AN EVALUATION OF THE INDEXING METHODS EMPLOYED IN A COMPUTERIZED INFORMATION SYSTEM USED IN THE AREA OF SPECIAL EDUCATION

Thesis for the Degree of Ph. D.
MICHIGAN STATE UNIVERSITY
Robert Don Moon, Jr.
1972

LIBRARY
Michigan State
University

This is to certify that the

thesis entitled

An Evaluation of the Indexing Methods
Employed in a Computerized Information System
Used in the Area of Special Education

presented by

Robert Don Moon, Jr.

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in Education

Major professor

Date February 21, 1972

O-7639

ABSTRACT

AN EVALUATION OF THE INDEXING METHODS
EMPLOYED IN A COMPUTERIZED INFORMATION SYSTEM
USED IN THE AREA OF SPECIAL EDUCATION

By

Robert D. Moon, Jr.

In 1966 The Council for Exceptional Children in cooperation with the Educational Resources Information Center established the CEC-ERIC Information Center. The Center was funded by the U. S. Office of Education's Bureau of the Handicapped. This study evaluates the indexing method used at the Center, compares the method with two alternative methods, analyzes the indexing vocabulary, describes changes in indexing procedures, and evaluates those changes. The data base for the indexing evaluation was 2100 abstracts contained in Volume I of Exceptional Child Education Abstracts (ECEA), a computerized journal produced from the Center's information files.

Three indexing methods were compared based on results of questions written by staff members acquainted with the Center's indexing procedures and by professional educators not familiar with the procedures. Each group wrote 105 logical search questions to retrieve target documents. All questions were used with each of the three indexing methods. The computer searches were made using Basic Indexing and Retrieval System (BIRS). Indexing Method 1 (the method normally used at the Center) extracted terms from titles of document surrogates and used ERIC

descriptors assigned by indexers. Indexing Method 2 used terms extracted from the titles and abstracts, and Indexing Method 3 used terms extracted from the titles and abstracts and ERIC descriptors assigned by indexers.

Estimated average recall for the Center staff was .73 for Method 1, .36 for Method 2, and .81 for Method 3. Estimated average recall for professional educators was .54 for Method 1, .77 for Method 2, and .80 for Method 3. Average Microprecision for the Center staff was .83 for Method 1, .76 for Method 2, and .74 for Method 3. Average Microprecision for professional educators was .94 for Method 1, .93 for Method 2, and .89 for Method 3. Six null hypotheses were tested at the .01 level to determine if there were significant differences between the search results of the CEC-ERIC staff and professional educators. These tests when based on estimated average recall indicated that the Center staff had significantly better results for Method 1, professional educators had significantly better results for Method 2, and there was no significant difference for Method 3. As measured by Average Microprecision the professional educators had significantly better results for all three methods. This data tends to suggest that the need for carefully controlled indexing languages is minimized in the field of education when sophisticated computer searching algorithms are available.

The vocabulary of the ERIC descriptors used to index Volume I of ECEA was compared with the vocabulary of the titles of abstracts in Volume I and an empirically based thesaurus developed by a retrospective search of five years' literature in special education. These comparisons implied that the vocabulary found in the ERIC descriptors had as much or

more similarity to the vocabulary based on the five-year retrospective search as did the titles.

A subjective analysis of the indexing terms used in Volume I of ECEA was performed by the indexing staff. This resulted in the establishment of a subset of the ERIC Thesaurus to be used in indexing future volumes of ECEA. The use of this reduced list was evaluated by processing 20 search questions on Volumes I and II of ECEA. In 19 of the searches precision was greater for documents retrieved from Volume II than from Volume I. In the one case where this was reversed the precision was almost identical and greater than .9.

AN EVALUATION OF THE INDEXING METHODS EMPLOYED IN A COMPUTERIZED INFORMATION SYSTEM USED IN THE AREA OF SPECIAL EDUCATION

By

Robert Don Moon, Jr.

A THESIS

Submitted to
Michigan State University
In partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

College of Education

1972

67-16,246

© Copyright by
ROBERT DON MOON, JR.

1972

ACKNOWLEDGMENTS

This study would not have been possible without the assistance and cooperation of many individuals. I wish to express my sincere appreciation to my doctoral committee not only for their assistance in this study, but also for their contribution to my graduate program--to Dr. Dale Alum for the insights he has given me into curriculum development; to Dr. Louise Sause for the rich understanding of children and their development which she has shared with me; to Dr. John Vinsonhaler for the opportunity provided me to work under his direction at the Information Systems Laboratory; and to Dr. William Walsh, chairman of my graduate committee for his able guidance in helping me to plan a very meaningful doctoral program.

Dr. Walsh has also been most helpful in making suggestions which have aided in clarifying the presentation of this study. His continued availability to answer questions and be of assistance despite his very busy schedule has been greatly appreciated.

The Basic Indexing and Retrieval System (BIRS) used in this study was developed under the direction of Dr. John Vinsonhaler. It was while working for Dr. Vinsonhaler at the Information Systems Laboratory that the study was first conceived. His continued technical assistance has also been greatly appreciated.

The majority of activities related to this study have taken place at the Council for Exceptional Children. It is difficult to conceive of an organization that could have provided a better environment for these

activities. Those involved in a special way at the Council for Exceptional Children have been Mr. William Geer, Executive Secretary; Dr. June Jordan, the first director of the CEC-ERIC Information Center; Dr. Don Erickson, present director of the CEC-ERIC Information Center; Carl Oldsen, the editor of ECEA: and his staff.

I am also indebted to Mr. John Hafterson who has contributed to this study both while he was at the Information Systems Laboratory at Michigan State and presently as a staff member at CEC. It has been especially helpful to have an individual with his technical capabilities with whom I could interact while I have been working on the project at CEC.

I am especially indebted to my wife Louise for typing and editorial assistance and to my children Bobby and Cami for their endurance and willingness to forego some activities while the study was being completed.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
	Problem	5
	Need for the Study	6
	Objectives of the Study	7
	Questions Examined	8
	Definitions and Acronyms	9
	Scope and Overview of the Study	14
	Overview of Procedures	14
II.	RELATED LITERATURE	16
	Systems and Systems Analysis	16
	What is a System?	17
	What is System Analysis?	20
	Information Retrieval Systems	23
	Indexing Methods--Content Analysis, Specification, and Control	31
	Indexing Languages and Retrieval Systems	32
	Methods of Machine Indexing	41
	Evaluation of Indexing Methods	59
	Descriptive Statistics for Document Retrieval	62
	Relevance Judgment	70

CHAPTER PAGE

II. (cont'd)

Comparison of Indexing Schemes 72

An Overview of BIRS--Basic Indexing and Retrieval System 76

The Executive Program--EXEC 77

Task Management Program--TASK 78

Translation Program--TRANS 78

Information File Maintenance Program--IFMP 80

Printed Indexing Program--PIP 80

Printed Listing Program--PLP 81

Descriptive Analysis Program--DAP 81

Description File Maintenance Program--DFMP 81

Description File Search Program--DFSP 82

Information File Retrieval Program--IFRP 86

Summary 87

III. THE DEVELOPMENT OF CEC-ERIC INFORMATION CENTER AND ITS PRESENT OPERATING STATUS 89

A History and Description of Central ERIC 90

Objectives of ERIC 91

The Growth of ERIC 93

The Future of ERIC 94

The Development of the CEC-ERIC Clearinghouse 95

The Early Operation of the Center 98

The Establishment of Data Processing Procedures 101

The Decision to Publish a Computerized Journal 101

CHAPTER PAGE

III. (cont'd)

An Overview of the Operating Procedures Used by the CEC-ERIC Information Center 104

Legend and Nomenclature 104

Model Developed for the CEC-ERIC Information Center . . . 107

Overview of the Information Center's Major Activities . . 107

Overview of Major Input and Output 110

Overview of Evaluation and Processing Modifications . . . 112

Overview and Model of the Information Center's Operation 112

The Publication of Exceptional Child Education Abstracts 115

Selective Publication 119

Descriptive Statistics About the Present Operating Status of the CEC-ERIC Information Center 121

The Center's Holdings--Types of Documents and Their Subjective Content 121

Acquisition and Processing Rates 122

Information Request Processing Statistics 124

Processing Costs 125

Summary 128

IV. PROCEDURES USED IN THE EVALUATION AND ANALYSIS OF THE INFORMATION CENTER INDEXING METHODS 130

The Evaluation of the Indexing Procedures Used in Volume I of ECEA 131

Questions Examined 132

CHAPTER PAGE

IV. (cont'd)

Selection of Target Documents 134

Preparation of Questions to Retrieve Target Documents . . 134

Relevance Judgments 136

Measurement Techniques Employed in the Indexing Evaluation 136

The Content Analysis of the Vocabulary Used in Indexing Volume I of ECEA 137

Compilation of Indexing Terms Assigned to Volume I of ECEA 139

Subjective Analysis by Indexers of Terms Used in Volume I of ECEA 140

A Comparison of the Word Vocabulary Used in the Indexing Terms of Volume I of ECEA with Words Extracted from the Literature 140

Analysis of the Vocabulary Used in Writing Questions to Retrieve Target Documents 143

Analysis of Changes in Indexing Procedures Between Volume I and Volume II of ECEA 143

Summary 145

V. RESULTS OF THE EVALUATION AND ANALYSIS OF THE INFORMATION CENTER INDEXING METHODS 146

A Comparative Evaluation of Three Indexing Methods 147

Questions Examined 148

Indexing Methods Compared 149

Results of the Comparison of Indexing Methods 149

CHAPTER PAGE

V. (cont'd)

Factors Important to the Analysis of Data Resulting from the Comparison of Indexing Methods 158

Analysis of the Question Vocabulary 163

Analysis of the Indexing Vocabulary Used in Volume I of ECEA 164

Notation 167

Results of Vocabulary Comparisons of Three Word Lists . . . 168

A Subjective Analysis of Terms Selected From the ERIC Thesaurus to Index Volume I of ECEA 171

Results of Subjective Evaluation of the ERIC Descriptors Used in Volume I of ECEA 172

The Effect of Indexing Procedure Changes in Volume II of ECEA 175

Results of Indexing Procedure Changes 176

Summary 177

VI. SUMMARY, CONCLUSIONS, RECOMMENDATIONS, AND IMPLICATIONS . . . 181

Summary 181

Procedures Used at the CEC-ERIC Center 181

A Comparison of Three Indexing Methods 182

Comparison of Vocabulary of Three Word Lists 183

Changes in Indexing Procedures 183

Conclusions 184

Results of Testing Null Hypothesis 1 184

Results of Testing Null Hypothesis 2 185

Results of Testing Null Hypothesis 3 185

CHAPTER PAGE

VI. (cont'd)

Results of Testing Null Hypotheses 4, 5, and 6 185

Results of Testing Null Hypothesis 7 185

Results of Testing Null Hypothesis 8 186

Results of Testing Null Hypothesis 9 186

Results of Testing Null Hypothesis 10 187

Interpretation of the Results of the Comparison of Three Indexing Methods 188

Reflections on Methodology Used in Comparing Indexing Methods 190

Interpretation of the Vocabulary Comparisons 192

Interpretation of the Effect of Changing Indexing Procedures 194

Recommendations 195

Data Related to Recommendation 1 195

Recommendation 1 196

Data Related to Recommendation 2 196

Recommendation 2 197

Observations Related to Recommendation 3 197

Recommendation 3 197

Observations Related to Recommendation 4 198

Recommendation 4 198

Observations Related to Recommendation 5 199

Recommendation 5 199

CHAPTER PAGE

VI. (cont'd)

Implications 200

The Use of Controlled Indexing Vocabularies 200

An Evolving Thesaurus 202

Selective Publication From Information Files 204

BIBLIOGRAPHY 207

APPENDIX A 214

LIST OF TABLES

TABLE		PAGE
3.1	An Analysis of Information Requests Processed by The CEC-ERIC Information Center During the First Quarter, 1971	126
5.1	Descriptive Statistics Resulting From the Evaluation of Three Indexing Methods	150
5.2	Data and Calculations Used in Testing Null Hypothesis 1 . .	155
5.3	Data and Calculations Used in Testing Null Hypothesis 2 . .	156
5.4	Data and Calculations Used in Testing Null Hypothesis 3 . .	157
5.5	Data and Calculations Used in Testing Null Hypothesis 4 . .	159
5.6	Data and Calculations Used in Testing Null Hypothesis 5 . .	160
5.7	Data and Calculations Used in Testing Null Hypothesis 6 . .	161
5.8	Data and Calculations Used in Testing Null Hypothesis 7 . .	165
5.9	Data and Calculations Used in Testing Null Hypothesis 8 . .	166
5.10	Data and Calculations Used in Testing Null Hypothesis 9 . .	170
5.11	Results of Indexers' Subjective Analysis of Terms Used to Index Volume I of ECEA	173
5.12	Data and Calculations Used in Testing Null Hypothesis 10 . .	178
5.13	Search Results of Twenty Questions Used on Volume I and Volume II of ECEA	179

LIST OF FIGURES

FIGURE		PAGE
2.1	Input, Processing, and Output	25
2.2	Input, Processing, and Output with Feedback	25
2.3	A Portion of the ERIC Thesaurus	38
2.4	Illustration of BIRS Word Extraction Techniques	49
2.5	Examples of Permuted or Key-Word in Context (KWIC) Indexes	53
2.6	The Partitioning of a Document Collection by A Search Question	64
2.7	A Comparison of Various Types of Recall and Precision Averages	67
2.8	An Overview of the Basic Indexing and Retrieval System (BIRS)	79
2.9	Examples of Search Questions	84
3.1	Flowcharting Symbols	105
3.2	Overview of Information Center Major Activities	108
3.3	Overview of Major Input and Output	111
3.4	An Overview of the Information Center's Evaluation and Systems Modification Components	113
3.5	An Overview and Model of the Information Center's Operations	114
3.6	Sample ECEA Abstract	116

FIGURE PAGE

3.7 Samples of ECEA Author and Subject Indexes 117

3.8 Subject Content Description of Information Center Holdings Based on 5715 Acquisitions in Volumes I & II of ECEA 123

4.1 A Description of Data and Descriptive Statistics Used in Comparing Various Indexing Methods 138

5.1 Number of Target Documents Retrieved by Three Indexing Methods 151

5.2 Average Microprecision for Three Indexing Methods 152

5.3 Number of Relevant Documents Retrieved by Each Indexing Method 153

1A Flowcharting Symbols 215

2A Overview of Information Center Major Activities 218

3A Overview of Major Input and Output 221

4A An Overview of the Information Center's Evaluation and Systems Modification Components 222

5A An Overview and Model of the Information Center's Operations 224

6A Acquisition Control and Document Management 226

7A File Maintenance 230

8A File Processing for Exceptional Child Education Abstracts 235

9A A Computer Search - Predefined Process 9 239

10A Information Request Processing 241

11A Procedures for Processing Selective Publications 244

CHAPTER I: INTRODUCTION

There is a valid concern over the scientist's ability to keep abreast of the rapid growth of knowledge in his discipline. The realization of this problem is not new. In 1936 historian H. G. Wells presented a paper entitled "World Encyclopaedia" which stated his concern about the ineffective use and lack of coordination of knowledge. In this paper he suggests, "Possibly all the knowledge and all the corrective ideas needed to establish a wise and stable settlement of the world's affairs in 1919 existed in bits and fragments, ..." He continues in descriptive terms to describe the human species as a "man of the highest order of brain, who through some lesions or defects or insufficiencies of his lower centres, suffers from the wildest uncoordinations; . . . ." Finally in his presentation Wells suggests a world encyclopaedia as a means to "solve the problem of that jigsaw puzzle and bring all the scattered and ineffective mental wealth of our world into something like a common understanding, ..."[1]

Subsequent to the presentation of H. G. Wells one finds in the literature with increasing frequency similar expressions of concern and suggestions for solutions. Alvin Weinberg succinctly summarizes this concern in the following statement:

---

[1]H. G. Wells, "World Encyclopaedia," World Brain (Garden City, New York: Doubleday, Doran & Co., Inc., 1938), pp. 3-35. Paper read at the Royal Institution of Great Britain Weekly Evening Meeting, Friday, November 20, 1936.

> The ideas and data that are the substance of science and technology are embodied in the literature; only if the literature remains a unity can science itself be unified and viable. Yet because of the tremendous growth of literature, there is danger of science fragmenting into a mass of repetitious findings, or worse, into conflicting specialities that are not recognized as being mutually inconsistent. This is the essence of the "crisis" in scientific and technical information.[2]

When one looks at the rate at which knowledge, or at least literature, is growing, the scope of the problem becomes staggering. As reported in 1965 there were approximately seven new papers published each year for every hundred previously published. Since 1860 the general trend indicated an exponential increase in the literature with the total literature doubling approximately every thirteen and a half years. There have been only a few noticeable interruptions in this growth. These occurred during World Wars I and II. Since World War I, with the exception of the World War II period, the rate of growth appears to be even more rapid than a doubling every thirteen and a half years.[3]

Individual disciplines have attempted to solve this problem with reference works such as Psychological Abstracts, Chemical Abstracts, and the Educational Index. Such reference works are a significant aid to the scientist; however, they are far from a total solution. As the literature has continued to increase, the volume of such journals has also increased; and the indexing of articles contained in the journals has become a more difficult problem.

When the indexing is too broad, the scientist is still confronted

---

[2]Alvin Weinberg, Science, Government, and Information: The Responsibilities of the Technical Community and the Government in the Transfer of Information, President's Science Advisory Council, (Washington: Government Printing Office, 1963), p. 7.

[3]Derek J. deSolla Price, "Network of Scientific Papers," Science, CXLIX (July 30, 1965), 510-515.

with an awesome volume of articles to review. Often he would like only specific articles which contain information about several specific categories. When this is the case, it is necessary for him to look at the intersection of lists of articles where each list may contain hundreds of individual articles.

Two advances in technology were recognized almost immediately to provide assistance in helping the scientist cope with the expanding volume of literature. These advances were microfilm and the computer. The contribution of microfilm was straightforward. It allowed a large volume of material to be stored in a small area and made it possible to have copies of documents for a small cost. Copyright laws are making it difficult to apply this medium to recent publications thus preventing it from reaching its potential effectiveness.

Being able to find specific information in an ever expanding volume of literature is a problem which remains whether material is put on microfilm or is in hard copy. It was almost immediately recognized that the computer provided a powerful tool to assist in solving this problem. Methods of indexing which could be used by the computer were experimentally examined before computer technology was capable of implementing them on large collections of documents. One of the first such experiments related to the Library of Congress and was conducted about 1953. This involved a comparison of proposed coordinate indexing with subject heading systems then in use by the Library of Congress. The experiment involved approximately fifteen thousand documents.[4] In 1954

---

[4]C. D. Gull, "Seven Years of Work on the Organization of Materials in the Special Library," American Documentation, VII (October, 1956), 320-329.

another series of experiments was initiated which continued over a period of years and now is collectively called the Cranfield Studies.[5] Following these projects an increasing number of experiments have attempted to determine what the most effective indexing methods are for computerized information retrieval systems.

Existing reference works such as Chemical Abstracts are using this new technology to make their services more effective. There have been increasing numbers of new information services taking advantage of this technology, with one of the most notable being related to medical literature, MEDLARS (Medical Literature Analysis and Retrieval System.) A number have also resulted from the aerospace research. In the area of education ERIC (Educational Resources Information Center) has developed. Other educational Information Centers are listed in the Directory of Educational Information Centers, published in 1969 by the U. S. Government Printing Office. (Document #FS5.212:12042.)

One of the major problems in dealing with the retrieval of information is the ambiguity of language. The most effective systems for retrieving information generally deal with types of information which have a very technical and well-defined nomenclature. A classic example of this is the area of chemistry. The need for eliminating the ambiguity of vocabulary in this area resulted in a conference held in 1930 to develop an effective system for naming chemical compounds.[6] This

---

[5]Charles P. Bourne, "Evaluation of Indexing Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: Interscience Publishers, 1966), I, 176.

[6]Commission and the Council of the International Union of Chemistry, "Definitive Report of the Commission on the Reform of Nomenclature of Organic Chemistry," Journal of American Chemical Society, LX (1933), 3905-25.

has helped Chemical Abstracts, related reference works, and their computerized retrieval systems to be one of the more effective information networks in existence.[7]

One has only to look at the literature concerning the research done on thesaurus refinement to become aware of the significance of this problem. The indexing systems used by various medically-oriented retrieval systems have been the focus of considerable research on this problem.[8] Despite the technical nature of medical terminology, the literature shows there is still a considerable problem in finding the best indexing methods.

The area of education has a nomenclature which is much less structured and more ambiguous than technical areas such as chemistry or medicine. One has only to examine the ERIC Thesaurus and observe the large number of broad terms, related terms, and narrow terms listed for a given concept to obtain a quick appreciation of the problem.[9]

Problem

The increasing amount of material being published in the area of education has made it important to find better ways to store and disseminate information.

---

[7]F. A. Tate, "Handling Chemical Compounds in Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: Interscience Publishers, 1967), II, 285-310.

[8]John O'Connor, "Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems," American Documentation, XV (April, 1964), 96-104; and Montgomery and D. R. Swanson, "'Machine' Like Indexing by People," American Documentation, XIII (October, 1962), 359-66.

[9]Thesaurus of ERIC Descriptors: Working Copy Descriptor Listing, ERIC Processing and Reference Facility (Bethesda, Md.: Leasco Systems and Research Corporation, August, 1971), pp. 1-244.

The Council for Exceptional Children has an Information Center (CEC-ERIC Information Center) which is a part of the ERIC system,[10] and it is envisioned that a major contribution of this study will be to describe how BIRS (Basic Indexing and Retrieval System)[11] was used in developing an information system used by the Center.

Specifically this study evaluates the indexing methods which are used at the CEC-ERIC Information Center and compares them with alternative methods available to the Center. The results are examined for any implications that suggest ways of improving the indexing methods used at the Center and for implications which might be generalized to the total field of education.

Need for the Study

The effectiveness of any information retrieval system directly relates to the indexing methods used. If the literature in the field of education is to be more effectively used by researchers, it is important to locate and identify studies which contribute to a significant understanding and improvement of the nomenclature. This study will examine indexing methods and describe the overall procedures used at the CEC-ERIC Center for processing information about special education. The results will be examined for implications which may contribute to further studies relating to the total field of education.

---

[10]"All About ERIC," Journal of Educational Data Processing, VII (April, 1970), 51-129; and June B. Jordan, "CEC-ERIC-IMC: A Program Partnership in Information Dissemination," Exceptional Children, XXXV (December, 1968), 311-313.

[11]John F. Vinsonhaler, John M. Hafterson, Stuart W. Thomas, Jr. (editors), Basic Information Retrieval System Technical Manual (East Lansing, Michigan: Information Systems Laboratory, College of Education, Michigan State University, 1970), Vols. I-XII.

Many information centers and systems have been developed to help cope with the rapid growth of knowledge. Various disciplines, including education, are using new methods involving computers and microfilm. While there is considerable documentation about the various ways in which these techniques are used, the documentation for any single system is often contained in many places, fragmented and sketchy.

The CEC-ERIC Center uses two computerized systems for information handling: (1) BIRS and (2) a commercial system for computer typesetting.[12] The Information Center has interfaced these in a unique manner, allowing for selective computerized publication.

The manner in which the CEC-ERIC Information Center is using computerized information retrieval and computerized publication has not been previously documented, nor have the indexing methods been evaluated. By describing the procedures used at the Center and evaluating the indexing methods, this study provides information which may contribute to a better understanding of educational nomenclature and the dissemination of educational information.

Objectives of the Study

This study has the following objectives:

1. To document the development of the information system used by the CEC-ERIC Information Center.
2. To document the manner in which the CEC-ERIC Information Center uses the BIRS system and other computerized programs.

---

[12]Exceptional Child Education Abstracts, II (November, 1970), see inside front cover.

3. To evaluate the indexing methods used by the CEC-ERIC Information Center.
4. To recommend means for improving these indexing methods.
5. To examine the results of the study for implications about how the CEC-ERIC Information Center might improve its overall effectiveness.
6. To examine the results of the study for implications concerning improvement of communication within the total field of education.

Questions Examined

In the evaluation of the indexing methods the following questions are examined:

1. How effective is the indexing method used by the Information Center for:
   a. CEC-ERIC staff who are familiar with the Information Center's indexing system?
   b. Professional educators who are not familiar with the Information Center's indexing system?
2. How effective is a computerized indexing method which extracts terms from the titles and abstracts for:
   a. CEC-ERIC staff who are familiar with the Information Center's indexing system?
   b. Professional educators who are not familiar with the Information Center's indexing system?
3. How effective is the indexing method used at the CEC-ERIC Information Center when combined with machine indexing of abstracts for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?

b. Professional educators who are not familiar with the Information Center's indexing system?

4. Is the vocabulary of the terms used in the indexing method employed by the CEC-ERIC Information Center found in the literature of special education?

Definitions and Acronyms

The field of computerized information retrieval has been rapidly developing since about 1953. As with many new technical areas there is a certain amount of ambiguity in the nomenclature. The alphabetized definitions in this section in most cases have been chosen to reflect a consensus of the literature; however, it is possible to find some of the same or very similar ideas represented in the literature by terms other than those in the list of definitions. The several mathematical definitions have been given without the use of mathematical symbols, but are consistent with definitions using symbolic nomenclature.

For the reader wishing a more comprehensive introduction to the nomenclature of information retrieval, articles or books by Bourne, Swets, and Lancaster may provide a quick introduction,[13] while a simple

---

[13]Charles P. Bourne, "Evaluation of Indexing Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor New York: Interscience Publishers, 1966), I, 171-190; Donald W. King, "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (Chicago: Encyclopaedia Britannica, 1968), 61-104; F. W. Lancaster, "MEDLARS: A Report on the Evaluation of Its Operating Efficiency," American Documentation, April, 1969, pp. 119-142; and John A. Swets, "Information-Retrieval Systems," Science, CXLI, (July 19, 1963), 245-250.

text on sets or symbolic logic should give the reader a further understanding of the mathematical terms.

The definitions and acronyms are meant to be a reference for the reader and in this document the terms are consistently used as defined unless an exception is noted. It may be helpful to a reader unfamiliar with information retrieval to read the definitions of the twelve terms in the following list in the numeric order specified.

1. Document Surrogate
2. Term
3. Indexing Method
4. Information file
5. Description file
6. Target Document
7. Hit
8. Precision
9. Recall
10. Set
11. Union
12. Intersection

BIRS The Basic Indexing and Retrieval System is a generalized system of computer programs designed for storing, indexing, and retrieving information.[14]

CEC The Council for Exceptional Children.

DAP The Descriptive Analysis Program (one of the BIRS Programs) aids the user with the task of indexing or classifying informational elements. DAP reads informational elements (abstracts) and searches them for descriptive terms or phrases.[15]

---

[14]John F. Vinsonhaler, John M. Hafterson, Stuart W. Thomas, Jr. (editors), Basic Information Retrieval System Technical Manual (East Lansing, Michigan: Information Systems Laboratory, College of Education, Michigan State University, 1970), Vols. I-XII.

