certain amount of control over the construction of chains is possible in some instances, allowing the programmer to minimize to some extent, delays due to equipment characteristics such as latency, select times, etc.

(1) Loading

(a) Clear device to "unoccupied" symbols.

(b) Reserve a field large enough to contain an address in each record. Put a special symbol in this field in every record. This field is called a "link field," the special symbol is called a "terminator."

(c) Generate all unique addresses from record "tags."

(d) Load one record at each of the computed addresses.

(e) For each remaining record:

(i) Compute its address.

(ii) Assign the record to an unoccupied location.

(iii) Go to the computed address and check the link field of the record stored there. If it contains a terminator, put the address assigned to the record to be loaded in the link field of the record at the computed address.

(iv) If it contains an address, go to the address and repeat the process until a terminator is found in the link field of some record. Replace that terminator with the address assigned to the record to be loaded. (The list of connected records which results from this procedure is a "chain.")

(f) Repeat step (e) until all records are loaded.

(2) Retrieval

(a) Compute address from tag given in transaction request.

(b) Read record contained at that address and compare tags.

(c) If tags are not the same, check link field for terminator. If found, requested record is not in the file.

(d) If terminator is not found, go to the address specified in the link field and repeat the process until either the desired record or a terminator is found.

(3) Maintenance

(a) To add a record.

(i) Same procedure as step (d) of loading.

(b) To delete a record:

(i) Follow steps (a), (b), and (c) of the retrieval procedure.

(ii) When the record is found, put the data contained in its link field into the link field of the record preceding it in the chain

(iii) Clear the location where the record was stored to unoccupied symbols.

(c) To update for changing priorities:

(i) Retrieve the entire chain of interest.

(ii) Rearrange the links of the chain so that the high usage records are the closest in the chain to the computed address of that chain.

(iii) Reload the records according to the revised address list.

B 4.5 The Tag-Address List Method

The file structuring technique to be described (see Table B-1) emphasizes control over the placement of records in the file. The particular discussion is directed specifically toward facilitating file maintenance. A different distribution of records, it can be readily seen, can be contrived which will facilitate retrieval (in the current context, this means minimizing disc latency). The merits of this tag-address list technique have not been fully determined as far as experience in using it is concerned. It would seem, however, that the control afforded the programmer in the structuring of a file is an important feature. In particular, though data is stored in desired locations, retrieval of the data is still random in the sense used here. Furthermore, it would seem that the structuring of the file would not be as sensitive to the particular address calculation scheme used as in other techniques. Finally, given only a reasonable distribution of addresses, retrieval is guaranteed in exactly two accesses, only one of which would involve the more costly track select operation. The addressing of the device assumes a hierarchy of "sector" as the most generic, then "zone," and finally "track" as the most specific.

(i) Loading

(a) Compute all unique sector-zone (not track) addresses.

(b) Beginning with the first track at a computer sector zone address, load tags, and tags only, from the records having that computed address, each together with an address assigned as follows:

(i) Start with the last track of the first "free" zone on the relevant sector (where a "free" zone is not a computed address) and assign sector, zone, and track. Distribute these addresses uniformly over the free zones, leaving, say, the first track in each for possible future storage of tag addresses, if such room is available.

(ii) If all free zones get assigned, start with the last tracks of the zones containing tag-addresses and distribute the assigned address among these

zones so that some room is left in each for possible future storage of tag-addresses, if such room is available.

(iii) If a sector gets fully packed, set sector number to an auxiliary sector (previously reserved for this purpose).

(iv) If a tag-address zone gets filled entirely with tag-addresses, the address computation is no good.

(v) Terminate each tag-address list with a terminating symbol followed by a count of the available characters left in the zone.

(vi) For each sector, store the next assignable address. (Since the record addresses are assigned in this method, the "next" assignable address should always be known).

(vii) Load the records.

(2) Retrieval

(a) Compute sector number and select if necessary. Compute zone.