[15]John F. Vinsonhaler, John M. Hafterson, and Stuart W. Thomas, (editors), Basic Information Retrieval System Technical Manual (East Lansing, Michigan: Information Systems Laboratory, College of Education, Michigan State University, 1970), I, 110.

DFMP The Description File Maintenance Program (one of the BIRS Programs) reads descriptions and access numbers of informational elements and stores them on the Description File (DFT) to provide an index to the contents of the Information File (IFT).[16]

DFSP The Description File Searching Program (one of the BIRS Programs) is designed to read queries; to search the DFT for relevant informational elements; and to store the access numbers of the most relevant elements on the Question File (QFT).[17]

Description File A file containing descriptions of the information found on an information file. The object of such a file is to help retrieve information from an information file. In this study the phrase "description file" will always refer to a computerized description file.

Document Surrogate A substitute or abridged representation of an original document.

ECEA Exceptional Child Education Abstracts, a journal of abstracts in the field of special education published by the Council for Exceptional Children.

ERIC The Educational Resources Information Center.

Estimated Average Recall The number of successful attempts to retrieve target documents by computerized searches divided by the total number of attempts. An attempt is a computer search to retrieve one target document.

---

[16] Ibid.

[17] Ibid.

EXEC The Executive Program (one of the BIRS Programs) is designed to store and retrieve system components by augmenting the supervisory monitor.[18]

Hit A document retrieved by a computer search which is considered to be relevant to the computer search question.

IFMP The Information File Maintenance Program (one of the BIRS Programs) maintains an Information File Tape (IFT) by reading and storing informational elements (abstracts) of arbitrary length. The IFMP may also be used to generate printed books.[19]

IFRP The Information File Retrieval Program (one of the BIRS Programs) is designed to read queries and access numbers from the Question File Tape (QFT); and to generate reports with the information elements (abstracts) read from the IFT.[20]

IMC/RMC An Instructional Materials Center/Regional Media Center.

Indexing Method A method, procedure, or algorithm for selecting and assigning terms to describe a document or document surrogate.

The indexing method may be a manual method involving human indexers who assign the terms or a computerized method which extracts or selects terms to be assigned to documents or document surrogates. The computerized method may assign terms from a predetermined list according to an algorithm or extract the terms according to an algorithm from any portion of the text of the document or document surrogate.

---

[18] Ibid. [19] Ibid. [20] Ibid.

Information File A file containing information. This file may or may not be stored on a computer storage device.

Intersection The intersection of two sets will be the set of all objects common to both sets.

PIP The Printed Indexing Program (one of the BIRS Programs) prepares a traditional subject index using informational elements read from cards or from the IFT.[21]

PLP The Printed Listing Program (one of the BIRS Programs) provides printed books, i.e., listings of abstracts, ordered by the contents of the abstracts. The books produced by PLP are similar to those produced by the IFMP, except that the latter are ordered by the Information File Tape (IFT) access number.[22]

Precision The number of hits divided by the total number of documents retrieved in a computerized search.

Recall The number of hits divided by the number of documents in the information file which are relevant to the search question.

Set A well-defined collection of objects.

Target Document A document which is randomly selected from an information file to be used as the basis for writing a search question.

Term A word or phrase assigned to describe (used to index) a document or document surrogate.

---

[21] Ibid.

[22] Ibid.

Union The union of two sets is a new set of all objects belonging to either or both of the original sets.

Scope and Overview of the Study

This study describes the procedures used at the CEC-ERIC Information Center and attempts to provide sufficient detail and organization so as to allow others to use them as a model. The study also evaluates indexing methods used at the CEC-ERIC Information Center. The methods are evaluated in the context of a data file containing information about special education and an information retrieval system based upon BIRS. The data base used in the initial evaluation of indexing procedures is 2100 documents contained in Volume I of Exceptional Child Education Abstracts (ECEA).

Overview of Procedures

The indexing methods used at CEC-ERIC Information Center involve manually selecting terms from the ERIC Thesaurus to index each abstract. The vocabulary of the terms assigned in Volume I of ECEA was compared with the collective vocabulary found in the titles of the 2100 documents. This vocabulary was also compared with the vocabulary of a thesaurus developed by Samuel Price for use in the area of special education.[23]

One hundred and five target documents (abstracts) were randomly selected from Volume I of ECEA and computer-searchable description files were created of these abstracts. To search these files questions were generated by members of the CEC-ERIC staff familiar with the in-house

---

[23]Samuel T. Price (comp.), Thesaurus of Descriptors for an Information Retrieval System in the Subject Matter Area of Special Education, (Normal, Illinois: Special Education Instructional Materials Laboratory, Illinois State University, January, 1970).

indexing procedures and by professional educators who had no knowledge of the ERIC Thesaurus or in-house indexing procedures. The search results for these questions were used to compare the effectiveness of various description files generated by different indexing methods.

A subjective analysis of the ERIC terms assigned to Volume I of ECEA was made by the CEC-ERIC indexers in an attempt to develop and refine the Thesaurus which has been used in indexing successive volumes. A series of search questions used by the Center to create selected bibliographies were used in searches of both Volume I and II of ECEA. The individual responsible for editing the bibliographies determined the relevance of documents retrieved. Precision of documents retrieved over Volume I was compared with the precision of documents retrieved over Volume II to determine if it had been possible to improve the in-house indexing procedures.

The theoretical base for the evaluation procedures used in this study is contained in the literature and is examined in detail in the following chapter.

CHAPTER II: RELATED LITERATURE

This chapter considers five specific areas of related literature. They are: (1) systems and systems analysis, (2) information retrieval systems, (3) indexing methods--content analysis, specification, and control, (4) the evaluation of indexing methods, and (5) an overview of the Basic Indexing and Retrieval System (BIRS). The review of related literature has been designed to stress the importance of viewing an information system as a whole, including its environment, rather than concentrating on specific components without attention to their context.

The chapter is organized so that it moves from general to specific, starting with broad concepts of systems and systems analysis, next moving to information retrieval systems as a particular category of systems, then examining specific processes in an information retrieval system that are most important to this study; namely, indexing methods and indexing evaluation, and finally giving an overview of BIRS--a specific example of a generalized computer information retrieval system. The overview of BIRS is included not only because it provides a specific example, but also because it is used by the CEC-ERIC Information Center and in the evaluation of the Center's indexing methods.

Systems and Systems Analysis

One has only to look at the literature in education, management, behavioral sciences, applied sciences, and other fields to discover that

the terms "system", "system concept", and "system analysis" are used repeatedly, often without definition. The frequent use of these terms in popularized reporting of science and technology would imply that they are important and that most people understand their meanings. However, an examination of literature related to system analysis reveals that there is not a consensus concerning the meanings of these terms and as defined by various experts the meanings embrace broad concepts.

What is a System?

The question "What is a system?" does not have a simple answer and might be responded to by the question "What kind of system?" Texts on systems analysis speak of natural, man-made, mathematical, physical science, and engineering systems. When the term "system" is used in relation to mathematics it most often means a set of rules. When it is used in connection with the physical sciences, or natural systems, it is taken to mean a portion of the universe around which an imaginary boundary has been drawn for the purpose of study. In engineering the word "system" is interpreted as "meaning an organized working total, an assemblage of objects united by a form of regular interaction or interdependence."[24]

While mathematical systems might be considered as man-made systems, they are distinctly different from the term system used in an engineering sense, in that a mathematical system deals only with ideas, whereas an engineering system usually deals with man-made systems involving real

---

[24]Dimitris N. Chorafas, Systems and Simulation (New York: Academic Press, 1965), p. 4.

objects. The broadest definition of a system found in this study of related literature defined it to be "a set of interacting elements."[25] This definition could include natural systems such as the solar system, or man-made systems with the exception of mathematical systems where the elements (ideas) do not interact. In a mathematical system the elements exist only in man's mind, and have no substance or energy of their own with which to interact.

In the science of information retrieval the concept of system or system analysis is most closely related to the terms as they are used in engineering. Following are seven selected quotations relating to the term "system" which have been grouped for convenient examination:

1. A system is broadly defined as "a group of interdependent elements acting together to accomplish a predetermined task."[26]

2. An integrated assembly of interacting elements designed to carry out cooperatively a predetermined function.[27]

3. There has been a growing realization of the existence of an identifiable science of systems, comprising a body of concepts, methods, and above all, a philosophy of treating the whole rather than bits and pieces. The new field of systems science is as yet only loosely defined and has different meanings in different contexts. Its ultimate domain is still seen only dimly as compared to traditional disciplines such as physics, mathematics, and engineering.[28]

---

[25]Harry J. White and Selmo Tauber, Systems Analysis (Philadelphia: W. B. Saunders Company, 1969), p. 4.

[26]Ibid., p. ix, (Preface by F. Gordon Smith).

[27]Harry J. White and Selmo Tauber, Systems Analysis (Philadelphia: W. B. Saunders Company, 1969), p. 3, citing R. E. Gibson, "A Systems Approach to Research Management, Part I," Research Management, V (1962), 215.

[28]White and Tauber, op. cit., p. 1.

4. A set of objects with relationships between the objects and their attributes.[29]

5. Although system is a general term used in many senses, it does convey a very important meaning not readily described in any other way. The word derives from a Greek verb meaning to place or set together, and Webster's New World Dictionary gives the definition: "A set or arrangement of things so related or connected as to form a unity or whole: as, a solar system, irrigation system, supply system."[30]

6. The concept of a man-made system usually includes the idea of optimizing certain parameters such as cost, efficiency, size, or reliability, in terms of criteria derived from externally imposed value systems. The value systems are subjective and are based on a variety of factors such as economic, social, or even political. Adjustments or trade-off values between such considerations as cost, reliability, and prestige are frequently necessary.[31]

7. A system can contain within its structure a number of subsystems each of which has all the attributes of a system when considered as an integrated collection of components.[32]

An examination of the above quotations reveals at least five important concepts:

1. The concept of a system relating to a whole as indicated by statements 1, 2, 3, 4, and 5.

2. The concept of interacting components or elements as indicated by statements 1, 2, and 4.

3. The concept of a system existing to accomplish a specified task of function as indicated by statements 1 and 2.

---

[29]A. Hall and R. Fagan, "Definition of a System," General Systems, Vol. V of Yearbook of the Society for General Systems, 1956), p. 18.

[30]White and Tauber, op. cit., p. 3.

[31]Ibid., p. 4.

[32]Ralph Deutsch, System Analysis Techniques (Englewood Cliffs: Prentice-Hall, Inc., 1969), p. 2.

4. The concept of optimizing according to predetermined parameters such as cost, efficiency, size, reliability, etc. as indicated by statement 6.

5. The concept that a system can contain subsystems each of which has all the attributes of a system is indicated by statement 7.

To summarize, the term system is used in many ways. It is used to refer to mathematical systems, physical or natural systems and non-mathematical man-made systems and in each of these areas the word "system" has a complex meaning with numerous implications. The meaning of the word which is most directly related to information science is when it is used as related to man-made mon-mathematical systems which include the five above-mentioned concepts. The fourth concept, the concept of parameters to be used in measuring the effectiveness of the system is important because it is this component that establishes criteria for evaluation, modification, and redesign--functions related to systems analysis.

What is Systems Analysis?

Defining the term systems analysis is not a simple task as emphasized by a statement in a recent book entitled Systems Analysis Techniques. The statement says:

> A reasonable expectation from a book on systems analysis would be to find an introductory section which defines the term in an unambiguous fashion. This starting point, while admittedly desirable, is not feasible because the wide spectrum encompassing the field of systems analysis is still in its infancy and formative state.[33]

An examination of references to the terms "system analysis," "system concept," "systems science," and "systems engineering" indicates

---

[33] Ibid., p. 1.

that there is considerable overlapping in the meaning of these terms, sufficient overlapping that it does not seem advisable to make a distinction in their meaning. The following are a number of selected quotations relating to these terms which are grouped for examination:

1. The new and promising discipline of system analysis seeks to determine the optimum means for accomplishing the task described in the problem statement.[34]

2. System analysis is an attempt to define the most feasible, suitable, and acceptable means for accomplishing a given purpose.[35]

3. System analysis is merely a study of a system--but it should be emphasized that one does not usually study a system as an end in itself. Rather the explicit motivation of any system study is to generate information so that a decision can be made.[36]

4. Systems analysis: The analytic study of system, where analytic is taken in its most general sense.[37]

5. Systems analysis, by its very meaning, cuts across academic departmental barriers; an interdisciplinary approach is therefore necessary, both in the marshaling of varied resources and in the manifold applications.[38]

6. Essentially the system concept is that of examining the overall interactions of a group of items rather than focusing attention on the operation of each of the component elements in turn.[39]

7. The system concept can be interpreted as stripping the non-essential details from a collection of interacting elements so that the structure of the interrelations is laid bare for study.[40]

---

[34]Chorafas, op. cit., p. ix. (Preface by F. Gordon Smith).

[35]Chorafas, op. cit., p. 2.

[36]Deutsch, op. cit., p. 8.

[37]White & Tauber, op. cit., p. 5.

[38]Ibid., p. 2.

[39]Deutsch, op. cit., p. 2.

[40]Ibid.

8. Systems science--The science that is common to all large collections of interacting functional units that are combined to achieve purposeful behavior.[41]

9. Systems engineering--A process in which complex systems are idealized, designed and manipulated by conscious rational processes based upon the scientific method.[42]

10. Definitions of systems science and systems engineering generally include requirements for utility or at least directed behavior. By contrast, definitions of system tend to be more abstract.[43]

The ten statements illustrate the difficulty originally expressed in defining system analysis and the overlapping of various terms using the word "system." These statements about system analysis, system concept, system engineering and system science appear to be typical of other statements found in the literature.

There are many concepts which are implied in the preceding statements. Specific concepts which can be identified in one or more of the statements are:

1. That system analysis involves the study of a system as indicated by statements 3, 4, and 7.

2. That system analysis involves finding a best, optimum, or most feasible, way to accomplish a given task as indicated in statements 1, 2, 8, 9, and 10.

3. That the study or analysis is motivated by a problem or need for information to make a decision as indicated by statements 1, 3, and 10.

---

[41]White and Tauber, op. cit., p. 3, citing Institute of Electrical and Electronics Engineers, Systems Science Committee, Charter.

[42]White and Tauber, op. cit., p. 3.

[43]Ibid., p. 4.

4. That the study or analysis focuses on interaction between components of a system and their relationship to the system as a whole rather than examining components in isolation as indicated by statements 6, 7, and 8.
5. That the study or analysis is rational or formalized as indicated by statements 4 and 9.
6. That the study or analysis usually involves an interdisciplinary approach as indicated by statement 5.

One reason for the variety of perceptions about systems analysis may be that it is interdisciplinary with some sources indicating that those doing systems analysis need to be "generalists" to insure an unbiased approach to the entire problem.[44] The judgments involved in determining which of many approaches should be used and what is representative data of the total available to be examined has led others to ask the question "Does the product of such an effort deserve to be called a science or an art?"[45] The answer to this question is not clear for there are obviously elements of both.

Information Retrieval Systems

Those involved with information science are primarily concerned with man-made systems designed to accomplish a specific task. A simple definition consistent with the concept of man-made systems defines a

---

[44]Deutsch, op. cit., p. 1; and White and Tauber, op. cit., p. ix.

[45]Chorafas, op cit., p. 3.

system as a "group of interdependent elements acting together to accomplish a predetermined task.[46]

A definition of system analysis given by Borko incorporates many of the concepts suggested by statements about systems and system analysis. This definition will serve as a reference for the term system analysis as it relates to this study.

> Systems analysis is a formal procedure for examining a complex process or organization, reducing it to its component parts, and relating these parts to each other and the unit as a whole in accordance with an agreed upon performance criteria. Systems design is a synthesizing procedure for combining resources into a new pattern.[47]

By substituting appropriate synonyms one finds the statement indicates that systems analysis includes: (1) a formal procedure for studying a system (a complex process or organization), (2) reducing the system to its component parts for convenient study. (3) relating these parts to each other and studying their interaction, (4) emphasis upon the unit (system) as a whole and the relationship of the interacting components with this whole, (5) the existence of a performance criteria by which the system may be judged.

Information retrieval systems in their simplest form might be illustrated by Figures 2.1 and 2.2.[48] While these diagrams are provided with references, it is probably unnecessary because of their common use. Claire K. Schultz has constructed a similar diagram

---

[46]Harry J. White and Selmo Tauber, Systems Analysis (Philadelphia: W. B. Saunders Company, 1969), p. 3, citing R. E. Gibson, "A Systems Approach to Research Management, Part I," Research Management, V (1962) 215.

[47]Harold Borko, "Design of Information Systems and Services," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, 1967), II, 37.

[48]White and Tauber, op. cit., p. 5.

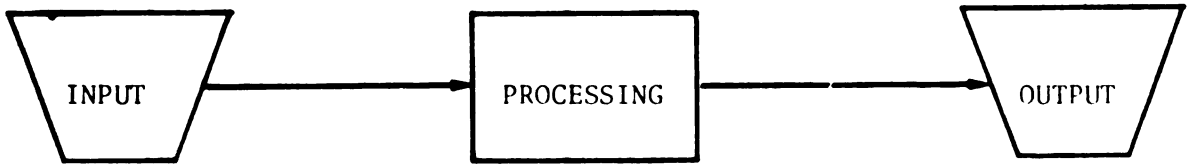


FIGURE 2.1

INPUT, PROCESSING, AND OUTPUT

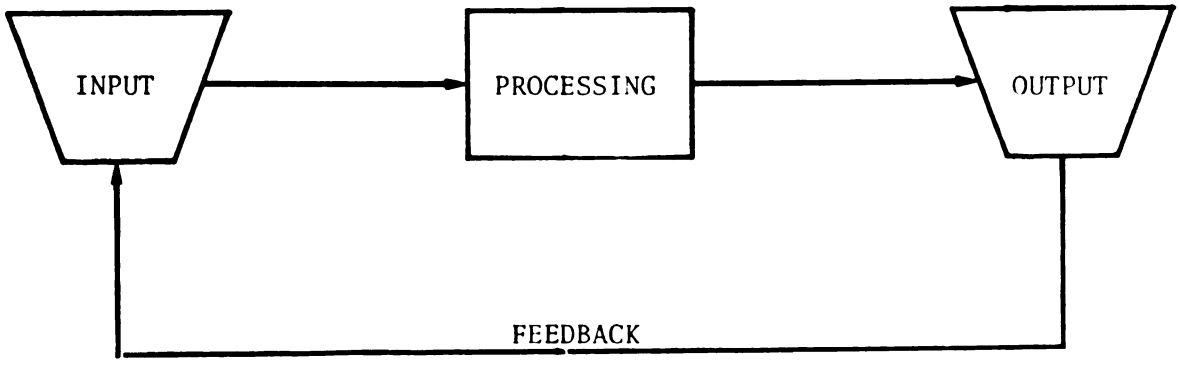


FIGURE 2.2

INPUT, PROCESSING, AND OUTPUT WITH FEEDBACK

entitled "Design of Basic Components of an Information Retrieval System" which also includes input, processing, and output. Under each of these three major components she has three subcomponents; materials, personnel, and equipment. She then proceeds to pose a series of questions for each of the categories which are designed to aid individuals in developing their own retrieval systems.[49] Input, processing, and output, with or without feedback for systems modification are perhaps as universal to data processing as any other single concept. This is presently illustrated by the categories found on an IBM Flowcharting template which includes symbols for input, output, various processing procedures, and feedback via program modification.[50]

Meadow states, "Information retrieval is the process of recovering information-bearing symbols from storage places in response to requests from prospective users of information or from libraries on the users' behalf."[51] Moreover, Artandi believes:

> Document retrieval systems may be viewed as consisting of four major elements: input to the system, the file(s) that (is/are) searched, searching methods, and output of the system. While each of these four elements is an essential part of any effective system, they are subject to differences in emphasis in various systems and there are differences in the theories and techniques that relate to them.[52]

Vickery similarly recognizes a number of components and channels between the user of the information and the store. He states,

---

[49]Claire K. Schultz, "Do-It-Yourself Retrieval System Design," Special Libraries, LVI (December, 1965), 721.

[50]IBM Flowcharting Template, Form X20-8020.

[51]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, 1967), p. 3.

[52]Susan Artandi, An Introduction to Computers in Information Science (Metuchen, N. J.: Scarecrow Press, Inc., 1968), p. 20.

"Retrieval is concerned with the structure and operation of devices to select documentary information from a store in response to questions."[53] A store can be a library, abstract journal, textbook, etc. A retrieval device can be an index in a book, library catalog, mechanical selector, or an electronic data processing device.

These definitions could be applied to Figure 2.1 by classifying input as the user request for information, processing as retrieval of information from the store, and output as the information which is provided to the user making the request. If to this analogy we add the classification of feedback which is concerned with the effectiveness of the information retrieved as well as the cost of retrieving the information we have in a simple form the components that are involved in the analysis, design, and redesign of information retrieval systems. As the components of various real or ideal information systems are examined, these elements will occur repeatedly with the modification that each of these broad components may involve multiple interacting sub-components.

Both Meadow and Vickery view information retrieval as part of a broader system of communication.[54] Meadow states ". . . information retrieval is part of a complex communication system existing between the authors of information-bearing documents and their readers."[55] In a diagram that Meadow states is highly over simplified he identifies a number of components and interactions between components of the

---

[53]B. C. Vickery, On Retrieval System Theory (London: Butterworths, 1965), p. 2.

[54]Ibid., p. 1; and Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, 1967), p. 3.

[55]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, 1967), p. 3.

information retrieval process as it relates to a library. The components are:

1. Authors and publishers generating documents.
2. Library management.
3. Indexers.
4. Indexing files.
5. Document files.
6. Search Assistance.
7. Patrons or users.

Among the interactions are:

1. Library management acquisitioning documents from authors and publishers.
2. Coordination of user needs and problems and interaction between library management and those involved in search assistance.
3. Coordination, classification and indexing techniques and interaction between library management and indexers.
4. Search requests from the patrons to those involved with search assistance.
5. Miscellaneous interactions relating to building index files and document files, and using these files to retrieve information for the patron or user.[56]

With the proper substitutions for example, computerized files for manual files, computerized operations for some of the operations done by indexers and information management for library management, the basic components could be generalized to the operation of various types of information retrieval processes.

---

[56]Ibid., p. 11.

Vickery identifies ten components of information retrieval which are very similar to the components and interactions found in the diagram provided by Meadow. The ten components identified by Vickery are:

1. A document store.
2. A description file.
3. A mechanism for indexing.
4. A mechanism for storing.
5. A mechanism for filing.
6. A mechanism for formalizing queries.
7. A mechanism for selecting appropriate documents to be retrieved.
8. A mechanism for retrieving the appropriate documents.
9. Rules for bibliographic description.
10. Rules for subject description.[57]

While not included in this list of ten components, he does mention in construction of an information system it is first necessary to select documents for inclusion in the store. With the inclusion of this element and substitutions of appropriate synonyms, the similarities are striking between the components found in the diagram suggested by Meadow, the list provided by Vickery, and a diagram suggested by Lancaster.[58] The major difference appears to be that the components suggested by Vickery and Lancaster pay less attention to the operation

---

[57]B. C. Vickery, On Retrieval System Theory (London: Butterworths, 1956), p. 11.

[58]Ibid.; Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, 1967), p. 11; and F. Wilfrid Lancaster, Information Retrieval Systems (New York: John Wiley & Sons, Inc., 1968), p. 4.

of an information depository than the components included in the diagram by Meadow.

Salton provides a table which describes various types of information centers and the differences in their functions.[59] In this table he refers to the functions described by the previous sources as well as to suggest functions that relate to the integration of information including bibliographies and analytical studies.

Kochen provides models for information services which use means of dissemination other than a user asking questions and receiving documents as a response. Included in Kochen's suggestions are procedures for evaluation and synthesis of information, tutorial information service, and standing request lists for specific types of information to aid in current awareness.[60]

Kochen's work is primarily an examination of the searching and dissemination process and does not consider the total information system in its context. Failure to examine a component in the context of a total system was typical of the majority of the literature reviewed. Aside from the articles by Kochen which focused on the retrieval process, works cited in this section were primarily chosen because they relate components and their interaction to the whole rather than focusing upon functions independently. In reviewing other literature, it appears that the components incorporated in the works cited here include most, if not all, of those components mentioned by articles and

---

[59]Gerard Salton, Automatic Information Organization and Retrieval (New York: McGraw-Hill Book Company, 1968), p. 6.

[60]Manfred Kochen, "Systems Technology for Information Retrieval," The Growth of Knowledge, Manfred Kochen, editor (New York: John Wiley & Sons, 1967), pp. 352-372.

books discussing specific functions. Not directly included in an overall description of most information systems was feedback or the evaluation necessary for continuing system modification, redesign and improvement.

Indexing Methods--Content Analysis, Specification, and Control

Vickery indicates that the key operation in retrieval is the description of what the documents are about. "This is the point at which research is most urgently needed for it is on adequate description that all ensuing operations in retrieval must rest."[61] Similarly Fairthorne has commented that indexing is "the basic problem as well as the costliest bottleneck in information retrieval."[62]

The first four volumes of the Annual Review of Information Science and Technology[63] discussed indexing methods under chapters entitled, "Content Analysis, Specification, and Control." These three terms express what some assert are the components of indexing: content analysis, the process of determining what a document is "about"; specification, a process of assigning indexing terms to describe the document; and control, the process of establishing and regulating the form and semantics of the descriptive labels making up the indexing

---

[61]B. C. Vickery, On Retrieval System Theory (London: Butterworths, 1965), p. 36.

[62]R. A. Fairthorne, Towards Information Retrieval (London: Butterworths, 1961), p. 136.

[63]Carlos A. Cuadra, editor, Annual Review of Information Science and Technology, Vols. I-IV, (New York: John Wiley & Sons, 1966-69).

language used for specification.[64]

Content analysis, whether done by indexers or by a computer is dependent upon the indexing language which controls the procedures used for the specification of indexing terms. For example, faceted classification languages aid the indexer in the analysis of the documents through grouping similar terms, and control the language by lists of terms with rules indicating how they are to be assigned to the different facets.[65] If the content analysis is computerized, the algorithms which are used to analyze the documents (determine what they are about and specify descriptive terms) are in essence an indexing language. The control of this language is established through the computerized algorithms and in some cases the interaction of these algorithms with authority lists stored in the computer.

Indexing Languages and Retrieval Systems

Vickery in discussing the Cranfield studies states that an indexing language

> . . . is significant in determining the performance of a retrieval system, not only as a result of actual intellectual arrangement of indexing, but also very significant on the output side. It is very significant in the question analysis and definition. So my point in this connection is simply that one cannot tear a classification or an indexing language out of the context of a total retrieval system.[66]

---

[64]P. Baxendale, "Content Analysis, Specification and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor, (New York: John Wiley & Sons, 1966), I, 71; and John R. Sharp, "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, 1967), II, 87.

[65]B. C. Vickery, Faceted Classification Schemes (New Brunswick, N. J.: Rutgers University Press, 1966), p. 1-108.

[66]Ibid., p. 15.

This section will attempt to categorize the various types of indexing languages and relate these to information retrieval systems having computerized components. Meadow suggests eight categories of languages: (1) hierarachical classification, (2) subject headings, (3) fixed key words, (4) free key words, (5) tagged descriptors, (6) faceted terms, (7) phrases, and (8) natural languages.[67] The first four of these range from very structured to very unstructured languages which have no rules for syntax while the next four languages are grouped according to increasing sophistication of syntax.

These categories are expressed in different ways by other authors and by no means represent the total possibilities; for example, Vickery talks about faceted classification schemes[68] which includes a combination of what Meadow describes as a hierarchical classification which has no syntax with a faceted structure which does have syntax--thus a combination of categories 1 and 6.

Some indexers might not consider the categories of "free key words" and "natural languages" as true indexing languages because they do not control the indexing terms which may be assigned. Compared to the six remaining "controlled" language categories, Hyslop assigns controlled languages to three categories: (1) classification schemes, (2) subject heading authority lists, and (3) thesauri.[69] Hyslop comments that

---

[67]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), p. 47.

[68]B. C. Vickery, Faceted Classification Schemes, (Vol. V of Systems for the Intellectual Organization of Information, ed. Susan Artandi (New Brunswick, N. J.: Rutgers University Press, 1966), pp. 1-108.

[69]Margorie R. Hyslop, "Sharing Vocabulary Control," Special Libraries, December, 1965, p. 708.

"there are innumerable specialized vocabularies of hybrids of these three types that defy any attempt to force them into general categories."[70] As can be seen while attempts have been made to place indexing languages into neat compartments there is not agreement upon the categories, and new cagetories can be made up through combinations of existing ones as indicated by Vickery's faceted classification schemes.

A detailed description of the various categories of indexing languages is beyond the scope of this review; however, an attempt will be made to describe briefly some of the categories mentioned.

Classification Schemes or Hierarchical Classification Schemes

Classification schemes are highly structured and show word associations by means of a hierarchy or family tree which leads the indexer from general terms at the top of the tree to successively more specific terms in succeeding lower levels. Usually there is a numerical or alphanumeric code which defines the unique term (branch of the tree) which can be translated into a unique word description. Examples of such indexing languages are the Library of Congress system, the Dewey Decimal Classification system, and the Universal Decimal Classification (UDC), a modification of the Dewey Decimal system.[71]

These systems commonly have at least three tools for the construction of indexes. First is a classification schedule which provides a visual map of the conceptual structure used to design the tree or

---

[70]8Ibid.

[71]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), pp. 22-25; and Marjorie R. Hyslop, "Sharing Vocabulary Control," Special Libraries, December, 1965, pp. 708-714.

hierarchy. The schedule is ordered according to the hierarchy established by the classification scheme. Second is an alphabetical index to the terms in the classification schedule. Each term that appears in the classification schedule appears in alphabetical order with its numerical code identifying its position in the classification scheme. Third is a set of rules which usually describe the structures of the classification scheme, its alphabetical index and how they are to be used in selecting terms and applying notation symbols to documents.[72]

Subject Headings or Authority Lists Subject or authority lists usually consist of alphabetical lists of words and/or phrases which are acceptable to use as indexing terms. If the list contains terms with more than one word, the term may appear alphabetically listed under each word in the term. In this manner, the array provides for limited word association by bringing together all terms containing a given word.

There are many possible variations on such indexing languages, including sets of rules which may establish various levels of indexing such as main entry terms plus various levels of subterms. If the list contains only single word terms it is sometimes called a key word index. While some specific indexing languages in this category contain various levels of indexing terms with rules for specification, they do not contain an alphanumeric code that is associated with the highly structured hierarchical classification schemes.[73]

---

[72]B. C. Vickery, Faceted Classification Schemes (Vol. V of Systems for the Intellectual Organization of Information, ed. Susan Artandi (New Brunswick, N. J.: Rutgers University Press, 1966), pp. 40, 41.

[73]Marjorie Hyslop, "Sharing Vocabulary Control," Special Libraries, December, 1965, p. 708; and Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), pp. 25-33.

The authority lists and classification schemes would be called precoordinate indexing languages, in that all of the word associations allowable have been established (precoordinated) before they are used by the indexer. However, the authority list allows for the addition of new terms with greater ease than the more highly structured classification schemes.[74] The fact that an authority list does not require that the total system is designed in advance is one of its advantages. As perceived by Meadow subject headings represent a ". . . loosening of the structure of a hierarchical language. Their use makes initial language design easier since there is less to predict, and makes future changes easier to implement because no elaborate structure need be disturbed by such a change."[75]

Thesauri Indexing Languages Indexing languages which use a thesaurus are a natural extension of subject heading languages. As with the subject heading languages they have an alphabetized list of words and/or phrases which may be used as indexing terms. However, associated with each term is a list of narrow terms, broad terms, and related terms. This provides a limited hierarchical structure and allows terms which are related to one another to be displayed together.[76] Hyslop says that, "although the hierarchies are not so discreetly displayed, they go beyond the confines of traditional classification array by

---

[74]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), pp. 25-30.

[75]Ibid., p. 26.

[76]Marjorie R. Hyslop, "Sharing Vocabulary Control," Special Libraries, December, 1965, pp. 708-710.

permitting any term to appear in as many hierarchies as may be appropriate. It is thus the more versatile of the three types of vocabularies in showing word association,"[77] Figure 2.3 is an example taken from the ERIC Thesaurus which shows five major terms and how they are displayed, with related terms designated by an RT, broad terms designated by a BT, narrow terms designated by an NT, and "Use For" designated by a UF.[78]

Free Key Word and Key Phrase Indexing The hierarchical classification, subject headings, and thesauri indexing languages are fixed in that the number of subjects which can be described is equal to the number of defined terms. These languages". . . are often called 'pre-coordinated' systems, in that whatever semantically meaningful descriptor combinations are allowed, have been made--the descriptors 'coordinated' to form terms by language designers."[79] The major distinction that exists between free key word or free key phrase indexing and pre-coordinated languages is the point where coordination (words or phrases grouped to form indexing terms) takes place. In the free key word or free key phrase languages this coordination can be done by the indexer or a computer, thus allowing the opportunity to form new terms (combinations of words) at the time that the indexing is done. In these languages the classes of words that may be used are generally not

---

[77] Ibid., p. 709.

[78] Thesaurus of ERIC Descriptors, Bethesda, Maryland: ERIC Processing and Reference Facility, Operated for U.S. Office of Education by Leasco Systems & Research Corporation, 1970), p. 82.

[79] Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), pp. 29, 30.

FAMILY ENVIRONMENT 160

UF Home
Home Conditions
Home Environment
BT Environment
RT Family (Sociological Unit)
Family Influence
One Parent Family
Permissive Environment

FAMILY FACTORS
Use Family (Sociological Unit)

FAMILY HEALTH 250

BT Health
RT Family (Sociological Unit)
Homemaking Education

FAMILY INCOME 220

BT Income
RT Family (Sociological Unit)
Family Resources
Family Status

FAMILY INFLUENCE 490

UF Home Influence
NT Fatherless Family
RT Family (Sociological Unit)
Family Counseling
Family Environment
Family Status
Fatherless Family
Motherless Family
One Parent Family
Parental Aspiration
Parent Attitudes
Parent Participation
Parent Reaction
Parent Role

FIGURE 2.3

A PORTION OF THE ERIC THESAURUS[80]

---

[80] Thesaurus of ERIC Descriptors, ERIC Processing and Reference Facility, Operated for U.S. Office of Education by Leasco System & Research Corporation, (Bethesda, Maryland: 1970), p. 82.

restricted except for exclusion of conjunctions, prepositions, articles and other non-content words.[81]

The major advantage of such a system is the ease with which new terms may be made up while a major disadvantage is a lack of control to aid the searcher and indexer in using the same language.[82]

Indexing Languages with Syntax Webster's Third New International dictionary, 1967 edition, defines syntax in three ways:

> 1: connected system or order : orderly arrangement : harmonious adjustment of parts or elements 2a: sentence structure : the arrangement of word forms to show their mutual relations in the sentence b: the part of grammer that treats of the expression of predicative, qualifying, and other word relations according to established usage in the language under study--compare MORPHOLOGY 3a: SYNTACTICS b: the area of Syntactics dealing specifically with the formal properties of languages or calculi--called also logical syntax.

An examination of these definitions might suggest that indexing languages having syntax would be able to show word relations or the role of specific words by the arrangement of the language (perhaps by where a specific word appears in a string of words) or through modifying symbols which are linked to specific terms.

Meadow describes languages which used tagged descriptors. ". . . a descriptor has affixed to it another descriptor to describe the first. The role of the affixed might be to classify the basic descriptor, denoting it as a proper name or attribute, or an activity."[83]

---

[81]F. Wilfrid Lancaster, Information Retrieval Systems (New York: John Wiley & Sons, Inc., 1967), pp. 21, 30.

[82]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1968), pp. 33, 34.

[83]Ibid., p. 33.

Vickery speaks of faceted classification schemes where the role of each facet might be identified by what he calls "facet indicators (somewhat comparable to role indicators)."[84]

A second type of syntactic mechanism which can be used in indexing languages is to have an indexing string of terms where the position of a given term in the string indicates the role that it plays. For example, in an inventory system, successive terms in a record might play the following roles: (1) item name, (2) style, (3) color, (4) quantity, (5) on hand, (6) unit price, and (7) total value. A system of this type might be designed in such a way that any term in the string could be used to arrange the total file in a specified alphabetic or numeric order in this illustration each one of the terms in the descriptive record would be called a facet.[85]

Another manner in which roles or facets may be defined is by indicating the portion of a document from which a term was extracted. This procedure is possible in the BIRS system by using field names; for example, searches may be done which indicate that a person is looking for the term "Brown" from the author field and that it has no meaning if it is in the descriptor field. In this cause the searcher wants an author by the name of Brown and finding the color "brown" in a portion of the text, the title, or the descriptor field would not be of value.[86]

---

[84]B. C. Vickery, Faceted Classification Schemes, (Vol. V of Systems for the Intellectual Organization of Information, ed. Susan Artandi. New Brunswick, N. J.: Rutgers University Press, 1966), p. 58.

[85]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), pp, 33, 34.

[86]John F. Vinsonhaler and John M. Hafterson (editors), Technical Manual for Basic Indexing and Retrieval System, BIRS 2.5, Appendix I (East Lansing: Educational Publications Services, College of Education, Michigan State University, January, 1969), pp. 2601-2641.

Another use of syntax also available in the BIRS system is the extraction of phrases from the text of a document or document surrogate. In this procedure all words are extracted except those contained in an exclusion list thus maintaining their positional (modification) relationships with the other words in the sentence. In this way, it is possible to search for terms with the syntax that is established by natural language through co-occurrence of terms.[87]

In general indexing languages involving syntax might be classified as languages where (1) the syntax occurs because of rules which are part of the indexing scheme and (2) languages where the syntax is the result of grammatical structure of the text from which the terms are extracted.

Methods of Machine Indexing

Indexing languages may be used by machines, human beings, or machine-human combinations in specifying the terms which describe what a document is "about." The variety of procedures used in machine indexing is illustrated in a state-of-the-art report by Stevens which includes a 662-item bibliography.[88] In a discussion of automated indexing Borko describes four types of procedures: (1) Statistical indexing, (2) permutation indexing, (3) citation indexing, and association indexing.[89] The discussion which follows includes these four categories and an added category for procedures which extract and

---

[87] Ibid., pp. 2401-2419.

[88] M. E. Stevens, Automatic Indexing: A State of the Art Report NBS Monograph 91 (Washington: National Bureau of Standards, March 1965).

[89] Harold Borko, Automated Language Processing (New York: John Wiley and Sons, Inc., 1967), pp. 100-114.

assign terms by algorithms which may use neither syntactical nor statistical analysis.

Statistical Indexing Luhn suggests a statistical procedure for determining the significance of words as they relate to the frequency of use. In his discussion he notes that words which occur very frequently usually have little descriptive significance and are often conjunctions, articles, or prepositions. It is also noted that words which occur very infrequently may also have little significance because they are not commonly used, are misspellings, or are infrequently-used synonyms for terminology more commonly used in the literature.[90]

Meadow, in discussing Luhn's model hypothesizes that "words are significant as subject descriptors, then, in proportion to the difference between their actual and expected frequency."[91] The Descriptive Analysis Program of the 2.0 version of BIRS provides one of the better examples of how this hypothesis can be applied in the design of computer programs.[92]

Simmons, et. al. describe a system which is divided into two portions, a program called Indexer which is used for indexing the full text and a program called Protosyntax I which is used in searching. In the example used to illustrate this system the total text of the Golden

---

[90]H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, II (1958), 159-165.

[91]Charles T. Meadow, The Analysis of Information Systems (New York: John Wiley & Sons, Inc., 1967), p. 100.

[92]John F. Vinsonhaler (editor), Technical Manual for Basic Indexing and Retrieval System, BIRS 2.0 (East Lansing: Educational Publications Services, College of Education, Michigan State University, 1968), pp. 1201-1221.

Book Encyclopedia was indexed. The procedures used by Indexer to process the text extracted all words not included in a list of approximately 300 articles, prepositions, conjunctions and other non-content words. The frequency of occurrence of each content word was determined and space allowed to assign sufficient VAPS numbers to identify the location of each occurrence of the word in the text. (VAPS numbers indicate the Volume, Article, Paragraph, and Sentence where a term occurred.)

The search program used the word index with its VAPS numbers to identify areas of the text that were likely to have a relationship to specific questions stated in English. Finally, the content and syntax of these portions of text were compared with the question to determine what part of the text should be identified as potentially useful.[93]

The search program was included in the above discussion of indexing techniques because a major portion of the content analysis was done by this program. Because the search program uses syntactical techniques, the inclusion of this example under the heading of statistical indexing may not be entirely satisfactory, but was done on the basis of Borko's example[94] and the fact that the program producing the index used only statistical and word extraction techniques.

Non-Syntactical, Non-Statistical Term Specification There are methods of computerized indexing that would be difficult to place in

---

[93]Robert F. Simmons, and Keren McConlogue, "Maximum Depth Indexing for Computer Retrieval of English Language Data," American Documentation, January, 1963, pp. 68-73; and Robert F. Simmons, Sheldon Klein, and Keren McConlogue, "Indexing and Dependency Logic for Answering English Questions," American Documentation, July, 1964, pp. 196-204.

[94]Harold Borko, Automated Language Processing (New York: John Wiley and Sons, Inc., 1967), pp. 100-104.

either a statistical or syntactical category. These methods are generally based upon algorithms which use an authority list for extracting terms and, in some cases, for substituting appropriate synonyms.

One such method described by Moon and Vinsonhaler is based upon the assumption that the terms found in the titles of scientific articles have high descriptive value. The first step in the procedures was to generate an authority list of all descriptive terms found in the titles of the document file. A descriptive term was defined to be any term that did not appear on an exclusion list which contained articles, conjunctions, prepositions, and other words that were considered to be of little content value. The second step was to index the document surrogates by extracting from desired portions of the text all terms which were both in the authority list and in specified portions of text. In the example cited, the portions of text used were the title and abstract of articles.[95]

Artandi describes a system used in medical articles that extracts all terms having characteristics which are considered to be unique to terms having descriptive content. The characteristics defined in the medical project were, ". . . length of the character strings (organic compounds have long names); an alternating string of numbers, letters, and dashes; the presence of Greek letters in the strings and the

---

[95]R. D. Moon and John F. Vinsonhaler, "The Title-Generated Thesaurus: A Practical Method for Automated Indexing," in Shultz, L. (ed.), Proceedings of the Sixth Annual National Colloquium on Information Retrieval - The Information Bazaar. The Medical Documentation Service of the College of Physicians at Philadelphia, 1969.

presence as part of the name such words as ethyl, methyl, propyl, etc."[96] She further states that the indexing algorithm is to satisfy the following requirements:

> . . . to recognize information in the text that should be indexed, to switch from a variety of text words to a controlled vocabulary, to create a standardized index record, to compute and assign weights to the index terms automatically, to create valid links between index terms, and to provide for expandability.[97]

The two indexing schemes cited are typical procedures that do not rely on the statistical frequency or syntax to control term specification, but instead identify terms with high content value by their location in the text (for example terms in titles, subheads, abstracts, conclusions, summaries, etc.), their characteristics, comparing them with an authority list, or through a combination of these.

While the procedures described by Artandi do not use statistical or syntactical analysis of the text to extract terms, techniques related to co-occurrence were applied to the terms extracted. An automatic algorithm generated links based on the assumption that "co-occurrence within a sentence is a satisfactory indication that the terms belong together within the context of the document"[98]

In the evaluation of this technique it was found that the closer the terms occurred together within a sentence, the greater the probability that the terms actually should be linked together. The data

---

[96]Susan Artandi, "Computer Indexing of Medical Articles--Project MEDICO," Journal of Documentation, September, 1969, p. 218.

[97]Ibid., pp. 214-223.

[98]Susan Artandi and Edward H. Wolf, "The Effectiveness of Automatically Generated Weights and Links in Mechanical Indexing," American Documentation, July, 1969, pp. 198-202.

indicated that the average number of terms between links that were judged relevant was 3.71 words while the average number of terms between links that were judged irrelevant was 7.08 words.

Procedures also automatically assigned weights based on the frequency of occurrence to words which possessed characteristics described in the extraction criteria. (It should be noted that no frequency or statistical measures were used in specifying which term could be used in indexing.) The evaluation indicated that weights assigned to terms by manual indexers were in agreement 71% of the time with those assigned by automatic procedures and that 72% of the links automatically assigned because of full text scanning were considered relevant.[99]

The above system illustrates the difficulty of attempting to classify automated indexing methods; for while the procedures for term specification did not use statistical or syntactical methods, additional procedures which weighted terms by frequency and/or linked terms by co-occurrence within a sentence were included in the total indexing scheme.

Word Association, Co-Occurrence, Links and Roles One weakness in some systems which extract key words from text is that they fail to maintain links between these words and the context from which they were extracted. When searching is done on these systems irrelevant documents are sometimes retrieved because the words used by the searcher do not play the roles expected. For example, a searcher might formulate a question that reads "train and coach" with the intention of retrieving information about "train coaches," but instead receives information on

---

[99] Ibid., p. 202.

how to train football coaches. If the words had been extracted in a manner that the words "train" and "coach" were mechanically linked or co-occurred in the phrase "train coach," this difficulty could have been avoided.

The problem described relates to precision, i.e. retrieving documents that are not relevant to the question and is sometimes solved by mechanically linking words together, by extracting phrases or co-occurring words, or attaching tags to words which indicate their role.

Doyle suggests a procedure for using statistics of word co-occurrence in the analysis of documents. He also suggests means by which association maps can be developed for frequently co-occurring words and presents two methods for using these maps in literature searching. As used by Doyle, co-occurring words are words which appear together (co-occur) in the text and association maps are maps which graphically represent relationships between words, developed through statistical analysis of co-occurrence. An analysis of text reveals that some words co-occur with many different separate words whereas some words may co-occur with only a few other words. This observation suggests that special significance may be placed upon words co-occurring with many other words. This is graphically displayed in Doyle's association maps.[100]

Borko indicates that computer programs exist which can analyze the co-occurrence of words and automatically draw association maps.[101] Dale

---

[100]Lauren B. Doyle, "Indexing and Abstracting by Association," American Documentation, October, 1962, pp. 378-390.

[101]Harold Borko, Automated Language Processing (New York: John Wiley & Sons, Inc., 1967), pp. 112-114.

and Dale also describe a retrieval model using association of words, or "clumping," which has been programmed for a digital computer. Based on the results of experiments on a small document set they suggest that the technique shows promise for larger document collections.[102]

Baxendale describes linguistic experiments at IBM that use syntactical procedures to identify and extract from selected portions of the document--such as the title, diagram captions, paragraph headings, and sentences of abstracts--words associated together as noun phrases. The sentence, "Since the _____ _____ was _____ _____ by the _____ _____, all _____ _____ must be _____," was used to illustrate how the syntax can predict where noun phrases might appear.[103]

The Descriptive Analysis Program of BIRS has a number of options available to the user including the ability to extract phrases in a way which obtains a result that has some similarity to the result described by Baxendale. By using an option to extract from selected portions of documents all words from sentences except those appearing on an exclusion list, it is possible to develop word strings to serve as descriptors.

To illustrate how this option may be used Figure 2.4 uses the sentence "The extraction of key words and phrases from selected portions of the text is very important to automated indexing," with an exclusion list which includes the words, the, of, and, from, is, very, and to. In Figure 2.4 the sentence is first shown with the words

---

[102]A. G. Dale and N. Dale, "Some Clumping Experiments for Associative Document Retrieval," American Documentation, January, 1965, pp. 5-9.

[103]P. B. Baxendale, "Autoindexing and Indexing by Automatic Processes," Special Libraries, December, 1965, p. 718.

underlined which would be extracted; then the sentence is shown with these words removed to illustrate the similarity to the approach described by Baxendale; and finally the extracted terms are shown as they would be recorded on a BIRS description file.

The extraction of words and phrases from selected portions of text is very important to automated indexing.

The _____ of _____ and _____ from _____ _____ of ____ is very _____ to _____ _____.

$extraction, words, phrases, selected portions, text important, automated, indexing$

FIGURE 2.4

ILLUSTRATION OF BIRS WORD EXTRACTION TECHNIQUES

The BIRS Description File Search Program allows phrases of up to ten words to be matched with these strings; however, the match must occur between $'s--in this case within a given sentence. If it is desirable to reduce the length of the string where a match can occur, it is possible for the Descriptive Analysis Program to insert additional $'s at punctuation marks such as commas, colons, and semicolons.

Procedures described in this section show how both statistical and syntactical methods can be used to extract words in ways that will help to maintain some of their original relationships. In the methods suggested by Doyle these procedures were used to develop association maps involving only word pairs. In methods described by Baxendale and in those used by the BIRS Descriptive Analysis Program it is possible to have words associated together in groups larger than pairs. Procedures suggested by Doyle and those described by Dale and Dale used statistical techniques; procedures described by Baxendale used a linguistic approach

which considered the syntax of the language, and the procedures described in the Descriptive Analysis Program used an exclusion list which could be developed to consider language syntax. In these methods words are linked by occurring together with their grammatical roles designated by their context.

Document Association and Citation Indexing Another type of association indexing relates to the clumping or grouping of documents according to the similarity of content. Jones and Needham report on a computerized program which applies the automated classification techniques associated with the "series of clumps" to document descriptions obtained from the ASLIB Cranfield Project. This particular program examines co-occurrence of terms based upon their appearing in different document descriptions rather than the text of the document. This results in clustering of related documents rather than a clustering of words. The authors indicate the evaluation of the programs and procedures is still being carried out and consequently draw no definite conclusions on the value of these procedures.[104]

A similar technique is reported by Perry which is based on "inclusion relationships existing between sets of features assigned to the document." This technique used sets of features to identify documents that have been indexed "by all or by only some of the total features of the sets." Perry calls the resulting index a "combined group

---

[104]K. Sparck Jones and R. M. Needham. "Automatic Term Classifications and Retrieval," Information Storage and Retrieval, June, 1968, pp. 29-31.

co-ordinate index," and indicates that where used on one set of holdings there have been demonstrated advantages.[105]

The above two techniques utilized indexing terms assigned to specific documents to generate the clumps of similar documents. Another technique, citation indexing, employs the references cited by an article to generate clumps of similar documents or do computer searching. In most reviews of literature a person will use citations only in a historical sense, i.e., he may find an article which is very relevant to the information he is seeking and examine the references cited in this article to obtain other articles. When this procedure is followed, all of the articles obtained from the reference will be chronologically older than the original reference.

It is often desirable to move chronologically forward in the literature by examining all articles which have cited a pertinent article. This is possible with computer programs and involves a procedure where an article is described by the articles it cites; i.e., the descriptive terms assigned to the article are citations to other articles. With this type of indexing the name of a specific article may serve as a computer search question to retrieve all articles which have cited that article.[106] Price and Schiminovich describe a study where a computer program was used to do bibliographic coupling (citation indexing), and they indicate that the clustering process evaluated appeared adaptable

---

[105]Peter Perry, "Combined Grouping for Coordinate Indexes," American Documentation, April, 1968, p. 142.

[106]Harold Borko, Automated Language Processing (New York: John Wiley and Sons, Inc., 1967), p. 108-112; and Charles T. Meadow, The Analysis of Information Systems (New York; John Wiley & Sons, Inc., 1967), pp. 86, 119, 120.

to future use in developing a computer generated classification scheme.[107]

Permuted Indexing The term "permuted indexing" is commonly used interchangeably with "Key-Word-In-Context" or KWIC indexing. In 1960 Luhn suggested that an index allowing the user to see key words in their context would be of value for disseminating new information. He also described procedures whereby terms might be extracted from machine-readable documents or document surrogates and a KWIC index generated automatically by computers.[108]

The term "permuted index" is descriptive of the format generated by many computer programs which do this type of indexing. Figure 2.5-A provides an illustration of how an article titled "The KWIC Index Concept" would be permuted and Figure 2.5-B then shows how it would be alphabetized and formatted. If the title "Indexing Consistency and Quality" were also included as part of the index, the merged output generated by the two phrases would appear as illustrated in Figure 2.5-C.

Also included, but not shown in Figure 2.5-C would be information identifying the document from which the titles were extracted. Luhn suggested an 11-character code which contained information concerning the name of the author or senior author, the year of publication, and the title of the document.[109] If the documents in a KWIC index were part of

---

[107]Nancy Price and Samuel Schiminovich, "A Clustering Experiment: First Step Towards a Computer-Generated Classification Scheme," Information Storage and Retrieval, August, 1968, pp. 271-280.

[108]H. P. Luhn, "Keyword-In-Context Index for Technical Literature," American Documentation, XI (1960), 288-295.

[109]Ibid., p. 271.