(b) Initiate read of tag-addresses.

(c) Compare for terminating symbol; compare for specified tag. If tag, read record. If terminating symbol, indicate "record not in file."

In general, at most one sector select will retrieve any record, unless an auxiliary sector has had to be employed.

(3) Maintenance

(a) To delete a record:

(i) Compute sector, select, compute zone.

(ii) Initiate read of tag-addresses.

(iii) Compare for tag. When found, clear tag to "unoccupied" symbols to indicate available "empty" location. Leave address alone.

(iv) Compute the track number of the track which contained the deleted tag (this can be computed from the number of characters read in up to the tag).

(v) Rewrite track.

(b) To add a record:

(i) Compute sector, select, compute zone.

(ii) Check "next available address" for the computed sector (see loading, step (vi).

(iii) If room available on sector, initiate simultaneous read of tag-address zone.

(iv) Compare for terminating symbol. When found, check for available characters in zone.

(v) If room available in zone, add new tag-address, re-insert terminating symbol and new count of available characters.

(vi) (Same as last two steps under "delete")

(vii) Write the record to the assigned locations.

(viii) Update the "next available address" for the sector.

(ix) If no more room available on sector, initiate simultaneous read of tag-address. Search list for address of "empty" location.

(x) If unoccupied locations found, update list and write record to empty locations.

(xi) If no unoccupied location found, set sector number to auxiliary sector, update list and write record.

(xii) If no more room available in zone because filled with tag-addresses, re-evaluate addressing functions.

(xiii) If no more room available in zone because records are stored in that zone, the first record in the zone must be relocated. This is done by computing its tag-address list address, finding the tag in the list, changing the associated address and loading the record into the new address. The new record is then loaded in the normal fashion.

(c) Updating for changing priorities:

(i) Store a bit in address to indicate priority.

(ii) Sort tag-address lists when required.

B.4.6 The Indirect Chaining Method

The technique of indirect chaining, sometimes called the List Structure Method, allows the same file to be structured in as many ways as there are retrieval keys of interest in the records of the file (see Table B-1). This adds new dimensions to the retrieval concept and can prove very powerful in less straightforward retrieval systems, where the same file may have a different significance in different retrieval situations. For example, a file may be pre-sorted for efficient report writing and yet still be structured to allow a good degree of random access to its contents. The main disadvantage, of course, is the room taken up for the head and link fields in the records. On the other hand, many multiple file systems carry along a great deal of redundant data in the records of the respective files in order to facilitate complex retrieval demands. Hence, the consolidation of records from various files and the secondary structuring of the consolidated file may actually save space. As a final comment, it is to be noted that when a record is retrieved from a multi-structured file, the addresses of other records of potential interest are immediately available in the secondary head and link fields of that record. This leads to the notion of so-called "list structures" which has proven very powerful in many applications.

(1) Loading

(a) Load the file in any desired fashion after first reserving in each record two fields capable of storing addresses. These fields are called the "head" field and the "link" field.

(b) Compute a new set of addresses for the stored records, using a secondary key different from that which was used to load the file.

(c) Form lists of those records which have the same computed secondary address. The first record in each such list is called the "head" of the list. (A list can consist of a head only.)

(d) In general, there will already be a record stored at a computed secondary address. In the head field of that record, store the primary address of the record which is the head of the list.

(e) If there is no record stored at the computed secondary address, load a dummy record containing the primary address of the record which is the head of the list.

(f) In the link field of the list head, store the actual address of the next record in the list. Continue this process with all the records in the list, thus forming a chain. Put a terminator in the link field of the last record in the chain.

(g) Perform these operations for all records in the file.

(2) Retrieval (on the secondary key)

(a) Compute address from (secondary) key.

(b) Use the address stored in the head field of the record located at the computed address to find the first record in the secondary chain.

(c) Proceed as in retrieval described for ordinary chaining using the secondary links.

(3) Maintenance (of the secondary structure)

(a) When a record is added to the primary structure:

(i) Compute address from (secondary) key.

(ii) At the computed address is a record, which, in general, contains in its head field the address of the head of a secondary list. Replace that address with the primary address of the added record.

(iii) Store the address which was formerly the list head in the link field of the added record (thereby making the new record the head of the secondary list).

Note: If a record is to be added to the primary structure and a dummy secondary record (see step (e) under loading) is already stored where the new record is to go, simply put the address contained in the head field of the dummy record in the head field of the new record and load the record, overlaying the dummy record.

(b) When a record is deleted from the primary structure:

(i) If the deleted record contains a secondary list head address, retain this address in a dummy record stored at the location of the deleted record.

(ii) If the deleted record contains data in the secondary link field, update the secondary chain according to the procedures given under maintenance for the Chaining Method.

Note: A record can be deleted from a multi-structured file only if it is no longer required in any of the structures, primary and secondary.

(c) When the primary file is updated for changing priorities:

(i) In general, the procedure described under "loading" will have to be repeated.

B.4.7 Inverted Lists

The AUERBACH Corporation has developed a technique, under the guidance of Dr. Jack Minker, which retrieves the appropriate records from random access with a minimum number of accesses, regardless of the logical complexity of the retrieval request. The technique, called the Inverted List Method, is designed for a retrieval system where response time is at a premium and where each query requests records which contain a number of attributes, each with a specific value. The volume of queries in such a system is normally high, and the price of increased file maintenance time must be an acceptable constraint.

The ideal toward which most retrieval systems strive is to retrieve from auxiliary memory only those records which precisely match the search criteria, and to do this in no more than one access per record. The Indirect Chaining (or List Structure) Method described above cannot do this because it must choose a single search attribute and retrieve all records in the file containing the attribute, comparing the records against the other search attributes. Even if the Indirect Chaining Method selects the attribute associated with the least number of records, a great many more records than are needed

are drawn from storage. The Inverted List Method attempts to reduce the number of accesses in the following manner:

(1) Loading

(a) Load the file in any order and at any address desired.

(b) As each record is loaded, extract from it the values of the attributes which will be used for searches at a later time.

(c) Place the address of the record within the list for each appropriate value by sequence according to address.

(d) Increase by 1 the counter at the head of each value list every time an address is added.

(e) When the file has been loaded, arrange the value lists in order within attribute.

(f) Store the value lists in random access memory, often preparing a value index of where they can be found.

(g) Store the value indexes by attribute in random access memory, after preparing an attribute directory of where they can be found. Include in the attribute directory the total number of records addressed by the value lists associated with that attribute.

(2) Retrieval

(a) Convert the Boolean expression of the query into Polish Notation.

(b) Using the attribute directory and the value indexes, read from storage one-by-one the value lists which pertain to the query, reading the lists with the least number of addresses first.

(c) If the Boolean relationship between two lists is a union, merge the two lists.

(d) If the Boolean relationship between two lists is an intersection, compare the addresses of the two lists and retain only matched addresses.

(e) Continue the process for all pertinent lists and read from storage only those records whose addresses are retained.

(3) Maintenance

(a) To delete a record:

(i) Read the record from storage.

(ii) Using the attribute directory and the value index, read each pertinent value list from storage.

(iii) Delete the record address from each pertinent value list and diminish its counter by 1 before replacing the list.

(iv) Diminish by 1 the attribute directory counter.

(b) To add a record:

(i) Write the record in any convenient address.

(ii) Add the address of the record to each value list which is relevant to the values in the record.

(iii) Augment the attribute directory and the value indexes by 1.

APPENDIX C. GLOSSARY

Abstract, n., a brief summary of the contents of a document. See also extract.

Abstract, descriptive, n., see abstract, indicative.

Abstract, indicative, n., an abstract which merely states in general terms what kind of information is contained in the document for the purpose of conveying an idea of the scope only. See also abstract, informative.