A

THE KWIC INDEX CONCEPT
THE KWIC INDEX CONCEPT
THE KWIC INDEX CONCEPT

B

THE KWIC INDEX CONCEPT
THE KWIC INDEX CONCEPT
THE KWIC INDEX CONCEPT

C

THE KWIC INDEX CONCEPT
INDEXING CONSISTENCY AND QUALITY
THE KWIC INDEX CONCEPT
INDEXING CONSISTENCY AND QUALITY
THE KWIC INDEX CONCEPT
INDEXING CONSISTENCY AND QUALITY

D

CONCEPT	1	THE KWIC INDEX CONCEPT
CONSISTENCY . .	2	INDEXING CONSISTENCY AND QUALITY
INDEX	1	THE KWIC INDEX CONCEPT
INDEXING . . .	2	INDEXING CONSISTENCY AND QUALITY
KWIC	1	THE KWIC INDEX CONCEPT
QUALITY	2	INDEXING CONSISTENCY AND QUALITY

FIGURE 2.5

EXAMPLES OF PERMUTED OR KEY-WORD IN CONTEXT (KWIC) INDEXES

an information file as in the BIRS system, this code could be replaced by an access number. The output illustrated in Figure 2.5-D is similar to the output the BIRS System would generate for a KWIC index if the access numbers 1 and 2 had been assigned to the two titles.

When this type of index is printed in the format of Figure 2.5-D rather than the permuted format of 2.5-C some call it a Key Word Out of Context (KWOC) index.[110] This does not seem entirely consistent with seeing the word displayed in its context. The term KWOC is used in the BIRS system to denote an index where the word or phrase appears with access numbers but without any context.[111]

The KWIC index has the advantage of allowing the user to see the syntactical structure of the language around the key word. It, however, has some obvious limitations; for example, if one were to apply this type of indexing to the total document, each line of the document might generate four to ten lines of output. Borko suggests that, to cope with this problem, titles or some other small portion of the text rich in descriptive words, might be the only part of the document indexed. To further reduce the amount of output he suggests excluding from the terms to be indexed classes of words such as conjunctions, prepositions, and articles.[112]

---

[110]Marguerite Fischer, "The KWIC Index Concept: A Retrospective View," American Documentation, April, 1966, pp. 63, 64.

[111]John F. Vinsonhaler and John M. Hafterson (editors), Technical Manual for Basic Indexing and Retrieval System, BIRS 2.5, Appendix I (East Lansing: Educational Publications Services, College of Education, Michigan State University, January 1969), pp. 2201-2219.

[112]Harold Borko, Automated Language Processing (New York: John Wiley and Sons, Inc., 1967), p. 104.

Trend in Machine Indexing In 1967 Sharp in the second Annual Review of Science and Technology noted that:

> . . . ten years later we seem to have reached a period of disenchantment, not with machine methods generally, but with the idea that it is going to be easy. The particular point we seem to have reached is the realization that statistical techniques for textual analysis are inadequate.[113]

Borko and Wyllys writing in the book Automated Language Processing are very candid about the problems that exist in using computers to extract descriptive terms from documents or in doing automated abstracting. Borko remarks, "Thus we see that automated classification clearly supplements but does not replace manual systems of classification." He further indicates that studies of automated indexing and classification are the results of "a real need" to improve the storage and retrieval of information and that while progress has been made much more needs to be done.[114] In summarizing what has been done related to automated abstracting, Wyllys comments, "If it seems that relatively little has been accomplished in the field, it should be realized that very few people have concerned themselves with automated abstracting.[115] While both Borko and Wyllys present interesting possibilities and point to a number of promising studies, neither indicate that there will be any dramatic solutions in the near future. Both present a picture which

---

[113]John R. Sharp. "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, (New York: John Wiley & Sons, Inc., 1967), II, 88.

[114]Harold Borko, "Indexing and Classification," Automated Language Processing (New York: John Wiley & Sons, Inc., 1967), pp. 122, 123.

[115]Ronald E. Wyllys, "Extracting and Abstracting by Computer," Automated Language Processing, Harold Borko, editor (New York: John Wiley and Sons, Inc., 1967), p. 160.

implies that solutions will come only through hard work.[116]

Salton takes a more positive view toward automated indexing than some. Observing a study which compared the National Library of Medicine's MEDLARS system which uses human indexers with the fully automatic SMART system he states,

> Fully automatic text analysis and search systems do not appear to produce a retrieval performance that is inferior to that obtained by conventional systems using manual document indexing and manual search formulations. While the manual indexing and search formulations can lead to exceptionally fine results when the indexer and/or searcher are completely aware of the relationships between the stored collection and the user needs, the search results are also very poor when the conditions are not met. The automatic process on the other hand, with its exhaustive input data and complex analysis methods performs very poorly only rarely, and may often produce completely satisfactory retrieval action.[117]

It should be noted that the SMART system is a laboratory system which utilizes "a variety of intellectual aids in the form of synonym dictionaries, hierarchical arrangement of subject identifiers, statistical and syntactical phrase generation methods and the like, in order to obtain content identification useful for the retrieval process,"[118] whereas MEDLARS is an operating system with a data base of over 500,000 documents.[119]

A significant change in the use of automated indexing may be

---

[116]Ibid., p. 127-179; and Harold Borko, Indexing and Classification," Automated Language Processing (New York: John Wiley & Sons, Inc., 1967), pp. 99-125.

[117]Gerard Salton, "A Comparison Between Manual and Automatic Indexing Methods," American Documentation, January, 1969, p. 70.

[118]G. Salton, E. M. Keen, and M. Lesk, "Design Experiments in Automatic Information Retrieval," The Growth of Knowledge, editor, Manfred Kochen (New York: John Wiley & Sons, Inc., 1967), p. 337.

[119]Gerard Salton, " A Comparison Between Manual and Automatic Indexing Methods," American Documentation, January, 1969, p. 70.

brought about through recent developments in hardware and software which permit more efficient use of on-line systems. Speaking to this point, Lancaster and Gillespie state,

> There appears to be a very dramatic re-awakening of interest (somewhat quiescent for a few years, with a notable exception of Salton's work) and the design of systems incorporating automated indexing, automated classification, or automated search elaboration. Undoubtedly this renaissance has been at least partially prompted by the availability of on-line processing capabilities.[120]

One of the major problems in analyzing large amounts of data has been the cost of putting this data into the computer. Commenting on this problem Taulbee states,

> Naturally all automatic indexing procedures depend upon the existence of some representation of the document in machine-readable form, but this should not be a particular difficulty as improved page readers become available and as more and more publications produce machine-readable copy as a by-product of the printing process. It is beleived that automatic indexing will be more economical and less time consuming.[121]

In attempting to cope with the large amounts of data and the complexities of content analysis, some large systems are utilizing indexers and abstractors to develop document descriptions which can be used for computer searching. Notable examples of such systems are the MEDLARS[122] and the ERIC[123] systems. In these systems the indexing terms and/or

---

[120]F. Wilfrid Lancaster and Constantine J. Gillespie, "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (Chicago: Encyclopedia Britannica, Inc., 1970), V, 39.

[121]Orrin E. Taulbee, "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (Chicago: William Benton, 1968), III, 120.

[122]F. Wilfrid Lancaster, "MEDLARS: Report on the Evaluation of its Operating Efficiency," American Documentation, April, 1969, pp. 119-142.

[123]Lee G. Burchinal, "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, April, 1970, p. 55-67.

abstracts are generated by human judgments with the help of various indexing or abstracting guidelines.

The problem which is commonly discussed concerning systems that employ human judgment is the effect of indexer inconsistency upon operating characteristics. Lancaster reporting on the Cranfield Project indicates that sixty per cent of failures to retrieve source documents could be traced to inconsistencies of indexing.[124] Zunde and Dexter cite various studies on indexing inconsistency where the percentage of common terms assigned by two indexers ranges from ten to eighty per cent of the total number of unique terms assigned by the indexers.[125] Many of the studies which examine indexer inconsistency do so by comparing results of two indexers, rather than examining the effect of the indexing upon precision and recall as done by the Cranfield Project.[126]

Cooper emphasizes the need to evaluate indexing. He is not concerned about how well human indexers agree with each other or with automated indexing, but rather about the effectiveness of a particular system in retrieving documents. He states, "The crucial question is therefore; what is the relationship if any, between the level of retrieval performance achieved when the indexing is done by that method?"[127]

---

[124]F. Wilfrid Lancaster, and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," American Documentation, January, 1964, p. 7.

[125]Pranas Zunde and Margaret E. Dexter, "Indexing Consistency and Quality," American Documentation, July, 1969, p. 60.

[126]F. Wilfrid Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," American Documentation, January, 1964, p. 7.

[127]William S. Cooper, "Is Interindexer Consistency a Hobgoblin?," American Documentation, July, 1969, p. 268.

The references cited in this section have been chosen because they appear to be representative of many possible citations relating to trends in automated indexing. These as well as other references appear to support that:

1. Progress in automated content analysis has not come as easily or as quickly as some had originally predicted.
2. There has been progress and there is continuing progress in automated content analysis.
3. Developments in hardware and software which have made interactive on-line systems more realistic and efficient have created a new interest in automated content analysis.
4. Some feel the continuing development and use of optical scanners and computerized typesetting will reduce the cost of putting large amounts of data into the computer for analysis.
5. Because the problem of inputting large amounts of data has not been totally solved, some large systems such as ERIC and MEDLARS are using human-machine partnerships where the indexing and abstracting necessary for computer searches is done by people.
6. Indexing methods should primarily be evaluated on how effectively the total system can retrieve documents.

Evaluation of Indexing Methods

In this section the factors involved in the evaluation of information systems are identified, and those procedures specifically related to evaluation of indexing methods are examined in greater detail. The procedures and descriptive statistics described in this section are

similarly discussed in a variety of sources. Closely related to this study is work done by Cleverdon as part of the Cranfield Project,[128] by Lancaster in the evaluation of MEDLARS[129] which built upon the Cranfield Project, and by Salton on the totally automated information retrieval system called SMART.[130] In many instances it would be possible to give multiple references for some of the statements which will be made; however, usually only a single reference will be given from sources related to one of these projects.

Salton identifies the following factors as being among those which are most important to information systems evaluation:

User population Type of user, rate of requests, etc.
Collection Coverage of collection, type of document available at input, reliability of abstracts, etc.
Indexing Type of indexers, level and accuracy of indexers, depth of indexing required, complexity of indexing languages, etc.

---

*128C. W. Cleverdon, F. W. Lancaster, and J. Mills, "Uncovering Some Facts of Life in Information Retrieval," Special Libraries, February, 1964, pp. 86-91; and Cyril Cleverdon, Jack Mills and Michael Keen, Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Part 1: Text, and Part 2: Appendices, (ASLIB Cranfield Research Project, Cranfield, Bedford, England, 1966), pp. 1-120 and 121-377.

*129F. Wilfrid Lancaster, "Evaluating the Performance of a Large Operating Retrieval System," Electronic Handling of Information, Allen Kent, Orrin E. Taulbee, Jack Belzer, and Gordon D. Goldstein, editors (Washington, D.C.: Thompson Book Company, 1967), p. 199-216; and F. Wilfrid Lancaster, Information Retrieval Systems (New York: John Wiley & Sons, Inc., 1968), pp. 1-217.

*130Gerard Salton, Automatic Information Organization and Retrieval (New York: McGraw-Hill Book Company, 1968), p. 6; Gerard Salton, "The Evaluation of Automatic Retrieval Procedures--Selected Test Results Using the SMART System" American Documentation, July, 1965, pp. 209-222; and G. Salton, E. M. Keen, and M. Lesk, "Design Experiments in Automatic Information Retrieval," The Growth of Knowledge, Manfred Kochen, editor (New York: John Wiley and Sons, Inc., 1967), pp. 336-351.

*See bibliography for additional material by each of these authors related to their projects.

Anaylsis and search Type of searching, power and complexity of search mechanism, search effort required, accuracy of search, etc.

Equipment and input-output Type of store, type of in-out equipment, type of form of output

Operating efficiency Cost considerations, service problems, time lag, and response time[131]

In choosing the factors that are most critical, Salton contends that the overriding consideration should be those factors that lead to user satisfaction, and that all other criteria are secondary or relate to this. Thus, criteria relating to the management of a system would be considered only in "relation to their effect on the user criteria."[132] The following criteria are similar to those identified by both Lancaster[133] and Cleverdon[134] as important to user satisfaction:

1. The relevance of the material contained in the information system to the user's overall needs.
2. The effort required by the user to request and obtain information from the system.
3. The average time that elapses between the time that a request is made and information is provided to the user.
4. The proportion of the total relevant material in the system which is retrieved in response to the user request (recall).

---

[131]Gerard Salton, Automatic Information Organization and Retrieval (New York: McGraw-Hill Book Company, 1968), p. 282.

[132]Ibid.

[133]F. Wilfrid Lancaster, Information Retrieval Systems (New York: John Wiley and Sons, Inc., 1968), pp. 33, 34.

[134]C. W. Cleverdon, Identification of Criteria for Evaluation of Operational Information Retrieval Systems, Cranfield College of Aeronautics, England, November, 1964.

5. The proportion of the total material retrieved in response to a user's request which is relevant to that request (precision ratio).

6. The form in which information retrieved by search requests is presented to the users.

In the evaluation of indexing methods the factors which would be of greatest importance would be those relating to a system's ability to retrieve documents in response to search questions. Of the factors discussed by Cleverdon and Lancaster, recall and precision ratios are the two criteria that relate most directly to indexing evaluation. Lancaster and Gillespie explain:

> Most investigators, from Cleverdon on, have expressed evaluation results in terms of twin variables of recall ratio (the number of relevant documents retrieved over the total number of relevant documents in the collection) and precision ratio (the number of relevant documents retrieved over the total number of documents retrieved), although both ratios are sometimes given other names in the literature. For most practical purposes, these ratios are perfectly adequate parameters for expressing the results of a search.[135]

Descriptive Statistics for Document Retrieval

When indexing is evaluated on the ability of search questions to retrieve documents, the descriptive statistics used involve four variables. Listed below are these variables which result from the partitioning of a document collection by a search.

a. Documents that are not retrieved and not relevant.

b. Documents that are not retrieved but relevant.

c. Documents that are retrieved and relevant.

---

[135]F. Wilfrid Lancaster and Constantine J. Gillespie, "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor, (Chicago: William Benton, 1970), V, 45.

d. Documents that are retrieved but not relevant.[136]

The partitioning illustrated by Figure 2.6 and the variables designated by a, b, c, and d will be the basis for the discussion of the statistics used to describe document retrieval.

Recall and Precision Ratios for a Single Search Question As related to Figure 2.6 the recall and precision ratios[137] for a single document are defined as:

$$\text{Recall} = \frac{\text{number of documents retrieved and relevant}}{\text{total relevant in collection}} = \frac{c}{b + c} \quad \text{(Formula 2-1)}$$

$$\text{Precision} = \frac{\text{number of documents retrieved and relevant}}{\text{total retrieved}} = \frac{c}{c + d} \quad \text{(Formula 2-2)}$$

Averages of Recall and Precision In evaluating indexing methods it is desirable to use the results of a number of searches to calculate an average value of precision and recall. In determining these averages one possible procedure gives equal weight to each question in a set of searches while a second gives equal weight to each document.

In the following formulas b, c, and d are used as illustrated in Figure 2.6; m stands for the total number of searches and i is an index number with different integer values from 1 to m to indicate results for different searches. For example, if i equaled 2 then c_2 would stand for the number of relevant documents retrieved by the second search. The method which gives equal weight to each search takes a

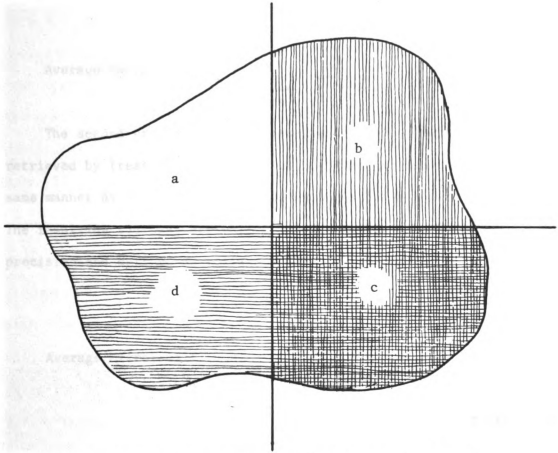
---

[136]Gerard Salton, Automatic Information Organization and Retrieval (New York: McGraw-Hill Book Company, 1968), pp. 282, 283.

[137]Ibid., pp. 283, 284.

Documents retrieved by the search question

Documents relevant to the search question



a = Documents that are not retrieved and not relevant

b = Documents that are not retrieved but relevant

c = Documents that are retrieved and relevant

d = Documents that are retrieved but not relevant

FIGURE 2.6

THE PARTITIONING OF A DOCUMENT COLLECTION BY A SEARCH QUESTION

mean of the recall ratios and precision ratios for each question with the averages referred to as Average Macrorecall and Average Macroprecision. Average Macrorecall and Average Macroprecision are defined as follows:[138]

$$\text{Average Macrorecall} = \frac{1}{m}\sum_{i=1}^{m}\frac{c_i}{b_i + c_i} \qquad \text{(Formula 2-3)}$$

$$\text{Average Macroprecision} = \frac{1}{m}\sum_{i=1}^{m}\frac{c_i}{c_i + d_i} \qquad \text{(Formula 2-4)}$$

The second procedure places equal weight upon the documents retrieved by treating the results of multiple questions in exactly the same manner as if they had resulted from one large search question. The resulting averages are called Average Microrecall and Average Microprecision and defined as follows:[139]

$$\text{Average Microrecall} = \frac{\sum_{i=1}^{m} c_i}{\sum_{i=1}^{m}(b_i + c_i)} \qquad \text{(Formula 2-5)}$$

$$\text{Average Microprecision} = \frac{\sum_{i=1}^{m} c_i}{\sum_{i=1}^{m}(c_i + d_i)} \qquad \text{(Formula 2-6)}$$

---

[138] Ibid., p. 299.

[139] Ibid.

Figure 2.7 illustrates what happens when a search question retrieves a comparatively large number of documents which are unrelated to the information request. As can be seen, questions 1, 3, 4, and 5 retrieved many less documents than question 2, which in the illustration had very poor recall and precision. An examination of the data reveals that the effect of question 2 which retrieved proportionately more documents than the other questions was greater on the Micro Averages than the Macro Averages. Determination of which average is most appropriate to evaluation depends on which is more important, the user request (in which case Macro Averages would be used) or distinguishing between relevant and nonrelevant documents (in which case Micro Averages would be used).[140]

Estimates of Recall In calculating recall in an experimental system with a small number of documents it is possible to look at every document in the file to determine its relevance to a specific question. With systems having thousands of documents this task is totally unrealistic, and procedures must be used to estimate values of average recall. There are at least five specific procedures mentioned in the literature for estimating recall or the number of documents in a file which are relevant to a specific question.

A first technique for estimating the number of documents in a file which are relevant to a question is to examine a random sample to determine what percent of the documents in the sample are relevant. For example, if in a 1% sample there were 5 relevant documents to a question it would be assumed that in the total file there would be 500 relevant

---

[140] Ibid.

	i	b_i	c_i	d_i	Recall $\frac{c_i}{b_i + c_i}$	Precision $\frac{c_i}{c_i + d_i}$
	1	3	7	7	.7	.5
	2	17	3	37	.15	.075
	3	1	9	6	.9	.6
	4	3	17	7	.85	.5
Total Searches=m=5	5	5	10	5	.67	.67
COLUMN TOTALS		29	46	62	3.27	2.345

$$\text{Average Macrorecall} = \frac{1}{5} \sum_{i=1}^{5} \frac{c_i}{b_i + c_i} = \frac{1}{5} (3.27) = .656$$

$$\text{Average Macroprecision} = \frac{1}{5} \sum_{i=1}^{5} \frac{c_i}{c_i + d_i} = \frac{1}{5} (2.345) = .469$$

$$\text{Average Microrecall} = \frac{\frac{\sum_{i=1}^{5} c_i}{5}}{\sum_{i=1}^{5} b_i + c_i} = \frac{46}{29 + 46} = \frac{46}{75} = .596$$

$$\text{Average Microprecision} = \frac{\frac{\sum_{i=1}^{5} c_i}{5}}{\sum_{i=1}^{5} c_i + d_i} = \frac{46}{46 + 62} = \frac{46}{108} = .426$$

FIGURE 2.7

A COMPARISON OF VARIOUS TYPES OF RECALL AND PRECISION AVERAGES

documents.[141] In large collections this, however, still has limitations. For if a 1% sample were taken of the MEDLARS system which contains over half a million documents it would require looking at more than 5,000 documents to estimate recall for a single question.

A second procedure involves using the retrieval of source or target documents as an estimate of recall. In this method source documents are used as a basis for writing questions to retrieve information from the file. Recall is then based on the assumption that the collection contains only source documents and is determined by the percentage of source documents retrieved when one question is written for each source document in a random sample.[142] For example, if 100 source documents were used in writing 100 different questions and 75 of these questions retrieved the specific source document used as a basis for writing the question, the estimate of average recall would be .75.

A third approach used by Lancaster in the evaluation of MEDLARS was based upon questions written by users. The user is asked to list documents that he knows are related to the specific question. The information from the title and author is used to determine if any of the documents listed by the user are in the information file and the percentage of these known relevant documents retrieved is used as an estimated recall. For example, if a user provided a list of documents related to a question, and if it were determined by using titles and authors that 20 of these were in the information file and 17 were

---

[141] Ibid.

[142] F. W. Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," American Documentation January, 1964, pp. 4, 8.

retrieved by the search question, the recall for that question would be estimated to be 17 ÷ 20 or .85.[143]

A fourth method is to use a KWIC index to examine the titles of the documents in the file and thus locate documents that are relevant to a specific question. These documents are then used as a basis for estimating recall for that specific question,[144] In using this method an assumption must be made that the documents not found by using the KWIC index will be retrieved in the same proportion as documents found by using the KWIC index.

A fifth technique that can be utilized with systems having a variety of automated search procedures is to use an information request to formulate questions that use as many of the different search procedures as possible. Results of these multiple searches for the same information request are examined and the aggregate of the relevant documents found is considered to contain all the relevant documents in the file.[145]

The method utilizing source documents was used by early studies of the Cranfield Project,[146] and subsequently criticized by Swanson because the question might not be typical of user's requests.[147] The procedure

---

[143]F. W. Lancaster, "Evaluating the Performance of a Large Operating Retrieval System" Electronic Handling of Information, Allen Kent, Orrin E. Taulbee, Jack Belzer, and Gordon D. Goldstein, editors, (Washington, D.C.: Thompson Book Company, 1967), pp. 201-204.

[144]Gerard Salton, Automatic Information Organization and Retrieval (New York: McGraw-Hill Book Company, 1968), p. 294.

[145]Ibid., p. 299.

[146]F. W. Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," American Documentation, January, 1964, p. 4-13.

[147]Swanson, Don R., "The Evidence Underlying the Cranfield Results," The Library Quarterly, January, 1965, p. 1-20.

was later defended by Cleverdon where he declared,

> Naturally, I accept that there must be a somewhat unnatural relationship between a question and a document on which it is based, but I am not prepared to concede that this relationship is such as to negate all the results. While we would not use questions of such type in a research test, I believe, for reasons argued elsewhere, that they can still be used satisfactorily in situations where time and cost are important considerations, as might be the case in an evaluation of an operational information-retrieval system, and I shall continue to believe this until experimental data are produced which show me to be incorrect.[148]

Relevance Judgment

One of the first questions that must be answered in the evaluation of an information system is the manner in which the relevance of documents to specific questions will be determined. Regardless of the objectivity used in other procedures there appears to be no way to do this without using human judgments. Because of the subjective nature of these procedures there has been considerable work to determine the effect of these judgments on the evaluation of information systems.

Cuadra, Katter, et. al. have made an extensive study of this area and the result is succinctly described in the abstract:

> Evidence has been developed that suggests that relevance judgments can be and are influenced by skills and attitudes of the particular judges used, the documents and document sets used, the particular information requirement statements, the instructions, and setting in which the judgments take place, the concepts and definitions of relevance employed in these judgments, and the type of rating scale or other medium used to express the judgments.[149]

---

[148]Cyril Cleverdon, "The Cranfield Hypotheses," The Library Quarterly, April, 1965, pp. 121-124.

[149]Carlos A. Caudra, Robert V. Katter, Emory H. Holmes, and Everett M. Wallace, Experimental Studies of Relevance Judgments: Final Report (Santa Monica, California: System Development Corporation, June, 1967), I-III.

This study points out that various factors can influence the relevance assessment of judges; however, other studies indicate that while there is variance between judgments, rankings of particular documents retrieved in response to a request tend to be the same for different groups of judges.[150] Salton feels that if consistent procedures are used when comparing various retrieval methods, the comparisons of the methods should not be affected because "bias introduced by individual faulty relevance judgments may be expected to be in the same direction for all methods." He does, however, acknowledge that absolute values for either recall or precision measures must be examined with considerable care because of their dependence upon subjective judgments.[151]

A study done by Lesk and Salton compared relevance assessments by individuals who compiled questions with relevance assessments of a second person who had compiled another request. The results indicated that while the overall agreement among relevance assessments was not high, this did not affect the relative performance of various retrieval methods. In other words, despite the inconsistency in judgment the comparative rankings of methods from poor to best were not changed.[152]

---

[150]Orrin E. Taulbee, "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (Chicago: William Benton, 1968), III, 107, citing Alan Rees, and Douglas G. Schultz, Principal investigators, A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Center for Documentation and Communication, School of Library Science, Case Western Reserve University, Cleveland, Ohio (October 1967, Vol. I (287 pp.): Vol. II, Appendices A-Q. Final Report to the National Science Foundation.

[151]Gerard Salton, Automatic Information Organization and Retrieval (New York: McGraw-Hill Book Company, 1968), p. 302.

[152]M. E. Lesk and G. Salton, "Relevance Assessments and Retrieval System Evaluation," Information Storage and Retrieval, December, 1968, pp. 343-359.

These studies tend to support the position that comparisons of different systems present difficulties because of the way relevance judgments affect numerical measures. Caudra, Katter, et. al. suggest that a variety of factors, including document sets, may influence relevance judgments. (See quotation on page 70.) If true this tends to make suspect comparisons when done on different information files.

Comparison of Indexing Schemes

Bourne in the first Annual Review provides a summary of the indexing evaluation projects prior to 1965. Most of these projects compare various indexing languages; however, says Bourne,

> One point becomes clearer after viewing this literature, namely that it is extremely difficult to make meaningful generalizations about the performance of various indexing systems. In almost all experimental reports, the investigator worked with an indexing language different than that of other experiments. Consequently, no one has ever had his test results verified, or expanded, or made more precise by another experimenter. Furthermore, the actual numerical values given for recall and relevance, or other factors for a particular indexing system would appear to have value only to that system.[153]

An examination of the literature for the most part still tends to support Bourne's position; however, there appear to be some other generalizations that are supported by a variety of articles. Of particular note are: (1) the fact that over-sophistication of indexing systems does not appear to be worthwhile, and (2) there exists an inverse relationship between recall and precision, such that in attempting to improve recall there is usually a decrease in precision and vice versa.

Sharp in the second Annual Review of Information Science and

---

[153]Charles P. Bourne, "Evaluation of Indexing Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, Inc., 1966), I, 180.

Technology, cites a series of authors to support his assertion that,

> The most useful product of any review is a positive conclusion that promises to have immediate application. An emergent feature of some of the investigations that are being carried out is that for many retrieval systems, attempts to over-sophistication are not worthwhile,"[154]

Among the studies cited by Sharp is the Cranfield Project which compared four indexing languages--U.D.C., alphabetical, facet, and uniterm--where Sharp quotes Cleverdon and Keen's conclusion that "single term indexing languages are superior to any other type,"[155] After citing a number of other studies Sharp concludes that the studies ". . . are all pointers to the probability that in small to medium systems where no special conditions exist, over elaboration of retrieval languages and their associated control devices cannot be justified,"[156] The fact that simple indexing methods, such as key-word indexes, appear to do as well or sometimes better than other indexing languages is important to machine indexing, because of their adaptability to computer programming.

The second generalization that appears appropriate is that there is a tradeoff between recall and precision. This generalization is supported by Lancaster and Mills in their report of the Cranfield Project[157]

---

[154]John R. Sharp, "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, Inc., 1967), II, 90.

[155]Ibid., p. 90 citing Cyril Cleverdon and Michael Keen, Factors Determining the Performance of Indexing Systems, Vol. 2: Test Results, ASLIB Cranfield Research Project, Cranfield, Bedford, England, 1966, pp. 1-299.

[156]John R. Sharp, "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, Inc., 1967), II, 90.

[157]F. W. Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," American Documentation, January, 1964, p. 9.

and also by Salton, Keen, and Lesk in their report on the results of the SMART system.[158]

Swets has recognized this tradeoff in his development of a descriptive statistic called "the normal operating characteristics curve" which considers the variables used in the calculation of both precision and recall.[159] This tradeoff can be seen in a subjective way by considering what happens when the total file is retrieved in response to a question. Obviously all documents that are in the file and relevant to that question are in the response, thus giving perfect recall. Yet the precision is reduced to the number of relevant documents in the file divided by the total documents in the file. This idea is clearly stated by Sharp. He says, "It is easy enough to ensure that all or most of what is of interest is recovered from a file by simply casting the net wide enough, but there is little purpose in doing this if the result is a set of retrieved documents of unmanageable size.[160]

In comparing various types of controlled indexing languages Hyslop concludes that a thesaurus type of control seems best for a computerized system.[161] Artandi concludes that many issues remain unsolved but "the advent of interactive systems has placed a new emphasis on the thesaurus

---

[158]G. Salton, E. M. Keen, and M. Lesk, "Design Experiments in Automatic Information Retrieval," The Growth of Knowledge, Manfred Kochen, editor, (New York: John Wiley & Sons, Inc., 1967), pp. 344-346.

[159]John A. Swets, "Information-Retrieval Systems," The Growth of Knowledge, Manfred Kochen, editor, (New York: John Wiley & Sons, Inc., 1967), pp. 174-184.

[160]John R. Sharp, "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, Inc., 1967), II, 115.

[161]Marjorie R. Hyslop, "Sharing Vocabulary Control," Special Libraries, December, 1965, pp, 708-714.

as an aid to the user as part of a user-oriented prompting apparatus," and later that ". . . operation of indexing, query formulation, and vocabulary building cannot be regarded as isolated activities, and that good systems design should provide for their meaningful integration."[162]

Thus, while Bourne's contention seems to be supported that indexing evaluation has been done in such a manner that it is difficult to compare systems,[163] there are some generalizations which can be made from conclusions dependent upon rankings or trends rather than a comparison of numerical values. These results tend to support the generalizations that:

1. There appears to be no clearly superior indexing scheme.
2. In determining the best procedures for any given system it is important to consider the total characteristics of the system being designed.
3. Some studies indicate that key-word systems which are applicable for machine use do as well as, or slightly better than, other more sophisticated indexing schemes.
4. Because of the apparent inverse relationship that exists between precision and recall, both statistics need to be reported if the evaluation of indexing methods is to be meaningful.

---

[162]Susan Artandi, "Document Description and Representation," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor, (Chicago: William Benton, 1970), V, 161.

[163]Charles P. Bourne, "Evaluation of Indexing Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (New York: John Wiley & Sons, Inc., 1966), I, 180, 181.

An Overview of BIRS--Basic Indexing and Retrieval System

The BIRS system is examined because it provides an example of a computerized information system and at present is the system used by the CEC-ERIC Information Center.

The term BIRS is used to refer to a series of computer programs written in Fortran IV which have been developed to assist in the computerized operations of information retrieval systems. "The distinctive characteristics of BIRS are generality (applicability to many types of information system problems), portability (ease of installation on many types of computer configurations), and usability (ease of usage by individuals lacking computer science training)."[164] The development of the programs has been jointly supported by Michigan State University and a series of federal grants from the U.S. Office of Education. The first grant was obtained in May of 1966, with continuing support from grants through March of 1971.[165]

An examination of the documentation reveals that the outward appearance of the system has remained reasonably constant; however, the actual development has been an evolutionary process resulting in a series of refinements and the ability of the programs to run on different

---

[164]John F. Vinsonhaler, The Information Systems Laboratory: A Progress Report for 1969 ISL Report No. 10. (East Lansing: Michigan State University, January 1970), P. 3.

[165]Ibid., p. 10.

computers.[166]

Considerable effort has been made to ensure that later BIRS versions have been built in modular form to permit the replacement of modules by more efficient operating programs without affecting the overall appearance of a system to the user.[167] The modular design and flexibility of the system makes this system ideal for those developing procedures for new information retrieval systems, as will be demonstrated in reporting this study.

The overall programs are divided into four major categories: (1) systems file maintenance, (2) information storage, (3) information indexing, (4) and information retrieval. The first, systems maintenance, is composed of three programs: The EXECutive program (EXEC), the TASK management program (TASK), and the TRANSlation program (TRANS). The single program which provides for information storage is the Information File Maintenance Program (IFMP). The four programs which make up information indexing are the Printed Indexing Program (PIP), the Printed Listing Program (PLP), the Descriptive Analysis Program (DAP), and the Descriptive File Maintenance Program (DFMP). The two programs which control information retrieval are the Descriptive File Searching Program

---

[166]Ibid., p. 2; John F. Vinsonhaler (ed.), Technical Manual Basic Indexing and Retrieval Systems, BIRS 2.0 (East Lansing: Educational Publications Services, College of Education, Michigan State University, January, 1968), pp. 101-111; and John F. Vinsonhaler and John M. Hafterson (eds.), Technical Manual for Basic Indexing and Retrieval System, BIRS 2.5, Appendix I (East Lansing: Educational Publications Services, College of Education, Michigan State University, January 1969), pp. 2001-3118.

[167]John F. Vinsonhaler and John M. Hafterson (editors), Technical Manual for Basic Indexing and Retrieval System, BIRS 2.05, Appendix I (East Lansing: Educational Publications Services, College of Education, Michigan State University, January, 1969), pp. 2004-2014.

(DFSP), and the Information File Retrieval Program (IFRP).[168]

Figure 2.8 provides a diagram of the BIRS system showing the individual programs and the relationship that exists between these programs. This diagram again illustrates the concept of input, processing, and output while the processing is here represented by the four fundamental operations, systems maintenance, information storage, information indexing, and information retrieval.

The Executive Program--EXEC The purpose of the Executive Program is to allow the user to use the various independent programs by means of a command language. The command language interacting with EXEC can call the various programs and indicate the specific task that these programs are to perform. Because of the Executive Program, the series of programs with their commands looks to the user like a powerful and simple programming language designed for doing information storage and retrieval.[169]

Task Management Program--TASK The Task Management Program is designed to permit the user to combine a series of command statements to form predefined operations. By calling for a specific task the user initiates the sequence of commands similar to the way one would use a subroutine in a programming language such as Fortran.[170]

Translation Program--TRANS This program is designed to translate the standard Fortran IV source programs into various machine-dependent Fortran dialects.[171]

---

[168] Ibid., p. 2012.

[169] Ibid., p. 2833.

[170] Ibid., pp. 2834-2838.

[171] Ibid., pp. 2839-3848.

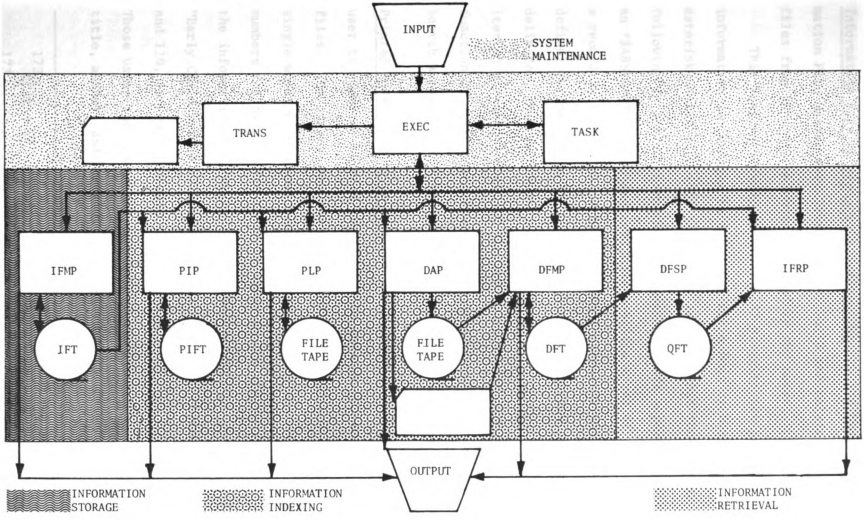


FIGURE 2.8
AN OVERVIEW OF THE BASIC INDEXING AND RETRIEVAL SYSTEM (BIRS)

Information File Maintenance Program--IFMP The purpose of the Information File Maintenance Program is to build and maintain information files from the data which is provided by the users.

This program as well as other BIRS programs can process textual information as it might be formatted for a book or article. The symbol asterisk-dollar sign (*$) is used to indicate that the word or phrase following is a BIRS command. Users may initiate separate records with an *$ABSTRACT command and designate specific types of information within a record by a field delimiter (a special character which the user may define) followed by a field name or blank. If there is a blank after the delimiter, the field must later be accessed by number with the n^{th} delimiter in a record indicating the beginning of the n^{th} field. The primary function of designating fields is to make it possible to index and/or search a portion of a record rather than the total record.[172]

Printed Indexing Program--PIP The Printed Indexing Program allows the user to generate printed indexes of the data stored on the information files. The program has the ability--at the option of the user--to index single words or phrases from any designated field. In the indexes numbers follow specific terms to indicate the abstracts (records) on the information file which contain that term. For example, if the term "Early Childhood Education" were followed by the numbers 4, 17, 25, 107, and 110, it would indicate that the term is in each of these abstracts. Those using the system for information center operations commonly index title, author, and descriptor fields.[173]

---

[172] Ibid., pp. 2101-38.

[173] Ibid., pp. 2201-19.

Printed Listing Program--PLP The purpose of the Printed Listing Program is to list the content of the information file in alphabetical order, according to a criteria determined by the user. Ordinarily the ordering is done on a single term per abstract--the first term of a specific field--so that each abstract will appear only once in the ordered list; however, it is possible, by a command, to remove this restriction so that an abstract will appear once for each different term in a specified field.[174]

Descriptive Analysis Program--DAP The commands and functions of this program are similar to the Printed Indexing Program in that the results of the analysis are used by the Description File Maintenance Program to build a description file. As in the Printed Indexing Program, words or phrases from any number of specific fields may be indexed.[175]

Description File Maintenance Program--DFMP The purpose of the Description File Maintenance Program is to use the results of the Descriptive Analysis Program in building and maintaining description files which are searched by the Description File Search Program. The program is very similar to the Information File Maintenance Program in that new files may be built, old files added to, or old descriptions replaced by new descriptions.

As with the information file where abstracts are sequentially stored, the description file has descriptions of abstracts stored to form a one-to-one correspondence between the description on the

---

[174] Ibid., pp. 2301-5.

[175] Ibid., pp. 2401-19.

description file and the abstract on the information file. The major purpose of the description file is to store the information needed for computer searching in a manner that will reduce the computer time required for searches.[176]

Description File Search Program--DFSP The following discussion of the Description File Search Program is more comprehensive than the descriptions of the other programs because of its relation to the evalution procedures described in Chapter 4.

This program allows a user to search descriptions of documents contained on the description file and generate a question file containing information about all documents which meet the criteria defined by a question. Available in this program is a language which allows for logical (Boolean) searches, relevance searches, and weighted relevance searches.

All questions in the BIRS program begin with an *$QUESTION appearing separately on the first line followed by lines containing the user's request for information--a logical expression or logical expressions separated by commas. The simplest logical expression is a word, the next simplest a phrase, with more complicated expressions being formed by joining and/or modifying words and phrases with the logical operators .AND., .OR., and .NOT.. Examples of logical expressions which might be used in a BIRS question are:

1. RETARDED
2. MENTALLY HANDICAPPED

---

[176]Ibid., pp. 2501-2519.

3. MENTALLY HANDICAPPED .AND. BLIND

4. MENTALLY RETARDED .OR. MENTALLY HANDICAPPED

5. .NOT. BLIND

6. (MENTALLY RETARDED .OR. MENTALLY HANDICAPPED) .AND. DEAF .AND. .NOT. BLIND

If expression 3 were used as a question a document would have to contain both the terms MENTALLY HANDICAPPED and BLIND before it would be retrieved. If expression 4 were used a document would be retrieved if it contained one or both of the terms MENTALLY RETARDED or MENTALLY HANDICAPPED. If expression 5 were used the document would not be retrieved if the term BLIND were in the document but would be retrieved if the term BLIND were not in the document.

When a user is uncertain of the order in which a computer will combine words or phrases to form a new expression, the user may indicate the order with parentheses. The parentheses are interpreted by the computer to mean that the words or phrases inside the parentheses are to be combined first. Expression 6 illustrates one of the most frequent uses--the combining of words or phrases by .OR. to indicate they have similar meanings. Expression 6 also illustrates the fact that when logical expressions are combined by .OR., .AND., and/or modified by .NOT. the resulting expression is also a logical expression.

Another command available in the BIRS language is *$COMMENT which causes all information between this command and the next line beginning with an *$ to be printed as a comment on the BIRS report. The *$COMMENT command, and the *$QUESTION command followed by a question are used in Figure 2.9 to illustrate some of the types of questions recognized by DFSP.

```
*$COMMENT      EXAMPLE 1
*$QUESTION
MENTALLY HANDICAPPED


*$COMMENT      EXAMPLE 2
*$QUESTION
MENTALLY HANDICAPPED .AND. BLIND


*$COMMENT      EXAMPLE 3
*$QUESTION
(MENTALLY HANDICAPPED .OR. MENTALLY RETARDED) .AND.
PHYSICALLY HANDICAPPED


*$COMMENT      EXAMPLE 4
*$QUESTION
(DEAF .OR. HARD OF HEARING .OR. HEARING IMPAIRED) .AND.
(BLIND .OR. VISUALLY HANDICAPPED) .AND. .NOT.
(MENTALLY HANDICAPPED .OR. MENTALLY RETARDED)


*$COMMENT      EXAMPLE 5
*$QUESTION
DEAF, BLIND, SPEECH HANDICAPPED, MENTALLY RETARDED


*$COMMENT      EXAMPLE 6
*$QUESTION
MENTALLY RETARDED=4, DEAF, BLIND, SPEECH HANDICAPPED


*$COMMENT      EXAMPLE 7
*$QUESTION
(MENTALLY HANDICAPPED .OR. MENTALLY RETARDED=4, (DEAF .OR.
HARD OF HEARING .OR. HEARING IMPAIRED), (BLIND .OR.
VISUALLY HANDICAPPED), SPEECH HANDICAPPED
```

FIGURE 2.9

EXAMPLES OF SEARCH QUESTIONS

Logical Search Questions The first four questions in Figure 2.9 use single logical expressions and are examples of logical or Boolean search questions. The fourth question illustrates a more complicated expression written to retrieve documents that are about hearing problems and visual problems, but not about mental handicaps. The multiple words or phrases used for each of the concepts in question 4 illustrate how a list of similar terms may be placed inside a parenthesis and combined with .OR.'s when it is not certain which of several terms might have been used to describe a document.

Relevance Search Questions Another type of search question, a relevance question, is illustrated by Example 5 of Figure 2.9. In this question documents containing all four terms would be retrieved first, followed by documents containing only three of the terms, then by documents containing only two of the terms and finally followed by documents containing just one of the four terms. By use of other commands it is possible for a user to specify the given number of terms a document must have before it is retrieved.

Weighted Relevance Searches By weighting MENTALLY RETARDED in Example 6 it is given more importance than the combination of the other three terms. If a command is given to indicate that documents are not to be retrieved unless they have terms with combined weights of five or more; it will guarantee that only documents having the term MENTALLY RETARDED and one or more of the other terms will be retrieved.

Example 7 illustrates a question which has used a number of logical expressions to expand each of the four concepts used in Examples 5 and 6.

By writing questions where lists of words or phrases having similar meanings are joined by .OR. the number of possible matches for each concept has been expanded.

DFSP also has available the relational and arithmetic operators; .EQ. (Equal to), .NE. (Not equal to), .GE. (Greater than or equal to), .GT. (Greater than), .LE. (Less than or equal to), .LT. (Less than), .PL. (Plus), .MI. (Minus), .DV. (Divided by), and .TM. (Times), which can be used for testing numerical values associated with descriptions. For example, if it were desired to retrieve all documents about the mentally retarded that had been published since 1969, the following question might be used:

```
*$QUESTION
MENTALLY RETARDED .AND. DATE .GE. 1969
```

The other arithmetic operators will not be illustrated because they do not directly relate to this study.[177]

Information File Retrieval Program--IFRP The Information File Retrieval Program enables the user to specify the form of output (access numbers only, specific lines of given abstracts, or the total abstracts) desired and have it printed on a high-speed line printer, remote terminal, or other output device.[178]

The user might view the multiple commands available to the BIRS programs as a command language with input to a single program--the Executive Program. In this sense, the user is writing a series of commands which with other appropriate data are used as input to be processed by the system which generates the specified output thus illustrating again the concept of input, processing, and output.

---

[177] Ibid., pp. 2601-30.

[178] Ibid., pp. 2101-25.

Summary

Information storage and retrieval systems represent one of many man-made systems designed to perform specific functions. As with other man-made systems they include interacting components, such that both the components and their interaction need to be considered in their design, analysis, and evaluation.

While not a sufficient condition alone for an information system to operate successfully, good indexing procedures are necessary and by some are considered to be the single most important component. The ultimate criteria in the evaluation of indexing methods and their interactions with a total information system is the system's ability to effectively retrieve information.

The descriptive statistics most commonly used to describe the results of searches are various types of precision and recall averages. Two types of averages involving precision and recall are commonly used; the first gives equal weight to search questions while the second gives equal weight to each document that is retrieved and/or is relevant to any of the questions.

In large systems it has been necessary to estimate recall because of the difficulty of examining the relevance of each document in a file to each search question. To cope with this problem a variety of methods have been developed for estimating recall with no single method or procedure used for all circumstances.

The BIRS system serves as an example of a computerized information retrieval system with its interacting components. This is the system which is used by the CEC-ERIC Information Center and which was used in the evaluation of their indexing methods.

Even when a single component is being analyzed or evaluated, it is important that it be considered within the context of the total system. The purpose of the following chapter is to describe the CEC-ERIC Information Center and its development so that the evaluation of the indexing method used at the Center may be considered in context.

CHAPTER III: THE DEVELOPMENT OF CEC-ERIC INFORMATION CENTER AND ITS PRESENT OPERATING STATUS

An information center is much more than a computerized system for storing and retrieving information. Thus, when evaluating a component of an information center's operations such as its indexing methods, it is important to have in perspective the total objectives and procedures of the center. The ability of the information center to meet its objectives is dependent upon a complex interaction of many variables, including the type of information it stores, the indexing methods it uses, the techniques available for retrieving information and the needs of its users.

The CEC-ERIC Center has its origin in the ERIC network and has been affected in a continuing manner by the procedures and objectives of that system. Because of this relationship it would be difficult to view the CEC-ERIC operations in perspective without some knowledge of central ERIC. King states, "An intricate part of information systems research involves a description of the entire system including its environment and component parts."[179] The major objective of this chapter is to describe the entire system including its environment and component parts. Specifically the chapter contains:

---

[179]Donald W. King, "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor (Chicago: William Benton, 1968), III, 63.

1. A description of central ERIC including its origin, objectives, and the relationship of CEC to the ERIC system.
2. A description of the historical development of the CEC-ERIC Information Center.
3. A description of the present operating procedures of the CEC-ERIC Information Center.
4. Descriptive statistics about the CEC-ERIC Center's present operating status.

A History and Description of Central ERIC

The ERIC system had its formal beginning in 1965 with its first publication, a Catalog of Selected Documents on the Disadvantaged.[180] In June, 1965 North American Rockwell was awarded the initial contract for the management of the central ERIC facility including major responsibility for processing related to the publication of Research in Education[181] (RIE). The first issue of RIE, ERIC's monthly announcement bulletin was published approximately four months later. In November, 1965 a contract was awarded to the Bell Honeywell Company to provide a service called Educational Document Reproduction Services (EDRS) which reproduced on microfiche and hard copy, documents which were difficult to obtain and not copyrighted.[182]

---

[180]Lee G. Burchinal, "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, April, 1970, pp. 59, 60.

[181]Ibid.

[182]Ibid.

Since 1966 local clearinghouses have been developed to handle information in specific topical areas and to provide central ERIC with abstracts of documents to be published in the ERIC journal, Research in Education. The documents which are not protected from reproduction by copyright laws are made available through EDRS in microfiche, hard copy or both.[183] The first eleven clearinghouses were established between March and June of 1966.[184]

Objectives of ERIC

The initial objectives of the ERIC program were stated as:

> Making hitherto unavailable or hard to find, but significant research and research related reports, papers, and other documents easily available to the educational community.
>
> Interpreting and summarizing information for many reports in ways that educational decision-makers and practitioners could understand and use the emerging results from the national R&D effort.
>
> Strenghtening existing channels of communication for putting R&D results into practice.
>
> Providing a base for developing a national education information network that can effectively link knowledge producers and users in education.[185]

Later objectives have been stated as:

1. Guarantee ready access to the world's English-language literature relevant to education. In information science terminology this is the documentation function of the program.

2. Generate new information products by reviewing, summarizing, and interpreting current information on priority topics.

---

[183] Ibid.

[184] Ibid.

[185] Lee G. Burchinal, "Evalaution of ERIC, June, 1968," available as ED 020 449 from the ERIC Document Reproduction Service. ERIC stands for the Educational Resources Information Center, a national education dissemination system designed and supported by the Office of Education, Department of Health, Education and Welfare.

This is the information analysis function of the system. Products include bibliographies, state-of-knowledge papers, critical reviews, and interpretive summaries.

3. Infuse information about educational developments, research findings, and outcomes of exemplary programs into educational planning and operations.[186]

A comparison of the early objectives with later objectives reveals a broadening of the scope of the ERIC system. In the first set of objectives the major emphasis was on making accessible previously "unavailable or hard-to-find" documents about significant research in education[187] whereas in the latter set of objectives the scope has been broadened to "guarantee ready access to the world's English-language literature relative to education."[188] Consistent with these broadened objectives ERIC in July of 1969 began publishing Current Index to Journals in Education (CIJE).

The following two ambitious principles have been presented as relevant to the broadened documentation objectives of ERIC:

1. Educators should be able to turn to one comprehensive source to identify current, significant educational documents on any topic of interest.

2. Educators should be able to obtain desired reports quickly, again from one source, regardless of where the report originated.[189]

---

[186]Lee G. Burchinal, "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, April, 1970, p. 56.

[187]Lee G. Burchinal, "Evaluation of ERIC, June, 1968," available as ED 020 499 from the ERIC Document Reproduction Service.

[188]Lee G. Burchinal, "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, April, 1970, p. 56.

[189]Ibid.

The name of any discipline could be substituted for "education" in the above stated principles, and one would have presented the desires of many researchers regardless of their field of endeavor. By expanding these principles to a variety of disciplines, one would in essence have the World Encyclopedia described by H.G. Wells.[190] The major problem in meeting these objectives is not related to technology but to copyright laws. In attempting to apply these principles, ERIC has met with continued and firm resistance by publishers of professional journals to include copyrighted materials in RIE or to produce these materials in microfilm or hard cover for EDRS.[191] Unless a solution is found to this problem the plan for educators to secure relevant information from a central source would seem impossible to implement.

The Growth of ERIC

Among the statistics used to measure an information network are the number of documents it disseminates and users it serves. An examination of such statistics leaves little doubt that the influence of ERIC is developing very rapidly. These statistics indicate that in July 1966, ERIC's document collection, (not including articles) had 1,746 documents; in January 1967, 1,839 documents; in July 1967, 3,551 documents; in Janaury 1968, 7,227 documents; and in June 1968, 12,324

---

[190]H. G. Wells, World Brain (Garden City, New York: Doubleday, Doran & Co., Inc., 1938), pp.3-34.

[191]Interview with Dr. Richard Dershimer, Executive Secretary of the American Educational Research Association, March 16, 1971.

documents.[192] The number of subscribers to Research in Education has grown from 209 in January, 1967 to 4,558 in May of 1968.[193] The number of documents sold in microfiche has increased from 328,000 in 1967 to more than 7 million in 1969.[194]

Since 1969 articles indexed for the Current Index to Journals in Education have been included in the ERIC collection. In 1969 about 12,000 articles from 220 periodicals were indexed in 1970 this increased to 18,000 articles from 500 journals.[195] The above statistics as well as data about the use of the ERIC clearinghouses[196] appears to strongly indicate that the impact of ERIC is growing.

The Future of ERIC

Some of the goals set by the ERIC staff for future development are: (1) the covering of new topical fields by improving existing clearinghouses and adding a limited number of new ones, (2) the expanding of the information analysis program, (3) the supporting of one-stop information centers with emphasis on the use of state agencies,

---

[192]Lee G. Burchinal, "Evaluation of ERIC, June 1968," available as ED 020 499 from the ERIC Document Reproduction Service, Fig. 1.

[193]Ibid. Fig. 3.

[194]"Development of ERIC Through December, 1968," Prepared under the direction of Lee G. Burchinal, Director, Division of Information Technology and Dissemination, Bureau of Research, U.S. Office of Education, Department of Health, Education and Welfare, Office of Education/Office of Information Dissemination, first printed in August 1969, Revised Rebruary 1970, Fig. 1.

[195]Lee G. Burchinal, "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, April, 1970, pp. 60-63.

[196]Statistics on ERIC accompanied by cover letter from Delmer J. Trester, System Coordinator, Department of HEW, Office of Education, addressed to Carl Oldsen, CEC-ERIC, February 16, 1971.