Abstract, informative, n., an abstract which comprehensively describes the principal factual conclusions contained in the document in order to provide information directly significant for the users' purposes. See also abstract, indicative.

Access, v.t., to communicate with a store in an information system for the purpose of either using or storing information in it. n., act of accessing.

Access, random, n., direct access of locations in a store without regard to any order, and not by way of any other locations in the store.

Access, sequential, n., access of locations in a store through a prescribed sequence. Pure sequential access would involve reading only one bit or code element at a time. In some systems, the sequence is a sequence of codes within which there is parallel access to individual code elements; e.g., bits on an eight-channel tape.

Accession number, n., an identifying number, usually in ascending sequence, assigned to each document entering an information system as part of the surrogation and cataloging function.

Address, n., a set of symbols which uniquely identifies a particular location in a store, often used to identify the location of stored data.

Announcement journal, n., a secondary source journal, containing abstracts, titles, indexes or a combination of all three, which is published as currently as possible with the primary source in order to provide a current awareness service to its readers.

Authority list, n., a relatively simple alphabetic list of descriptions of ideas or viewpoints that are likely to be found in literature being indexed together with acceptable index terms for each. See also thesaurus.

Batten card, n., an interior-punched card which can be matched against similar cards so that the locations of the cards which have holes in common can be visually determined.

Bibliography, n., a published list of references or citations, each of which is relevant to a predetermined topic.

Boolean statement, n., a symbolic statement in the algebra of George Boole which expresses logical connections between classes (in the form of AND, OR and NOT). The statement may be used to formulate an inquiry for the purpose of searching a file.

Cataloging, v.t., the part of the surrogation function which involves the assignment of permanent accession and/or class numbers; the recording of titles, authors, and sources; the control of the source name vocabulary; and other initial processing of a document.

Categories, n., see indexing criteria.

Character, n., see code character.

Citation, n., the description of a prior document which relates to the document being written by the author; usually in the form of author, title, journal, volume, number, date and pagination.

Citation index, n., a listing arranged by author giving the items which have cited the author's work.

Class, n., the group of all things which can be described by an index term.

Classification, n., a system of interrelated classes, arranged in an apparently natural or arbitrary order in a chain or lattice, so that each class includes or is included in another class, or both.

Code character, n., a configuration of code elements, which together represent a symbol such as a number or letter of the alphabet, used in a code.

Code element, n., a discriminable phenomenon, such as a hole or notch in a punched card or a magnetic spot on tape, which may be used as a component of a code character.

Coding, direct, n., a kind of coding in which each code character is composed of a specified arrangement of code elements and in which the code elements are not superimposed to represent simultaneously more than one code character; or in which each code word is composed of a specified arrangement of code characters and in which the code characters are not superimposed to represent simultaneously more than one code word.

Coding field, n., an interval of time or space reserved for coding elements in which their presence and order are distinguishable and can be kept constant. The coding interval can be temporal, such as the interval during which bits are being transmitted over a wire, or spatial, such as the interval in which bits are recorded on magnetic tape.

Coding, fixed field, n., a kind of coding in which parts of a coding field are dedicated for particular classes of codes, which, if present, are recorded in the dedicated position.

Coding, free field, n., a kind of coding in which codes may be entered in any part of the coding field, although sometimes the order of recording the classes of codes and how they are to be justified may be specified.

Coding, generic, n., a kind of coding in which code characters are used in dedicated parts of a coding field to indicate hierarchical order and the relation of inclusion among classes.

Coding, superimposed, n., a kind of coding in which code elements are combined to represent simultaneously more than one code character; or in which code characters are combined to represent simultaneously more than one code word.

Collate, v.t., to create a file by interfiling the file elements, such as cards, or records; based on specified relationships among the elements. For example, file elements may be interfiled according to the numeric sequence of a particular code.

Combination, n., any of the various groupings into which a specified number of distinguishable objects may be arranged without regard to order. For example, the dual combinations of A, B, and C are AB, AC, and BC. A combination is contrasted with a permutation in which the order of the objects is regarded.