(4) the expanding of acquisition efforts with emphasis on the use of state agencies to review reports and enter them in the ERIC system, (5) the further developing of computer searching capabilities including on-line terminal access to the ERIC file, (6) the using of commercially supported products to disseminate information contained in the ERIC files, and (7) the developing of means to insure that valid research data and information is applied to the improvement of educational programs.[197]

The Development of the CEC-ERIC Information Center

In the early part of 1966 The Council for Exceptional Children submitted a proposal to the U. S. Office of Education titled "Handicapped Children and Youth ERIC Clearinghouse and Research Dissemination Project." The project began July 1, 1966 with Dr. June Jordan as the first director of the CEC-ERIC Information Center. There were three main programs:

Program I. Operation of an ERIC Clearinghouse for Handicapped Children and Youth.

Program II. Expansion of the Clearinghouse activities and the development of materials interpreting research results.

Program III. Implementations of research findings into educational programs for handicapped children.[198]

---

[197]Lee G. Burchinal, "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, April, 1970, pp.60-63

[198]June B. Jordan, "Handicapped Children and Youth ERIC Clearinghouse on Research Dissemination," a proposal submitted to the U. S. Department of Health, Education and Welfare, Bureau of the Handicapped, (1966), p. 1.

The functions of the ERIC Clearinghouse (Program I) were to include:

> (a) identifying significant research and research related literature not readily available to the consumer, (b) abstracting and indexing such literature, (c) maintaining Clearinghouse material in a storage and retrieval system, and (d) participating in the development of an educational thesuarus.[199]

The activities of Program II and III are an expansion of the normal activities of an ERIC Clearinghouse. These expanded activities included: (a) abstracting additional significant research relative to the education of handicapped children, (b) coordinating the activities of the Handicapped Children Instructional Materials Centers, (c) developing special materials to describe educational practices based upon research, and (d) implementing the use of developed materials through workshops and demonstrations for teachers, supervisors, and teacher educators.

The specific objectives set forth in the proposal were:

Objectives

> The general objective of this proposed project is to provide a central, comprehensive clearinghouse information center in the education of handicapped children and youth. This center would be concerned with collecting, abstracting, testing, and evaluating literature and materials as well as developing materials, interpreting research, and disseminating information.
>
> Objectives--Program I. ERIC Clearinghouse: Handicapped Children and Youth. The Clearinghouse would serve as one of the US Office of Education's ERIC satellites and would follow procedures required of such a unit. Specific objectives include:
>
> 1. Identify and collect research and research related literature not readily available for wide dissemination.

---

[199] Ibid.

2. Evaluate the above material in terms of professional value.
3. Identify items for storage in handicapped area Clearinghouse (material with limited interest) and items for storage in central ERIC.
4. Prepare abstracts of above material and index.
5. Serve as a depository for Handicapped Clearinghouse materials.
6. Develop a retrieval system for location of Clearinghouse materials.
7. Coordinate efforts with related professional activities--Instructional Materials Centers, CEC's regular publication program (Exceptional Children Abstracts and Biennial Review of Research on Exceptional Children).
8. Disseminate information on the operation and ERIC acquisitions to the profession.
9. Provide an information service and copies of ERIC materials (hard copy and/or microfiche) upon request.
10. Participate in the development of the Office of Education's educational thesaurus (PET).
11. Provide for continuous evaluation of the effectiveness of the Clearinghouse operation.

Objectives--Program II. Expansion of Clearinghouse Activities and Development of Materials Interpreting Research Results. This aspect of the project is concerned with expanding the collection of literature, coordinating the efforts of Handicapped Children Instructional Materials Centers, and interpreting research results to the practitioner. Specific objectives include:

1. Provide in a central source a "library" of information on significant literature and materials related to the education of handicapped children.
2. Serve as the central communications center for the various Handicapped Children Instructional Materials Center located throughout the country.
3. Survey the research related to the education of handicapped children and identify results that have implications for classroom teaching.
4. Translate above research information into educational practices and develop materials (literature, films, video tapes, etc.) which would illustrate desirable educational practices indicated by research evidence.
5. Provide for a continuous review of research literature to identify that which has relevance for classroom instruction of handicapped children.

Objectives--Program III. Implementation. The purpose of Program III is to implement research findings into the classrooms for handicapped children. Specific objectives include:

1. Use materials developed in Program II in workshops with practitioners.

2. Cosponsor workshops with special education departments in local and state school systems and in college teacher preparation programs.[200]

The original proposal estimated that approximately 4,000 documents would be processed during the first year and 5,000 during the succeeding years. This was to be done by a staff which included a project director, an editorial secretary and two stenographer-clerks. Considerable importance was placed on field abstracting with the assistance of three university personnel at one-fourth time each and six graduate students at one-half time each. During the second year there was to be an additional staff associate at a professional level, the equivalent of one-fourth time of a university person and two graduate students on a one-half time basis. At the end of the second year, Programs II and III would also include three full-time equivalents on the central staff and four full-time equivalents as part of the field staff.[201]

The Early Operation of the Center

In March of 1967 Volume I, No. 1 of a tabloid entitled Clearinghouse on Exceptional Children announced to potential users the opening of the ERIC Clearinghouse on Exceptional Children.[202] Beginning with the second issuse of this tabloid was printed as a section of the journal Exceptional Children and sent to its approximately 40,000 subscribers with reprints available for special dissemination.[203] Since

---

[200]Ibid., pp. 2-3.

[201]Ibid., pp. 4-6.

[202]Clearinghouse on Exceptional Children, March, 1967, p. 1.

[203]"Clearinghouse on Exceptional Children," Exceptional Children, Summer, 1967, p. 693.

Volume II, No. 1, it has appeared under a new title "ERIC Excerpt."[204]

In the early part of 1967 arrangements were made for various organizations to do field abstracting. Included were graduate students and faculty from the University of Minnesota, the Alexander Graham Bell Association, and volunteers from the Association for the Gifted. In July of 1967 the first abstracts prepared by the CEC-ERIC Information Center appeared in the ERIC journal Research in Education.[205]

By the fall of 1967 CEC-ERIC Clearinghouse had processed over 100 abstracts into the ERIC collection, had approximately 900 additional documents in some stage of abstracting or indexing, and was receiving an additional 150 to 200 new items per month.[206]

The October issue of "ERIC ExCerpt" reported that the major focus of the staff would be on the following activities:

1. Processing abstracts into the Central ERIC collection.
2. Building an extensive computer bank of abstracts of literature and instructional materials.
3. Publishing special education abstracts on a regular basis as a companion piece to Research in Education.
4. Developing bibliographies on special topic areas in special education.
5. Disseminating information on the use of instructional materials through a department in Exceptional Children, "IMC Network Report."
6. Preparing films or video tapes interpreting research in terms of educational practice.[207]

---

[204]"ERIC ExCerpt," Exceptional Children, October, 1967, pp. 143-148.

[205]Research Education, July, 1967.

[206]"ERIC ExCerpt," Exceptional Children, October, 1967, p. 143.

[207]Ibid., p. 144.

At the time these statements were made it was unknown what the exact nature of the extensive computer bank of abstracts would be, how it would be used, how the abstracts not in ERIC would be published, and how the bibliography series would be produced. In April of 1968 the following quotation appeared:

> Within the next few months plans of ERIC-CEC include the development of numerous bibliographies on various special education topics and expansion of services to individuals by arranging for computer assisted searches on highly specific questions.[208]

At the time this statement was made the Instructional Materials Center at Michigan State University was using the BIRS programs to perform many of the operations desired by the CEC-ERIC Information Center. The major problem involving the use of BIRS programs to assist the CEC-ERIC Information Center was that the programs had been developed for a Control Data Corporation 3600 Computer located at Michigan State University. A similar computer was not available in the Washington area nor did The Council for Exceptional Children have trained staff necessary to maintain and run the programs.

In February of 1968 Dr. John Vinsonhaler received a grant titled "Improving the Dissemination of Instructional Materials, USOE Grant to develop BIRS for Information Management." One objective of the grant was to continue the development of BIRS so that it could run on a variety of computers including IBM system 360, models 40 and above. In the spring of 1968 negotiations were completed for the CEC-ERIC Information Center to use an IBM 360, model 40 located at George

---

[208]"ERIC ExCerpt," Exceptional Children, April, 1968, p. 633.

Washington University, thus making it possible for the Center to use the BIRS program.

The Establishment of Data Processing Procedures

One of the major advantages of the BIRS programs are the many alternatives they provide for maintaining and processing information files. This flexibility is especially useful when a user is not certain of the type of processing needed.

A weakness of the Information Center was the absence of an individual with a combination of skills necessary to develop procedures for computer processing. After making unsuccessful attempts to find an individual locally, help was requested from the BIRS project. In response, Dr. John Vinsonhaler, the MSU project director, provided the Center with the half-time consulting services of one of his staff members. The person provided was the author of this thesis who at that time was working as a computer specialist with the BIRS project.

The Decision to Publish a Computerized Journal

It was originally felt by some CEC-ERIC staff members that a good way to meet the goal of providing information to the center's users in as effective and efficient manner as possible would be for the Information Center to disseminate computer-readable information files to various computer centers.

In examining this and alternative methods the following factors and their interactions were examined.

1. The type of information disseminated by the CEC-ERIC Center.
2. The technological capabilities in the Washington, D.C. area.

3. The manner in which data was being processed for use on the ERIC system.
4. The sophistication of users.
5. The locations of users and cost of mailing information.
6. The similarity of questions asked by the various users.
7. The time required to implement various alternatives.
8. The cost of the various alternatives.

During the time when alternatives were being analyzed, computer files containing representative information processed by the CEC-ERIC Center were established at Michigan State and George Washington Universities. These files served as a basis to examine various processing procedures and acquaint the Information Center staff with some of the potential problems of a computerized system.

In the early fall of 1968 the author met with Dr. June Jordan, Director of the CEC-ERIC Center, and Mr. William Geer, the Executive Secretary of The Council for Exceptional Children, to consider the alternatives. In this discussion the difficulty of maintaining searchable computer files at various installations around the country was considered and rejected on the basis that:

1. There were only a limited number of centers with the necessary equipment, personnel, and willingness to serve users.
2. The geographic distribution of users versus the potential locations of such centers would greatly limit the number of individuals who would have access to these services.
3. It would be difficult to provide different centers with the type of professional educational staff necessary to make the procedure effective.

A second alternative considered was the possibility of an on-line system. This was rejected because of:

1. The cost of providing communication lines.
2. The cost of terminal rentals.
3. The cost of having a computer or a portion of a computer dedicated to the information file.
4. The limited number of locations where terminals could be maintained with the available financing.

The third alternative, considered and accepted, was to maintain the computerized information file in such a manner that it would allow for computer-controlled publication of a journal as well as selective publication of any portion of the information files. This alternative made it possible to:

1. Publish a journal directly from the computer information files using computer-controlled typesetting.
2. Use the computer to generate author, subject, title, or other types of indexes to use in the journal.
3. Use computer searches to aid in answering difficult user requests.
4. Use computer searches to help organize and select documents about special topics which relate to commonly asked questions.
5. Use the computer to index and to control typesetting so that annotated bibliographies of the selected documents could be printed at a minimal cost.

At the close of the meeting, Mr. Geer indicated that the Center should move as rapidly as possible to develop procedures for publishing

a computerized journal containing abstracts about exceptional children. This decision resulted in the publication of a new journal, Exceptional Child Education Abstracts, which first appeared in April, 1969.

An Overview of the Operating Procedures Used by the CEC-ERIC Information Center

The purpose of this section is to survey the procedures used at the Information Center, and relate the indexing process to the other procedures, thus describing the context in which the indexing evaluation took place. While this section gives an overview, a more complete discussion has been provided in Appendix A for those desiring additional information about procedures used at CEC-ERIC Information Center.

Legend and Nomenclature

The symbols used in the diagrammatic representation of CEC-ERIC's processing are commonly used in computer program and systems flowcharting. The description of the symbols found in Figure 3.1 are those given on the cover of the IBM flowcharting template, form X20-8020. In addition to the symbols described in Figure 3.1 the following alphanumeric legend is used to identify specific symbols in various figures:

1. C(N) stands for Connection number N where N may be any number.
2. IP(N) stands for Input number N.
3. OP(N) stands for Output number N.
4. PP(N) stands for Predefined Process N.

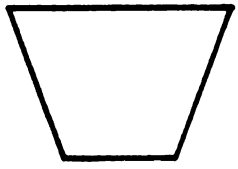
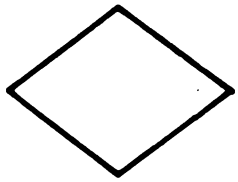
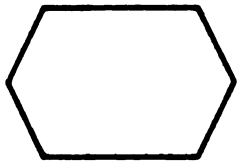
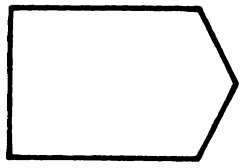
SYMBOL	REPRESENTS
	INPUT/OUTPUT Any function of an input/output device (making information available for processing, recording processing information, tape positioning, etc.)
	DECISION The decision function used to document points in the program where a branch to alternate paths is possible based upon variable conditions.
	PREDEFINED PROCESS A group of operations not detailed in the particular set of flowcharts.
	PROGRAM MODIFICATION An instruction or group of instructions which changes the program.

FIGURE 3.1

FLOWCHARTING SYMBOLS

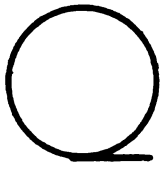
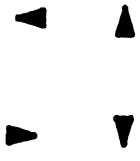
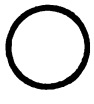
SYMBOL	REPRESENTS
	DOCUMENT Paper documents and reports of all varieties.
	**MAGNETIC TAPE**
	**FLOW DIRECTION** The direction of processing or data flow.
	**CONNECTOR** An entry from, or an exit to, another part of the program flowchart.

FIGURE 3.1 (cont'd)

5. SB(N) stands for Symbol N of a given figure. This notation will be used when it is necessary to identify a given symbol for discussion which is not specified in another way.

Model Developed for the CEC-ERIC Information Center

In the original proposal for the Information Center Dr. Jordan included objectives that provided guidelines for the Center's development. A diagrammatic representation of a model for the Center's operation was contained in that proposal. Considering the lack of information about potential technological components, the model demonstrated remarkable insight and still bears considerable resemblance to the present functioning of the Information Center.[209]

Overview of the Information Center's Major Activities

The overview in Figure 3.2 is an outgrowth of the conference between the author, Dr. Jordan, and Mr. Geer where the decision was made to publish Exceptional Child Education Abstracts and includes changes that have been made in the procedures as a result of that decision. This simplified overview divides the processing into six major activities: document acquisition, document management, file maintenance, file processing, information processing, and evaluation with system modification.

The core of activities shown in Figure 3.2 is presented with greater detail in the later diagrams, Figures 3.3, 3.4, and 3.5,

---

[209]June B. Jordan, "Handicapped Children and Youth ERIC Clearinghouse on Research Dissemination," a proposal submitted to the U. S. Department of Health, Education and Welfare, Bureau of the Handicapped, (1966), p. 5.

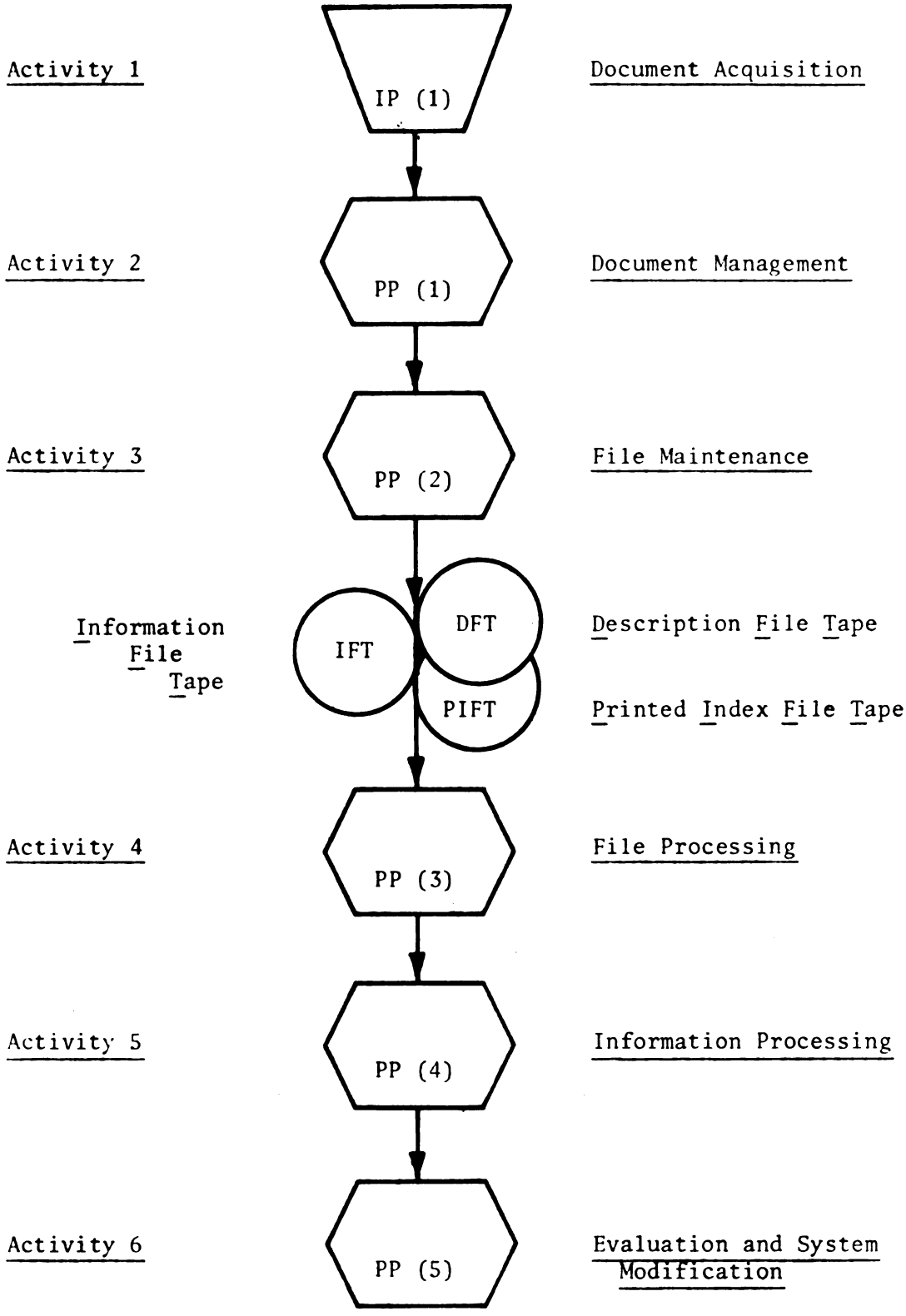


FIGURE 3.2

OVERVIEW OF INFORMATION CENTER MAJOR ACTIVITIES

and related discussions. These activities are found in most information centers utilizing computer processing; however, the specific steps and resulting products may vary considerably from center to center.

A brief description of each of these six major activities follows:

Activity 1 - Document Acquisition This activity includes the selection of documents which will be bought or acquired by other methods so that they may be examined to determine if they are appropriate for inclusion in the Information Center holdings.

Activity 2 - Document Management This activity includes examining documents to determine if they should be included in the Information Center data bank, the abstracting of documents, the indexing of documents, and the cataloging of documents.

Activity 3 - File Maintenance This activity includes punching document surrogates, storing the document surrogates on a computerized information file, and preparing computerized description files and printed index files.

Activity 4 - File Processing This activity includes computer processing of files to organize the information in a form that will be more useful and easier to disseminate.

Activity 5 - Information Processing This activity involves processing user requests, providing users with information, publishing new documents from the information contained on the computer files, and providing information to be used in evaluating the system. The activities in this section are primarily manual activities but they

may initiate computer file processing (activity 4) as one of several steps in a procedure.

Activity 6 - Evaluation This activity involves examining the procedures used by the Center and, if appropriate, modifying these procedures to make the total operation of the Information Center more effective.

Overview of Major Input and Output

Figure 3.3 provides an overview of the input to the Information Center and the output generated by processing this input. As illustrated documents are acquired (IP (1)) and processed in the document management activities (PP (1)) to generate copy for Research in Education (OP (1)) and Current Index to Journals in Education (OP (2)). All documents which will become part of the Information Center holdings, including those which are processed for RIE and CIJE, are then passed to file maintenance processing (PP (2)). In the file maintenance activity the documents are put in computer-readable form and various computer files are generated. These computer files provide input for the file processing (PP (3)) which generates the output for Exceptional Child Education Abstracts (OP (3)) and output for selected publication (OP (4)). This output is in a form that allows for computer typesetting, computer generated indexes, and printing with a minimum effort. ECEA and the selected publications in turn become input for Information Processing (PP (4)). These publications and other diagramed input (IP (2), IP (3), and PP (3)) are used in providing information to users (OP (5)), in assisting staff members to generate new documents (OP (6)), and as input to the evaluation component (PP (5)).

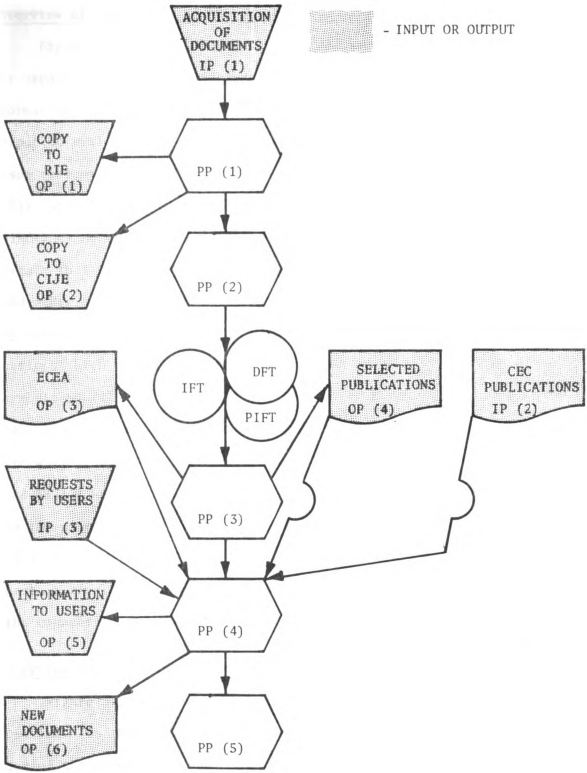


FIGURE 3.3

OVERVIEW OF MAJOR INPUT AND OUTPUT

Overview of Evaluation and Processing Modifications

Figure 3.4 provides an overview of the continuing evaluation which is used to monitor and, if appropriate, modify the processing of the Information Center so that it may more effectively meet its objectives. Input to evaluation is provided from information processing (PP (4)), user evaluation (PP (6)), the project officer and advisory board (PP (7)), and the Instructional Materials Center/Regional Media Center (IMC/RMC) Network (PP (8)). The arrows going in both directions indicate that there is an interaction between the evaluation component and other components. The input from the various sources is processed by the evaluation component to determine if there are system modifications which should be made. The decision process is illustrated by symbols SB (1), SB (2), and in the system modification occurring to PP (4). The numbers 1, 2, 3, and 5 appearing withing parentheses opposite arrows indicate that the same series of symbols; namely, SB (1), SB (2), and SB (3) would appear at these points. This would also be connected to the preprocessing procedures PP (1), PP (2), PP (3), and PP (5) as is done in the later diagram Figure 3.5. If no change is made this fact is provided as input to the evaluation component as indicated by the connection C (1) to PP (5).

Overview and Model of the Information Center's Operation

Figure 3.5 provides an overview of the Information Center's processing. In this overview six major activities can be seen in the Center of the flowchart. The input and output operation shown in Figure 3.3 are present as well as the evaluation procedures indicated in Figure 3.4. The model as presented indicates a continual flow of input,

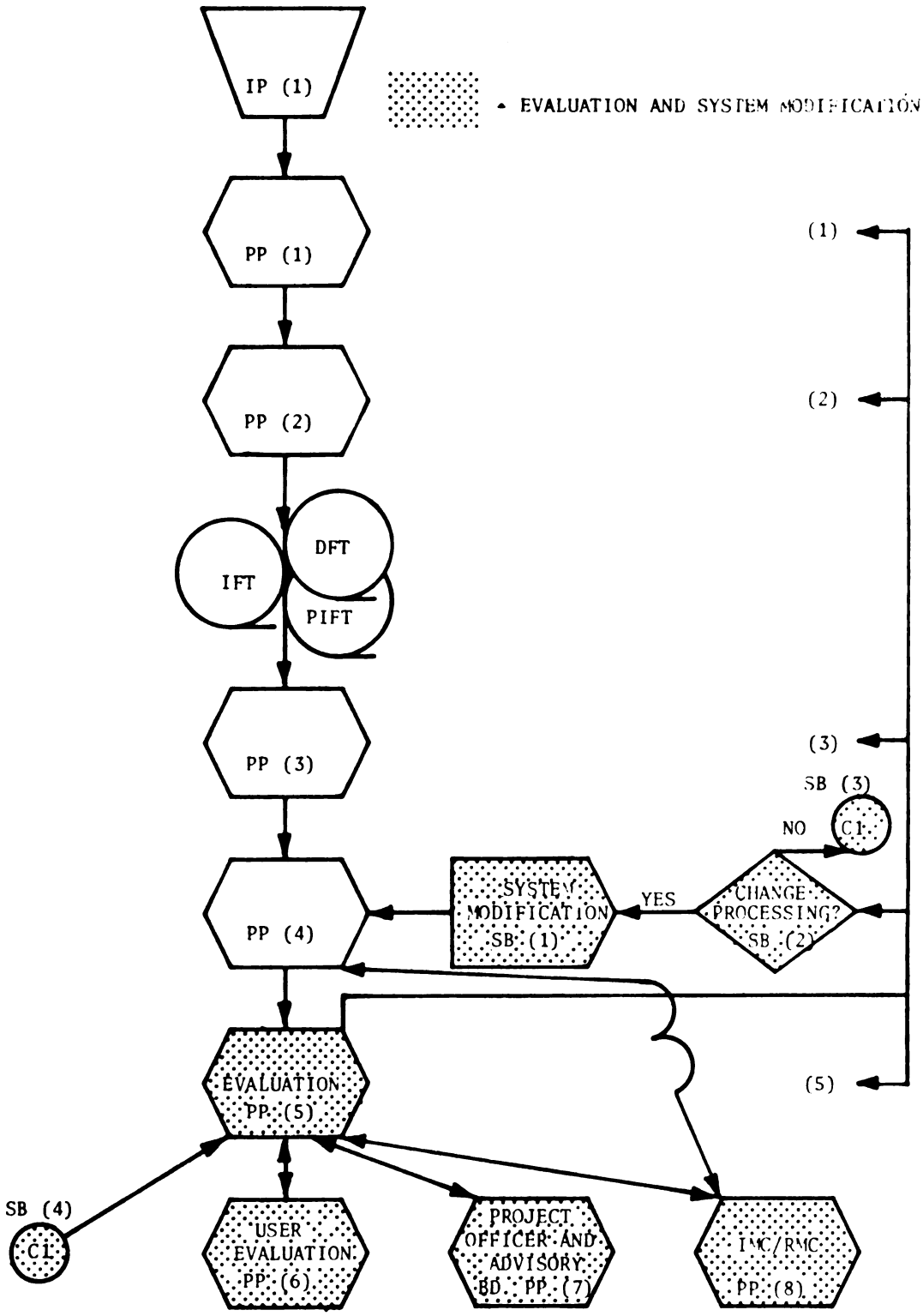


FIGURE 3.4

AN OVERVIEW OF THE INFORMATION CENTER'S
EVALUATION AND SYSTEMS MODIFICATION COMPONENTS

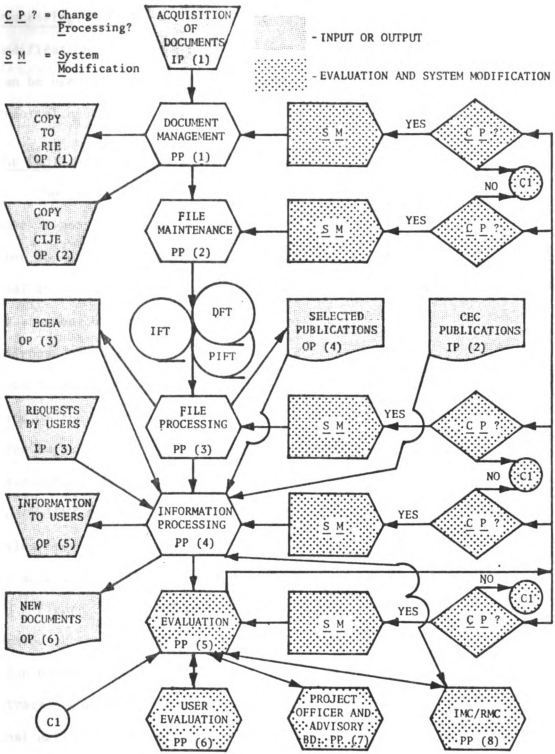


FIGURE 3.5

AN OVERVIEW AND MODEL OF THE INFORMATION CENTER'S OPERATIONS

processing, output, and evaluation, resulting in appropriate systems modification. Figure 3.5 and the simplified Figures 3.2, 3.3, and 3.4 can be used as a reference to the more detailed steps involved in the Information Center's operations which are discussed in Appendix A.