Communication, n., a two-way exchange of information (as opposed to transmission which is one-way).

Complement, n., <u>see</u> logical complement.

Completeness, n., the number of documents retrieved during a search which contain even a slight amount of relevance to the originating query.

Constraint, n., any condition which limits the operation of a particular system.

Coordinate indexing, n., a method of analyzing and describing items of information so that retrieval is performed by the logical operations of the product, sum and complement on the codes in the store.

Correlation, n., instructive information which results from the compilation, analysis and evaluation of other information for the purpose of answering specific questions within a specific range of inquiry.

Criteria, n., see indexing criteria.

Critical review, n., a comprehensive report by an expert in a field of learning which analyzes and evaluates a specific document with reference to the field.

Current awareness, n., the process of presenting recently published information in such a way as to draw the attention of certain interest groups to information relevant to the interest; emphasis is on rapid publication and on the use of a selected, rather than an exhaustive, list of documents.

Data, n., material which is easily quantifiable, non-abstract, and which can be formatted.

Descriptor, n., a word or group of words which characterize, define, or determine a general set or class.

Descriptive abstract, n., see abstract, indicative.

Dictionary, n., an alphabetic list of words with generally accepted, explicit definitions and sometimes with etymologies and pronunciations. In information storage and retrieval usage, dictionary is sometimes a near synonym for authority list. See also thesaurus, authority list.

Direct coding, n., see coding, direct.

Dissemination, n., the distribution of documents or facts to the people who are interested in such information.

Document retrieval, n., the function performed by an IS&R system which provides as an output one or more documents or document surrogates which may be relevant to a request.

Edge-notched card, n., a card having coding area around its perimeter into which holes are punched or notched to record coded data. An edge-notched card is contrasted with an interior-coded punched card.

Extract, n., a brief statement of the contents of a document in the form of direct quotations from the document. See also abstract.

Fact retrieval, n., the function performed by an IS&R system which provides as output specific answers to inquiries, such as the name of a person who has certain characteristics, rather than a document or a document surrogate.

File, n., see store.

File, inverted, n., a file in which the stored items are grouped by each index word that describes the item, and the groups are arranged in order by the index terms. In such a file, the items indexed by more than one index term are reproduced so that they can each be included in several groups, each group of items represented by one of the index terms used to describe the items in the group. See also file, linear.

File, linear, n., a file in which the index terms are grouped by the item they describe, and the groups are arranged in order by the items. See also file, inverted.

Fixed field coding, n., see coding, fixed field.

Free field coding, n., see coding, free field.

Free indexing, n., the intellectual process of indexing documents whereby the index terms are chosen from words in the document rather than from a controlled list of authorized words.

Generic coding, n., see coding, generic.

Guide, n., see access guide.

Hierarchy of classes, n., the relation of inclusion between classes, in which each higher class includes lower classes and each lower class is included in a higher class.

Index, v.t., to select on the basis of content analysis and to record one or more terms in natural language or code to describe the data contained in a document or part of a document. n., the record of documents or parts of documents and the terms selected to describe the data contained in each document or part of document. adj., pertaining to an index.

Indexing, coordinate, n., see coordinate indexing.

Indexing criteria, n., words of a higher level of generality than most of the words in an access guide; as such, these words serve as names of categories into which the other words may be grouped.

Index, integrated, n., an index in which the index terms are recorded on a storage medium in physical proximity to the described data or copy of the described document.

Index, permuted, n., see permuted index.

Index, separate, n., an index in which the index terms are recorded on a storage medium, which is physically remote from the described data or copy of the described document. The index terms are related to a surrogate sometimes only an accession number in the remote location.

Index term, n., a word in natural language or code which described a document or part of a document and which refers to something which is referred to in the document or part of document.

Indicative abstract, n., see abstract, indicative.

Information, n., material which is conceptual, descriptive, often judgmental, usually narrative and which is not easily quantifiable or formattable.