The Publication of Exceptional Child Education Abstracts

The original reason for placing abstracts on the computer was to make it possible to use a computer to assist in searching the information found in the abstracts. As previously indicated it was the original intent that the computer searchable files would be made available to a number of centers; however, an analysis of the potential effectiveness of this means for disseminating information led to the decision to publish ECEA. In the system design which followed, an examination of the technological capabilities available in the Washington area suggested that the most efficient way to have both computer searchable files and publish the abstract journal was to use computer controlled typesetting. Thus in this process the computer information files became the data source for both computer searching and printing of abstracts in ECEA.

Figure 3.6 illustrates a typical abstract set in the format generated by the computer controlled typesetting and provides a description of the various fields (types of information) contained in the abstract. The abstracts are stored on the information files in a manner that makes it possible to generate printed indexes for any of the fields identified in Figure 3.6 and to use some or all of the fields to generate computer-searchable description fields.

Figure 3.7 provides examples of portions of subject and author indexes which were indexed by the computer and output so that another

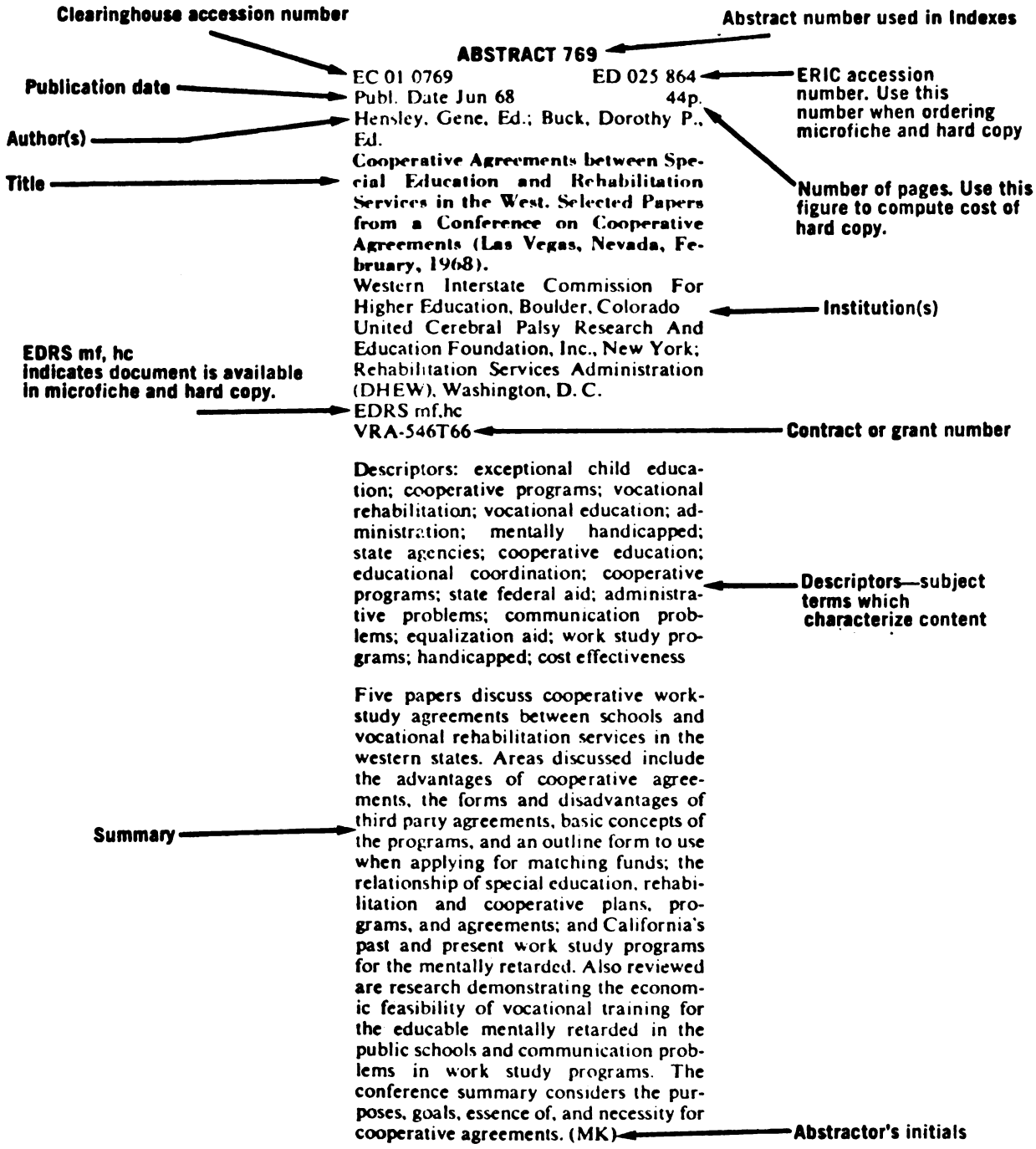


FIGURE 3.6

SAMPLE ECEA ABSTRACT

AUTHOR INDEX

Abbott, Margaret 314.
Abel, Georgie Lee And Others 421.
Abraham, Willard 51.
Ackerman, Nathan W And Others 55.
Adamson, T M 266.
Adler, Alfred 835.
Adler, Edna P, Ed 530.
Adler, Lenore Loeb 730.
Adler, Manfred 747.
Adler, Sol 661.
Ahlersmeyer, Donald E 214.
Aichhorn, August 141.
Albee, George W 903.
Aldrich, Robert A 718.
Alkema, Chester Jay 892-893.
Allen, K Eileen And Others 392.
Allen, Robert M 811.
Alonso, Lou And Others 609.
Alterman, Arthur I 85.
Amos, William E, Ed 574.
Anant, Santokh S 446.
Anderson, Donald T And Others 506-507.
Anderson, Jackson M 112.
Andrew, Gwen 751.
Annand, Douglass R 592.
Antinoro, Frank 407.
Apffel, James 235.
Arcieri, Libere And Others 859.
Armenti, Simma 20.
Armstrong, J D 727.
Arnold, Godfrey E 662.
Arthur, L J H 521.
Artuso, Alfred A And Others 548.
Asp, Carl W 18.
Attwell, Arthur A 704.
Ayers, George E, Ed 100.
Babow, Irving 869.
Bakwin, Harry 490.
Bakwin, Ruth Morris 490.
Banas, Norma 289-292.
Bandura, Albert 157.
Banks, Olive 77.
Bannatyne, Alex And Others 322, 379, 862.
Barbara, Dominick A, Ed 675.
Barden, John 223.
Barker, Roger G And Others 331.
Barman, Alicerose 508.
Barnard, James W, Ed And Others 865.
Barnes, Douglas And Others 540.
Barraga, Natalie 130.
Bartel, Nettie R 910.
Baruch, Dorothy W 741.
Bates, Karla K 20.
Baughman, M Dale, Ed 681.
Baumrind, Diana 958.
Beales, Philip H 739.
Becker, Howard S 670-671.
Beckett, Peter G S 496.
Beery, Keith E 203.
Beischer, N A And Others 265.
Bender, Ruth E 528.
Bennett, Merilyn Brottman 571.
Benson, F Arthur M, Ed 43.
Bentley, Ernest I 202.
Berger, Kenneth W 348.
Berger, Regina 939.
Bergsma, Daniel, Ed 922.
Berko, Frances G And Others 468.
Bernard, Jessie, Ed 65.
Bernardo, Jose R 318.
Berndt, Lois A 857.
Berner, George E And Others 948.
Bersani, Carl A 123.
Bettelheim, Bruno 212.
Bilovsky, David, Ed And Others 179
Bindman, Arthur J And Others 766.
Birch, Herbert G 955.
Birch, Herbert G And Others 594.
Blackhurst, A Edward 891.
Blair, John R 853.
Blank, Marion 575.
Blatt, Burton And Others 767.
Blessing, Kenneth R, Ed 35.
Block, James D And Others 280.
Blom, Gaston E 917.
Blom, Gaston E And Others 378.
Blue, C Milton 215.
Blum, Evelyn R 913.
Blum, Richard H And Others 950.
Bond, Guy L 996.
Bonner, J And Others 71-72.
Bonner, Ruth E 447.
Booker, Margaret 245.
Boone, Daniel R 585.
Boothroyd, Arthur 583-584.
Borowitz, Gene H And Others 907.
Bowden, M G, Ed 682.
Bowe, Frank 88.
Bowling, Wallace Lee 481.
Boyle, John 702.
Braaten, June 519.
Bradfield, Robert H, Ed 980.
Braley, William T And Others 352.
Bralley, Ralph C 357.
Braun, Samuel J 101.
Brearley, Molly, Ed 443.
Brenner, Harold J 371.
Brewer, Earl J, Jr 34.
Brewer, Jennie 676.
Breyer, Norman L, Ed 445.
Bricker, David D 390.
Bricker, William A 390.
Bright, George M 195.
Brill, Richard G 778.
Brittain, W Lambert 114.
Broadhead, Geoffrey D 282.
Broderick, Carlfred B, Ed 65.
Brogden, J D 556.
Brown, Doris V 946.
Brown, Sheldon S, Ed 183.
Brubaker, R S, Ed 492.
Brueckner, Leo J 996.
Bruininks, Robert H 870.
Brumbaugh, Florence N And Others 553.
Bryant, John E 205.
Buchholz, Sandra 125.
Buchwald, Edith 499.
Buktenica, Norman A 196, 466.
Bumstead, Richard 742.
Bunger, Anna M 712.
Burke, Donald A 425.
Burke, Douglas J N 777.
Burns, Robert C 106.
Burris, W R, Ed 604.
Burrows, Nona L 33.
Bush, Wilma Jo 368.
Butler, Katharine G 344.
Butler, Lucius 353.
Cain, Leo F 638.
Callaway, W Ragan 972.
Calovini, Gloria 623.
Calvert, Donald R 347.
Camp, Shirley L And Others 877.
Caplan, Gerald, Ed 997.
Carney, James R 542.
Carter, Charles H 812.
Carter, Darrell B, Ed 937.
Cass, Marion T 653.
Cassidy, Jean Trotter 695.
Cawley, John F 192.
Cawley, John F And Others 744.
Cazden, Courtney B 576.
Cegelka, Patricia A 470.
Cegelka, Walter J 470.
Chaiklin, Joseph B 417.
Chapman, Myfanwy E 818.
Chasey, William C 533.
Chazan, Maurice, Ed 66.
Chess, Stella, Ed 75.
Chethik, Morton 397.
Chisum, James 514.
Chomsky, Noam 369.
Christian, Floyd 306.
Christoplos, Florence I.
Clarcq, J R 780.
Clarizio, Harvey F 933.
Clark, Charlotte R 236.
Clark, Margaret M 836.
Cleland, Charles C 476.
Clyne, Max B 690.
Coggan, William G, Ed 185.
Cohen, Herbert J 758.
Cohen, Lisa 508.
Cohen, William J 949.
Cohler, Bertram J And Others 283.
Cohn, Maxine D 688.
Coleman, Jack L And Others 312.
Coles, Robert 529, 738.
Collins, James L 851.
Comly, Hunter H 894-895.
Conant, James Bryant 441-442.
Connelly, Elva A 511.
Conners, C Keith 277.
Connor, Leo E 333.
Conrad, R 351.
Contrucci, Victor J And Others 36.
Cooke, Robert, E, Ed 461.
Copel, Sidney L 809.
Cordova, Hector L 993.
Corrado, Joseph 568.
Cortazzo, Arnold D 915.
Cotten, Paul D, Ed 603.
Cowie, Valerie A 998.
Cox, Richard C 204.
Cox, T, Ed 67.
Craft, Michael 497.
Craig, William N 33, 851.
Crammatte, Alan B 931.
Cratty, Bryant J 143, 626, 642, 708-710.
Crawford, Gladys H 774.
Crickmay, Marie C 660.

FIGURE 3.7

SAMPLES OF ECEA AUTHOR AND SUBJECT INDEXES

SUBJECT INDEX

Ability Grouping 370.
Ability Identification 613.
Abortions 538.
Abstract Reasoning 478, 579.
Abstraction Tests 478.
Abstracts 240.
Academic Ability 566, 764.
Academic Achievement 87, 92, 156, 202, 238, 253, 260, 262-264, 324, 387, 430, 522, 549, 557, 566, 745, 764, 838, 859.
Academic Promise Tests 478.
Acceleration 156, 367, 370.
Accreditation (Institutions) 971.
Acculturation 180.
Achievement 158, 643, 914.
Achievement Gains 92.
Acoustic Phonetics 405-406.
Acoustics 727.
Activities 352, 996.
Activity Level 229.
Activity Units 636, 678, 707.
Adjective Check List 728.
Adjustment Problems 331, 436, 787.
Adjustment (To Environment) 24, 26, 35, 99, 154, 283, 318, 331, 362, 372, 689, 706, 718.
Adler, Alfred 835.
Administration 112, 127, 156, 271, 333, 385, 420, 486, 496, 587, 595, 619, 623, 641, 850, 916.
Administrative Change 767.
Administrative Organization 38, 77, 333, 428, 435, 587, 590, 593, 595, 639.
Administrative Policy 17, 38, 435, 554, 595, 600, 611, 620, 969.
Administrative Principles 271.
Administrator Attitudes 588.
Administrator Evaluation 271.
Administrator Guides 33, 435.
Administrator Intern 850.
Administrator Role 155, 333, 595, 623.
Admission Criteria 38, 126, 435, 600.
Adolescence 278, 831, 964.
Adolescents 11, 48, 76, 79, 89, 122, 133, 163, 195, 201, 216, 226, 236, 244, 254, 336, 375, 402, 448, 454, 482, 496, 695, 696, 815, 821, 837, 911, 960.
Adoption 538.
Adult Education 712.
Adults 23, 26, 89, 104, 246, 275, 334, 336, 359, 405-406, 411-412, 414, 432, 734, 764, 784, 846, 869.
Advanced Placement 553.
Affective Behavior 426, 897, 988.
Africa 716.
After School Activities 536.
Age Differences 278, 280, 284, 366, 583.
Age Groups 607.
Agencies 22, 815, 822, 891, 971.
Agency Role 269, 546, 597.
Aggression 141, 183, 302-304, 436, 762.
Agriculture 485.
Alcoholism 79, 107, 717.
Algebra 150.
Amblyopia Ex Anopsia 983.
American Indians 148, 668, 940.
American Literature 84.
Amphetamines 189.
Amputees 102, 178, 182, 386, 572, 803.
Anatomy 572, 648, 830.
Ancillary Services 385.
Anesthesiology 401.
Animal Behavior 461.
Anne Sullivan Macy Service For Deaf-Blind Persons 23.
Annotated Bibliographies 8, 59, 68, 144, 204, 220, 240, 309, 668-669, 796, 822, 935, 962.
Annual Reports 29, 40, 338, 340, 587.
Anomalies 83, 182, 187-188, 358, 722-725, 922.
Anxiety 104, 107, 160, 273, 473, 482, 666, 731, 741, 745, 809, 813, 881, 909.
Anxiety Scale For The Blind 482.
Aphasia 119, 180, 330, 411, 413-414, 537, 648, 656-657, 662, 665, 685, 784, 799.
Apraxia 248.
Architectural Barriers 424.
Architectural Programing 318, 848.
Architecture 318.
Area Centers For Services To Deaf Blind Children 625.
Arithmetic. See Mathematics.
Arizona 420.
Arkansas 625.
Arkansas School For The Deaf 73.
Art 106, 183, 730, 892-893.
Art Education 114, 618, 892.
Art Materials 114, 604, 618, 893.
Art Therapy 106, 892-893.
Articulation (Speech) 32, 113, 206, 208, 366, 409-412, 416, 492, 578, 660, 662, 924.
Asian History 999.
Aspiration 989.
Assistive Devices 102.
Associa-Math Program 289-291.
Association (Psychological) 730.
Associative Learning 16, 262, 264.
Asthma 182.
Athletics 82, 570.
Attendance 745.
Attendant Training 100, 568.
Attention Span 175, 398, 469.
Attitude Tests 535, 615.
Attitudes 55, 172, 255, 269, 375, 493, 506-507, 541, 615, 696, 768, 881, 931, 999.
Audio Equipment 67, 73-74, 186, 719, 868, 1000.
Audiology 120, 621, 797, 943.
Audiometric Tests 98, 349, 515, 656, 943, 959.
Audiometry 515, 621, 943, 959, 1000.
Audiovisual Aids 32, 132, 144, 149, 186, 216, 236, 286, 410, 414, 663, 668, 773, 822, 877.
Audiovisual Centers 50, 353.
Audiovisual Instruction 288, 585, 824.
Audition (Physiology) 297-298, 492, 530, 739, 797.
Auditory Agnosia 194.
Auditory Perception 98, 180, 311, 348, 366, 368, 407, 413, 417, 455, 471, 520, 543, 701, 912, 982, 1000.
Auditory Tests 98, 297, 415, 797, 943, 959.
Auditory Training 18, 44, 180, 791.
Aural Learning 524.
Aural Stimuli 98, 229, 471.
Aurally Handicapped 4, 7, 14, 18, 31, 33, 40-41, 44, 57-58, 73, 74, 84-88, 92, 98, 113, 132, 146-147, 149-152, 171, 181, 202, 208, 246, 253, 258, 294-298, 309-310, 331, 337, 347, 349-351, 373, 407, 415, 417, 423, 498, 512, 515-516, 528, 530, 551, 564, 583-585, 607, 616, 624, 656, 682, 700, 712, 719, 735, 739, 761, 763, 773-780, 787, 789, 791-792, 797, 824, 838, 851, 868, 931, 943, 952, 968, 970, 974, 981, 1000.
Australia 167, 692, 722, 984.
Authoritarianism 255.
Authors 84.
Autism 45, 222, 287, 487, 793-794, 813, 982.
Autobiographies 362, 503, 580.
Autoinstructional Programs 144.
Beginning Reading 69, 329, 512, 525, 527, 732.
Behavior 8, 277, 379, 436, 495, 617, 689, 978, 980.
Behavior Change 1-2, 8-9, 20, 43, 153, 157, 207, 228, 243-244, 251-252, 354, 383, 388-392, 394-396, 445-446, 596-597, 605, 617, 699, 715, 737, 752, 767, 771, 783, 821, 823, 827, 835, 842, 852, 857, 878, 898, 933, 949, 960, 980.
Behavior Patterns 123, 129, 243, 277-278, 303-304, 418, 439, 455, 469, 752, 762, 776, 779, 907.
Behavior Problems 65, 261, 278, 281, 387, 392, 395, 445, 448, 462, 468, 473, 490, 514, 605, 617, 670, 699, 704, 737, 745, 771, 827, 835, 878, 930, 933, 962.
Behavior Theories 6, 157, 239, 388, 396, 446, 597, 842, 897, 917, 980.
Behavioral Objectives 617, 946.
Behavioral Sciences 157, 229, 270, 393, 395.
Bender Gestalt Test 376, 382.
Bias 577.
Bibliographies 9, 31, 53, 144, 219-220, 241, 245, 309-310, 621, 668-669, 820-822, 879, 935.
Bibliotherapy 669.
Biochemistry 83, 308, 462, 808, 830.
Biographies 24, 111, 509-510, 612, 643, 700, 741, 831.
Biological Influences 265-266, 268, 814, 827, 972.
Biological Sciences Curriculum Study 165.
Biology 165, 889, 922.
Birth Defects 182.
Blackman, Leonard S 913.
Blind 191, 269, 475-478, 655, 703.

FIGURE 3.7 (cont'd)

computer could control the phototypesetting. Not illustrated in Figure 3.7, but included in Exceptional Child Education Abstracts since Volume III, No. 2 is a title index.

Selective Publication

The rapid expansion of knowledge has made it increasingly apparent that not only must better ways be found to store and retrieve information, but that also better ways must be found to organize knowledge. The computerized search provides a powerful tool for bringing together documents in a file that have similar information.

While computer searches can be used to retrieve information, the cost of retrieving information increases with the size of the file. The fact that the more such techniques are needed (for larger files) the more it costs provides an interesting paradox; however, not without solution. Analysis of user requests at the ERIC Information Center has indicated that there are often categories of similar requests. By categorizing the requests, it is possible to break large files into smaller subfiles by use of computer searches, thus reducing the cost of additional searches in special topic areas.

The manner in which the files are prepared for the CEC-ERIC Information Center not only makes it possible to create new subfiles, but to publish these subfiles. Thus, if there are a number of requests that could be answered by using the same document, it is possible to directly publish these documents using computer typesetting and a very inexpensive offset process. As of August, 1971 The Council for Exceptional Children had 59 separate bibliographies which have been published in this manner. The latest operating statistics indicate that

approximately one half of all user requests for information are being answered by the use of one or more of these bibliographies.[210] This procedure provides savings by:

1. Using a single search to organize information to answer many user requests.
2. Reducing personnel time for processing requests that can be answered by the printed bibliographies.
3. Reducing the cost of mailing a printed bibliography versus a computer printout abstract.

(The printed bibliographies may have as many as ten abstracts on the same amount of paper required to print a single abstract on the computer.)

As used by the CEC-ERIC Information Center the process of selected publication has provided a powerful technique for the organization of knowledge and a reduction of costs when compared with running individual computer searches. Thus by analyzing user needs it is possible to use the computer to serve individuals collectively with considerable reduction in cost as compared with using the computer to serve them individually.

This section has provided a brief overview of the Information Center operations which places the indexing process and evaluation in a total context. Because of their importance to the Information Center operation and their unique nature the most detail has been provided about the printing of ECEA, the selected publication of annotated bibliographies, and their use in answering information requests.

---

[210] Statistical information based on an analysis by Carl Oldsen, ECEA Editor, and his staff.

Descriptive Statistics about the Present Operating Status of the CEC-ERIC Information Center

The previous section described the procedures presently used at the Information Center and provided a model for continuing evaluation, development and modification of the operating system. This section describes the following categories of statistical information related to the present operating status of the Center:

1. The Center's holdings--types of documents and their subject content.
2. The rate of acquiring and processing documents.
3. Information request processing.
4. Operating costs.

The major objective of this section is to provide descriptive statistics about rate and scope of operations under normal conditions. Many changes in processing have occurred since the Center began operation; however, the changes have become less frequent since the publication of Exceptional Child Education Abstracts began in April, 1969. For this reason the information discussed is primarily concerned with data gathered since the initial publishing of ECEA, with greater emphasis placed on the more recent data.

The descriptive statistics presented are taken from data collected by the Information Center to monitor its operations and costs described in accounting or budgetary records.

The Center's Holdings--Types of Documents and Their Subject Content

It is the present policy of the Information Center that all documents acquired and processed will eventually become part of ECEA. Documents which were acquired by the Center before the publication

of ECEA, and felt appropriate for ECEA were included in Volume I and II. The documents not used in ECEA were discarded; thus the Center's total holdings are described by the abstracts in issues of ECEA.

An analysis of Volume I and II of ECEA indicates that 36.9% of the abstracts are of journal articles, 12.4% of research reports, 5.6% of curriculum guides, and 45.1% of books and other non-periodic documents. The information in Figure 3.8 describes the subject content of the 5,725 acquisitions that have abstracts in Volumes I and II of ECEA. No document was assigned to more than one of the categories represented in Figure 3.8 even though it contained information concerning multiple categories.[211]

Acquisition and Processing Rates

The early operations of the Information Center were not typical of normal processing rates because ordering and processing included documents found in the literature prior to the current year. Because of this the first two volumes of ECEA not only contain material from 1969-1970, but also considerable information published before these years. Beginning with Volume III almost all the information abstracted is recent material.

Carl Oldsen, ECEA editor, indicates that the Center is attempting to examine all sources of potentially relevant documents to special education and that processing has reached a steady rate of about 250 documents per month. Of these 250 documents approximately 50 were

---

[211]Statistical information based on an analysis of 5,715 acquisitions in Volumes I and II of ECEA performed by Carl Oldsen, ECEA Editor, and his staff.