Information, digital, n., any coded representation which can be processed by machines without first requiring a transformation into machine language.

Information, graphic, n., information which may be stored in visual form, such as in books, on microfilm, etc.

Information storage & retrieval, n., a term generic to all variations of the problems of storing, locating and selecting information of any kind, whether it is in graphic or digital form and whether the desired output is a document or a specific fact.

Informative abstract, see abstract, informative.

Input, n., whatever is entered into a system; specifically, in an information system, data or documents to be stored, the index terms to describe them, or search questions to retrieve them. v.t., to enter into a system.

Integrated index, n., see index, integrated.

Interior punched cards, n., a card used for searching which has coded data punched within its interior, as contrasted with an edge-notched card.

Inverted file, n., see file, inverted.

Item, n., the unit of data, such as a document, abstract, bibliographic citation, or accession number, which is stored in an information system for possible future retrieval.

Key word, n., a word, chosen from a document text or title, which is used in free indexing to describe the content of the document.

KWIC index, see permuted index.

Linear file, n., see file, linear.

Link, n., a number used to associate index terms which relate to the same topic within a document, thereby avoiding erroneous logical intersections of indexes during searches.

Literature, separate, n., individual items, such as books, reports and reference volumes, which usually contain information of lasting value.

Literature, serial, n., collections of articles published in journals which are generally of a partial, ephemeral or interim nature.

Logic, n., 1. the intellectual order of a system as contrasted with the physical form in which the intellectual system is applied. 2. a way of reasoning.

Logical complement, n., the class which has as members all elements except those that are members of a specified class. The logical complement of class A is expressed as $\bar{A}$.

Logical product, n., the class which has as members all elements which are members of all specified classes. The logical product of classes A, B, ... N is expressed as $A \cap B \cap \ldots \cap N$.

Logical sum, n., the class which has as members all elements which are members of any specified classes. The logical sum of classes, A, B, ... N is expressed as $A \cup B \cup \ldots \cup N$.

Meaning, n., is, in a narrow sense in information storage and retrieval, the relation of formal equivalence between symbols, which relation implies substitutability of equivalent symbols and nonsubstitutability of non-equivalent symbols. In a broad sense, see semantics.

Memory, n., a physical structure in which data can be recorded.

Microfiche, n., any transparent photographic film in flat form containing multiple negative or positive replica microreproductions of graphic records arranged in a grid pattern by rows. Microfiche may be any of the following: unit, chip, sheet, jacket, slide, or aperture.

Microfilm, n., any transparent photographic film in roll, strip or scroll form containing negative or positive microreproductions of graphic records. Microfilm is usually in reel form as distinguished from microfiche, which is usually in sheet form. v.t., to make microphotographs.

Microform, n., a generic term referring to any miniaturized form containing micro images.

Noise, n., from the viewpoint of an information system, retrieved data which is not relevant to the purpose of the user of the system; from the viewpoint of the user of the system, retrieved data which is not pertinent to the user's purpose; trash. See also redundancy.

Order, n., any fixed temporal or spatial sequence, frequently conforming to a conventionally accepted sequence of symbols such as numbers or alphabetic letters.

Output, n., whatever is produced by a system; specifically in an information system, data in any form provided by a system to a user in response to his question. See also selection.

Peek-a-boo cards, n., an inverted file of interior-punched cards, where each card represents an index term, and where the document numbers are coded in the form of punched holes in the card, such that logical intersections may be observed visually.

Permutation, n., any of the various ordered groupings into which a specified number of distinguishable objects may be arranged. For example, the dual permutations of A, B and C are: AB, BA, AC, CA, BC, and CB. A permutation is contrasted with a combination in which the order of objects is disregarded.

Permuted index, n., an index in which key terms are selected from the title of a document and are displayed in the index in exact context with other words in the title. An index entry is made for each key word in the title.

Pertinent, adj., refers to retrieved information which is useful to the user for his purposes. See also relevant.

Product, n., see logical product.

Punched card, interior-coded, n., a card, having a coding area within its body, into which holes are punched to record data. An interior-coded punched card is contrasted with an edge-notched card.

Query, see question.

Question, n., a set of terms in natural language or code, related according to a specified search strategy, which is compared with the index terms in the store during a search.

Random access, n., see access. random.

Recall, see completeness.

Redundancy, n., the quantity of transmitted information in excess of the necessary minimum, the purpose of which excess is to ensure against the loss of necessary information; trash. See also noise. Redundancy may be introduced in one part of an information system in order to reduce redundancy in another part of the system. For example, redundancy might be introduced into a search question involving a logical product in order to more precisely specify the information being sought so that the selected information will be less redundant.

Reference, see citation.

Reference tools, n., compilations of raw data or information which have been arranged in an orderly fashion, but have not been analyzed or evaluated.

Relevant, adj., refers to retrieved information which is related to the purpose of the user. Although relevance is a necessary condition for pertinence, relevance does not imply pertinence. For example, if a scientist were making a literature search in a certain field and the information system responded by giving him copies of some of his own writings in the field, these writings would be relevant because they were related to the field of interest of the user, but they would not be pertinent because the user already knew about them. See also pertinent.

Retrospective search, n., the process of reviewing all present and past information in order to select those items which are relevant to the inquiry which precipitated the search; emphasis is in the quality and completeness of the review and of the presentation of the search product.

Role indicator, n., a number or symbol designating the syntactic role which is attributed to an index term.

Search, v.t., to match successively the index terms of a store with a question in order to find documents or data which are described by the set of index terms equivalent to the question.

Search question, n., see question.

Search strategy, n., a definition of a set whose members are specified terms related according to a specified logical pattern, which set is to be used as a search question.

Selection, n., the act in response to a search by which a system physically indicates the documents or data described in index terms meeting the specifications of a search question.

Selective dissemination of information (SDI), n., a dissemination scheme which screens all input documents and automatically sends to each information user all documents relevant to his work.

Semantics, n., the study of the relation between symbols and the reality they denote; for example, the relation between natural language words and the reality they represent.

Separate index, n., see index, separate.

Sort, v.t., to divide a file according to specifications against which the file elements are matched; for example, file elements having '9' in the last position of a particular code may be separated from all other file elements.

Specificity, n., the degree of descriptive detail of the information retrieved.

Storage and retrieval device, information, n., a broad term which can include any physical structure built to perform some function involved with the storing or retrieving of information; for example, a typewriter or digital computer. Context must be relied upon to give the scope of the term in each particular situation in which it is used.

Storage and retrieval system, information, n., a set of related objects and processes whose integrated purpose is to store information so that it can be retrieved later and to retrieve information when desired.

Storage medium, n., see memory.

Store, n., the set of all data and index terms which have been physically recorded in some memory, such as magnetic core, disks or tapes, or punched cards.

Subject heading, n., a group of words describing a broad field of interest which is used to index documents at a more generic level than by descriptors or key words.

Superimposed coding, n., see coding, superimposed.

Surrogate, n., a substitute for a document, such as an abstract, a bibliographic citation, or an accession number.

Syntax, n., the study of the relation between the order of natural language words and the reality they denote because of their order.

Term, index, n., see index term.

Thesaurus, n., a list of vocabulary terms which have been authorized for use in an information retrieval system, together with a description of the hierarchical and semantic relationships between the terms, and a definition of the terms to the extent required.

Threaded file, n., a technique of structuring a search file by means of linking the address of all records which contain the same index term.

Transmission, n., a one-way transfer of data or information (as opposed to communication which is a two-way exchange of information).

Uniterm index, n., a manual searching method using an inverted file of term cards posted with the document numbers associated with the term cards.

Word, n., 1. a specific combination of one or more characters in a specific order. 2. specifically in computers, a space in computer memory capable of holding a specified number of characters.