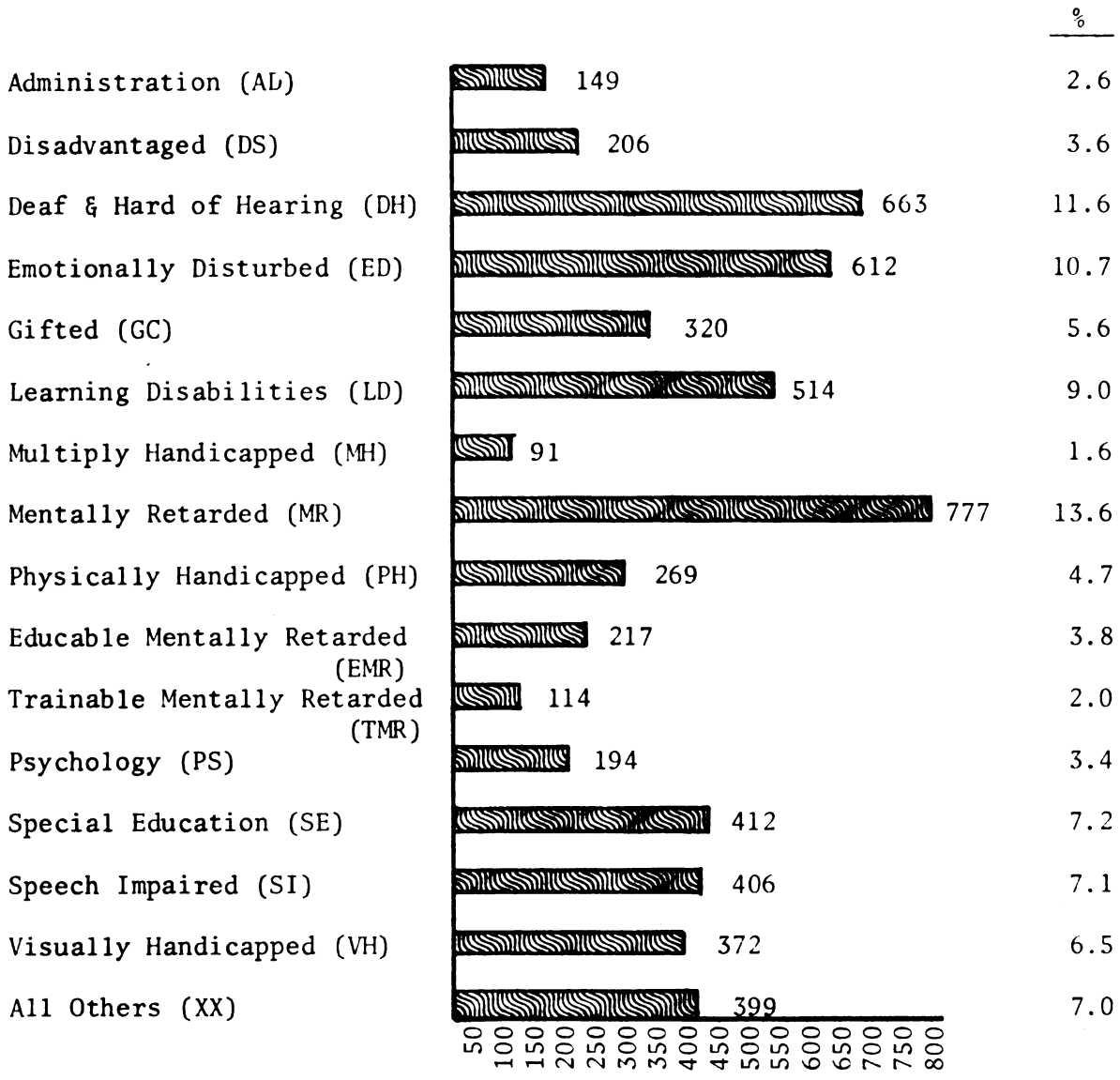


FIGURE 3.8

SUBJECT CONTENT DESCRIPTION OF INFORMATION CENTER HOLDINGS
BASED ON 5715 ACQUISITIONS IN VOLUMES I & II OF ECEA

processed for inclusion in Research in Education (RIE) and between 50 and 75 for inclusion in Current Index to Journals in Education (CIJE). All of the documents processed are eventually included in ECEA; thus, in an average issue of ECEA which appears quarterly one may expect to find about:

1. 150 document surrogates which also appear in RIE.
2. 200-225 document surrogates which are indexed only in CIJE.
3. 400-425 document surrogates which appear in neither RIE or CIJE.

Thus, of a total of 750 abstracts appearing in each issue of ECEA about 600 do not appear in RIE; however, about 225 of the 600 are indexed in CIJE.

Information Request Processing Statistics

While the processing of additions to the Center's holdings appears to have reached a stable rate, the number of user requests for information appears to be increasing as more individuals learn about the Center's capabilities. During the year of 1970 approximately 6,400 requests for information were received and processed by the Center. Of these about 21.4% were processed during the first quarter, 30.8% during the second quarter, 22.3% during the third quarter and 25.5% during the fourth quarter.[212] During the first quarter of 1971 2,176 information requests were processed as compared to 1,380 during the first quarter of 1970. If it were assumed that the first quarter represented a fourth of the information requests that will be processed during 1971,

---

[212]CEC-ERIC Information Center, "Processing Costs & Formulas," an unplublished summary prepared by the Center, September, 1970, under the direction of Carl Oldsen.

the projected total would be about 8,700 requests, a projected increase of about 36% over the previous year. Even though the number of requests are increasing, the Center has been able to cope with this without increasing staff through greater use of the computer and the computer-generated bibliographies.

Table 3.1 indicates the number of requests received, the type of responses made, and the type of users making the requests during the first quarter of 1971. There are more responses than requests because some requests have several questions which require different responses. The information related to user categories, types of responses, and the way in which requests were received appears to be similar to previous quarterly reports except for a general increase in all categories.

Processing Costs

The figures that are presented in this section have been taken from operations for the calendar year 1970 which included the majority of the processing for Volume II of ECEA. In determining the cost for various operations all salaries, supervisory time, rental of office space, supporting services and miscellaneous overhead items were included. For example, to obtain the cost of acquisitioning documents, the total number of documents processed in Volume II of ECEA was divided into the total cost of salaries, purchasing the documents, supervisory overhead, and supportive clerical functions. The total cost for these functions was $36,150 divided by 3,615 documents--resulting in an average cost of $10 per document.

Cost of Abstracting, Indexing, and Cataloging Calculating abstracting, indexing, and cataloging costs as described above results

TABLE 3.1

AN ANALYSIS OF INFORMATION REQUESTS PROCESSED BY THE CEC-ERIC INFORMATION CENTER DURING THE FIRST QUARTER, 1971

Total Requests Made to Clearinghouse During Report Period

	Jan.	Feb.	Mar.	Total for Qtr.	% for Qtr.
Phone	50	57	49	156	7.7
Letter	573	601	816	1990	91.3
Visits	14	7	9	30	1.4
TOTAL	637	665	874	2176	100
Types of Responses					
Reference - nonsubject	425	295	260	980	16.6
Reference - subject	566	667	908	2141	36.4
Spot bibliographies & literature searches	597	625	725	1947	33.1
General question on ERIC	5	17	27	49	.8
Other (Including mailing list)	235	266	271	772	13.1
TOTAL	1828	1870	2191	5889	100
General Breakdown of Users					
Teachers	103	86	81	170	12.4
Teacher educators	59	73	78	210	9.6
Supervisors & Consultants	17	9	16	42	1.9
Psychologists & Social Workers	11	11	6	28	1.3
Educational decision makers	88	54	73	215	9.9
Research & Development Specialists	13	10	7	30	1.4
Information professionals & dissemination specialists	51	45	58	154	7.1
Professional organizations	48	42	44	133	6.1
Students	166	254	399	819	37.7
Federal Gov't. & Public Agencies	11	17	24	52	2.4
Parents	19	14	29	62	2.8
Unidentified	51	51	59	161	7.4
TOTAL	637	665	874	2176	100.0

in the following: (1) The cost of abstracting, indexing, and cataloging 3,615 documents (not including special processing for RIE or CIJE) averaged $22.20 per document. (2) The cost of special processing of 600 documents for use in RIE, statistical reports prepared for ERIC, and all supportive functions averaged $24.50 per document. (3) The special processing of 1,000 documents which were indexed for use in CIJE averaged $7.60 per document. (4) If the special processing related to CIJE and RIE were included in the total cost of abstracting, indexing, and cataloging, and distributed over the total 3,615 documents, the average of these functions was $27.50 per document.

Cost of Answering Search Requests The cost of answering 6,000 search requests averaged $10 per request or $5 per response. The reason for the lower cost per response is because on the average a user's request for information requires two responses, i.e., an average of two questions are asked in a single user request for information. These costs include all personnel costs, overhead figures, computer costs for running searches and mailing costs. They do not include the cost of documents or special materials that were sent in response to requests. Obtaining an accurate value for these materials is difficult because many of them are obtained free; however, estimates suggest the value of materials sent averaged $3 to $5 per response.

Costs Related to Printing ECEA Of particular interest to the computerized publication operation is the cost of printing ECEA. The cost based on a per abstract average for printing 2,000 journals was reported as:

ECEA, per abstract

Keypunch	- Computer	- Photodata*	- Printing		
1.40	5.00	2.10	3.15	=	$11.65

ECEA - Printing
$3.15 per abstract
$14.30 per page

Photo-Data typesetting
$2.10 per abstract
$8.00 per page

Computer Time
$5.00 per abstract[213]

Based on 3,615 abstracts and 2,000 copies of Volume II of ECEA, the printing costs average approximately $21 per single copy of one volume. Included in the computer costs were costs for building information files; generating special author, title, and subject indexes for ECEA; and building computer searchable description files for use in answering user requests for information.

The figures presented are pessimistic in that they have attempted to include every item that might possibly be related to the specified cost. In attempting to compare these figures with similar operation, it would be necessary to have information about how both sets of figures were calculated as well as information concerning the quality and quantity of services offered by the different information centers.

Summary

The CEC-ERIC Information Center had its origin with the Educational Resource Information Center (ERIC) Network and was established and began operation early in 1967. Many of the procedures developed at

---

[213] Ibid.

*Computer controlled typesetting.

the Information Center were influenced by the development of ERIC, including the indexing processes which use the ERIC Thesaurus.

In meeting the unique needs of this specific Information Center, procedures for selective publication utilizing computer searches and computer-controlled phototypesetting have been developed. These procedures augment computer and hand searching which are ordinarily used in answering requests at information centers.

The number of documents being processed has become relatively stable and appears to be approximately 3,000 documents per year (all documents processed have abstracts in ECEA). The number of requests for information has been steadily increasing; however, it appears there is not sufficient data to estimate the future rate of information request processing.

The costs which are described are those most common to information center processing or those unique because of the computerized publication done by the CEC-ERIC Information Center. The process involved in calculating these costs attempted to take a conservative approach which included all related and overhead costs.

CHAPTER IV: PROCEDURES USED IN THE EVALUATION AND ANALYSIS OF THE INFORMATION CENTER INDEXING METHODS

The first two of six objectives stated in Chapter I--to document the development of the information system used by the CEC-ERIC Information Center and to document the manner in which CEC-ERIC Information Center uses the BIRS system and other computerized programs[215]--were accomplished in the previous chapter. The information provided in this documentation described the environment in which the third objective, the evaluation of the indexing methods used by the CEC-ERIC Information Center, took place. The objective of this chapter is to describe the procedures which were used in this evaluation, with the results and interpretation of the evaluation being reported in the following chapter. The remaining three objectives related to recommendations for the Information Center's operation as well as implications for similar studies are examined in the last chapter.

The procedures described in this chapter had two distinct phases which took place simultaneously. One phase involved the evaluation of the indexing methods used in Volume I of ECEA as determined by measures of Average Macroprecision, Average Microprecision and estimates of average recall for questions written to retrieve randomly selected target documents.

---

[215]Additional information concerning the Information Center's computer processing is found in Appendix I.

A second phase of the procedures involved:

1. A comparison of the vocabulary of the terms used in indexing Volume I of ECEA with (a) the vocabulary of the collective titles of the document surrogates included in Volume I, and (b) the vocabulary of the Thesaurus developed by Samuel Price[216] for indexing documents in special education.
2. An analysis of the frequency with which different indexing terms were assigned to the documents abstracted for Volume I of ECEA.
3. An analysis of ambiguity in assignment of terms having similar meanings.
4. The development of a subset of the ERIC Thesaurus to be used in indexing of successive volumes of ECEA.
5. A preliminary evaluation of the effect of applying refined procedures and a subset of the ERIC Thesaurus to Volume II.

The Evaluation of the Indexing Procedures Used in Volume I of ECEA

Presently all requests for information which require computer searches are processed by one of several Information Center staff members who translates the user request into a computer-searchable question. These staff members and those involved in indexing and abstracting have become familiar with the terms in the ERIC Thesaurus which were used in indexing Volume I of ECEA. The computer indexing now used in preparing description files for searches uses terms assigned by the indexer and

---

[216]Samuel T. Price, Thesaurus of Descriptors for an Information Retrieval System in the Subject Matter Area of Special Education (Normal, Illinois: Illinois State University, Special Education Instructional Materials Laboratory, 1970), pp. 1-465.

terms extracted from the titles. The BIRS programs have the ability to use other computer indexing methods where success in computer searching may be less dependent upon a person's knowledge of the types of terms assigned by the Center's indexers.

Questions Examined

The following questions were examined taking into consideration the above conditions:

Question 1 As measured by Average Macroprecision, Average Microprecision, and estimated average recall, how effective is the indexing method used by the Information Center for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?
b. Professional educators who are unfamiliar with the Information Center's indexing system?

Question 2 How effective is a computerized indexing method using terms extracted from the title and abstracts for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?
b. Professional educators who are unfamiliar with the Information Center's indexing system?

Question 3 How effective is the indexing method used at the Information Center when combined with machine indexing of abstracts for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?
b. Professional educators who are unfamiliar with the Information Center's indexing system?

These questions correspond to Questions 1, 2, and 3 on pages 8 and 9 of the introduction. In the computerized indexing methods considered, the following three sources of indexing terms were used:

Source 1 Indexing terms selected from the ERIC Thesaurus to describe the documents abstracted for Volume I of ECEA. These terms were manually selected by indexers and are called descriptors.

Source 2 Words extracted by the computer from titles of the documents whose abstracts were included in Volume I of ECEA.

Source 3 Words extracted by the computer from the abstract (summary) of the document surrogates.

When the computer is used to extract terms from the title or abstract, all terms not appearing on an exclusion list containing such words as a, an, the, and, or, at, by, of, with, and other non-descriptive articles, adjectives, conjunctions or prepositions are included as indexing terms. The three indexing methods employed in this evaluation used the following combinations of terms selected from the above sources:

Indexing Method 1 Terms from titles and descriptors.

Indexing Method 2 Terms from titles and abstracts.

Indexing Method 3 Terms from titles, abstracts, and descriptors.

Indexing Method 1 corresponds to the indexing procedures examined in Question 1, Indexing Method 2 to those examined in Question 2, and Indexing Method 3 to those examined in Question 3. The evaluation of these indexing methods used measures of Average Macroprecision, Average Microprecision, and estimates of average recall. These measures were

used to compare the search results of questions written to retrieve target documents by CEC-ERIC staff versus those written by professional educators.

Selection of Target Documents

The data base for the evaluation of the indexing procedures used at the CEC-ERIC Information Center was 2100 abstracts (The information contained in each abstract is illustrated in Figure 3.9, page 133) conitained in Volume I of ECEA. A stratified random sample of 105 documents was selected by using random number tables to specify five documents from each sequence of 100 documents. For example, five abstracts were randomly selected from abstracts 1 through 100, five from 101 to 200, five from 201 to 300, and so on. The documents thus selected were placed on an information file which was indexed to create a computer searchable description file for each of the three methods described in the previous section.

Preparation of Questions to Retrieve Target Documents

Two sets of 105 questions were written to retrieve the 105 target or source documents. The first set of questions was written by ten CEC-ERIC Information Center staff members who were familiar with the indexing procedures at the Center. The second set was written by seven professional educators from Andrews University who were not familiar with either the indexing procedures used at the Information Center or the ERIC Thesaurus which was used by indexers. Each of the professional educators had a doctor's degree in education and was familiar with the terminology used in the field of special education.

The types of questions written by the two groups were logical search questions using the operators .AND., .OR., and .NOT. to combine words or phrases in the manner described in pages 82 through 86. Both groups were given training and technical assistance until they were able to demonstrate a proficiency in writing logical search questions.

The 105 document surrogates which served as the basis for writing search questions contained only a title and a summary of a document. For each surrogate given to an individual he was instructed to write one question to retrieve the document described or similar documents. Those writing questions were told that the descriptions contained in the title and summaries represented ideal answers to requests for information that individuals might have sent to the Information Center. Neither group had any prior information concerning the manner in which the documents had been indexed, nor were they assured that the document surrogates which served as a basis for writing the question would be in the file that would be searched. Both groups knew that multiple indexing methods had been employed in creating the computer searchable description files but were unaware of how those files were made.

The basic reason for having two groups write questions was to provide one set of questions generated by individuals familiar with the Center's indexing procedures (the CEC-ERIC Information Center staff members) and a second set by individuals who were not familiar with the indexing procedures. The evaluation was in no way meant to examine the skill of the individuals in writing questions. The training was designed to develop the skills of two groups of individuals so that the major difference as related to this study was their degree of knowledge about the indexing language used at the Information Center.

Relevance Judgments

Three judges familiar with writing logical search questions and with educational literature were used to rate the relevance of the responses to specific questions. In rating the relevance they gave a rating of 0 if the document surrogate retrieved by a question had no relationship to that question, 1 if it had a moderate relationship to the question, and 2 if it had a very direct and obvious relationship to the question. When rating the relevance they were provided with the two sets of questions with a list of documents each question retrieved. They were instructed to read a single abstract and then compare it with all questions which had retrieved that abstract; thus, Abstract 1 was read and rated for all questions retrieving Abstract 1 followed by Abstract 2 and so on.

The sum of the ratings of the three judges had possible values ranging from zero to six. An abstract was considered to be relevant to a question if the combined score of the judges was three or greater and if a document obtained a rating of 3 it was the result of 3 one's. Thus if a document obtained a rating of 3 by means of one judge rating it 0, a second rating it as 1, and a third rating it as 2, it was not considered as relevant.

Measurement Techniques Employed in the Indexing Evaluation

The units of measurement used in this study were based on the retrieval results for the search question. Each question was considered as a question written either by a person who was familiar or one who was not familiar with the Information Center's indexing methods. Aside from these differences all questions were considered to have been written by individuals with equivalent skills.

Figure 4.1 illustrates the type of data that was collected and the descriptive statistics used in measuring the effectiveness of the previously described indexing methods. The procedures used for calculating Average Microprecision and Average Macroprecision are described in the related research. The calculations of estimated average recall were obtained by dividing the number of target documents retrieved by the total number of target documents which could be retrieved. For example, if 80 of a possible 105 target documents were retrieved the estimated recall would be 80 divided by 105.

For each indexing method a chi-square test was applied to determine (1) if the number of target documents retrieved with one set of search questions was significantly different from the number of target documents retrieved by the other set of questions and (2) if the ratio of relevant to non-relevant documents retrieved with one set of questions was significantly different from the ratio retrieved by the other set. The null hypothesis in each case asserted that there was no significant difference at the .01 level.

The Content Analysis of the Vocabulary Used in Indexing Volume I of ECEA

Because of the relationship which exists between the Information Center and the ERIC Network the descriptive terms assigned by indexers have come from the ERIC Thesaurus. In addition to the terms from the Thesaurus other "identifier" terms which usually contained information such as names of institutions, names of specific tests, or geographic location were assigned.

The ERIC Thesaurus was developed by a group of experts from many

	CEC-ERIC STAFF	PROFESSIONAL EDUCATORS	DESCRIPTION OF DATA AND STATISTICS
METHOD 1			Total No. of Documents Retrieved
			No. of Relevant Documents Retrieved
			Average Microprecision
			Average Macroprecision
			No. of Target Documents Retrieved
			Estimated Average Recall
METHOD 2			Total No. of Documents Retrieved
			No. of Relevant Documents Retrieved
			Average Microprecision
			Average Macroprecision
			No. of Target Documents Retrieved
			Estimated Average Recall
METHOD 3			Total No. of Documents Retrieved
			No. of Relevant Documents Retrieved
			Average Microprecision
			Average Macroprecision
			No. of Target Documents Retrieved
			Estimated Average Recall

FIGURE 4.1

A DESCRIPTION OF DATA AND DESCRIPTIVE STATISTICS USED COMPARING VARIOUS INDEXING METHODS

areas of education[217] with the result that there are often terms having closely related meanings. Because of the multiple possibilities for indexing the same concept, two indexers have often used different terms for indexing the same idea, thus making the retrieval of documents more complicated.

The way in which the ERIC Thesaurus was developed also raised questions about how well the terms selected from that Thesaurus to index Volume I of ECEA represent the vocabulary used in the field of special education. Specifically the procedures in this section were developed to examine the question, "Is the vocabulary of the terms used in indexing Volume I of ECEA found in the literature of special education?"

Compilation of Indexing Terms Assigned to Volume I of ECEA

As a part of the indexing procedures used in preparing ECEA for publication, indexing terms are selected from the ERIC Thesaurus and assigned to the descriptors field. A list of these terms assigned by the various indexers to abstracts contained in Volume I was compiled, including information concerning when the term was first used, and the number of times the term was used. For example, "16--MEDICAL RESEARCH--166" would indicate that the term MEDICAL RESEARCH was used 16 times in the 2100 abstracts found in Volume I and that it was first used in Abstract No. 166.

---

[217]James L. Eller and Robert L. Panek, "Thesaurus Development for a Decentralized Information Network," American Documentation, July, 1968, p. 213-220.

Subjective Analysis by Indexers of Terms Used in Volume I of ECEA

The list of terms used in Volume I of ECEA served as the basis for a subjective analysis by the indexers. To assist in this analysis each term was put on a single IBM card with the information concerning the number of times the terms was used and the abstract to which the term was first assigned. Also a Key-Word-In-Context index was prepared to aid in grouping terms with similar meanings.

The indexers examined each term that had been assigned to Volume I of ECEA, comparing that term with similar terms. If the indexers could not establish significant differences in the meaning of similar terms, a decision was made concerning which of the terms should be kept on a list for use in future indexing. If it was felt by a professional staff member that one term was generally used more often than the other in the special education vocabulary, this term was kept. If, however, there was no preference, the term which had been used most often in the indexing of Volume I was kept. The result of these procedures was a subset of terms used in indexing Volume I of ECEA which has served as an authority list for indexing successive volumes of ECEA.[218]

A Comparison of the Word Vocabulary Used in the Indexing Terms of Volume I of ECEA with Words Extracted from the Literature

Two lists of words extracted from the literature of special education were used for comparisons of words found in the descriptive terms used in indexing Volume I of ECEA. The first list of words was the collective vocabulary found in the 2100 titles of documents abstracted for Volume I of ECEA. The second list of words was the vocabulary of

---

[218]Thesaurus for Exceptional Child Education (Arlington, Virginia: CEC-ERIC Information Center on Exceptional Children, 1971), 12 pp.

the terms in a Thesaurus prepared for use in special education by Samuel Price. This Thesaurus was developed by using a computer to extract terms from the literature of special education, thus providing the Thesaurus an empirical base.[219] The words found in the titles and in the Thesaurus of Samuel Price were compared with the words in the indexing terms used by the Information Center by means of intersections of the various lists.

Proper nouns, conjunctions, prepositions, and other non-descriptive function words were removed from the word lists before comparisons were made. The letters A, B, C, and D are used as follows to stand for lists involved in the comparisons or other types of analysis:

A = All content words found in the titles of Volume I of ECEA

B = All content words found in the Thesaurus developed for special education by Samuel Price.

C = All content words found in the indexing terms assigned to documents in Volume I of ECEA.

D = All content words found in the reduced list of indexing terms developed through the analysis of the indexing terms used in Volume I of ECEA.

In addition to these word lists, lists of word roots were generated by a special computer program. This program was able to take a list of words and reduce each word to a root which is used in computer searches involving the BIRS programs. For example, ACCELERATED, ACCELERATING, and ACCELERATION would all reduce to the root, ACCELERATE. When one of the above lists is reduced to root form, it will be referred to with an

---

[219]Samuel T. Price, "The Development of a Thesaurus of Descriptors for an Information Retrieval System in Special Education," (unpublished doctoral dissertation, University of Pittsburgh, 1969), abstract.

R in front of the letter; thus the list of all roots from list A will be referred to as RA, from B as RB, from C as RC, and D as RD. Using this terminology with the symbol $\cap$ to stand for "intersection", $\cup$ to stand for "union" and n(S) to stand for the number of objects in the set, the following ratios were examined:

1. $\frac{n(A \cap B)}{n(A \cup B)}$

2. $\frac{n(RA \cap RB)}{n(RA \cup RB)}$

3. $\frac{\text{The number of words in A} \cup \text{B which also have a root in RA} \cap \text{RB}}{A \cup B}$

4. $\frac{n(A \cap C)}{n(A \cup C)}$

5. $\frac{n(RA \cap RC)}{n(RA \cup RC)}$

6. $\frac{\text{The number of words in A} \cup \text{C which also have a root in RA} \cap \text{RC}}{A \cup C}$

7. $\frac{n(B \cap C)}{n(B \cup C)}$

8. $\frac{n(RB \cap RC)}{n(RB \cup RC)}$

9. $\frac{\text{The number of words in B} \cup \text{C which also have a root in RA} \cap \text{RC}}{B \cup C}$

The ratio of terms found in the intersection versus the number of terms found in the union of two lists was used as a basis for comparing the similarity of the vocabularies of various lists. Specifically, the ratios involving the intersections of the two lists which were extracted from the literature was used as a basis for determining what proportion of similar words or roots might be expected in such intersections. This result was then compared with the result of other intersections involving the non-empirically-based words from the indexing terms used in Volume I of ECEA.

Analysis of the Vocabulary Used in Writing Questions to Retrieve Target Documents

Two comparisons were made of the question vocabulary used by the professional educators with the question vocabulary used by the CEC-ERIC staff. The first examined the proportion of phrases (terms with more than one word) versus single word terms. The second examined the proportion of question terms contained in the ERIC descriptors used by each group. A chi-square test was used to determine if there was a significant difference at the .01 level between the two groups for each of the comparisons. Because the professional educators had not seen the ERIC terms used to index Volume I of ECEA their use of these terms would tend to support the assertion that these terms are used in the field of special education.

Analysis of Changes in Indexing Procedures Between Volume I and Volume II of ECEA

As a result of the analysis done by the indexers on the vocabulary used for indexing Volume I of ECEA some changes in indexing procedures were implemented in Volume II. These procedures involved reducing the number of terms assigned to index a specific document and using a more controlled list of terms for indexing later portions of Volume II. The list of terms used was the subset developed in the analysis of the terms used in Volume I. The results of this analysis was not available until approximately the first half of Volume II had been indexed; thus this list was used only in the last half.

In indexing Volume I there had been a general rule that an indexer should assign any indexing term which was possibly related to the content of a document even if the document contained only a small

amount of information related to the term. The results of computer searches had indicated that many documents were retrieved which did not contain sufficient information about the question to be useful. Subsequently, indexers decided that in Volume II a term would not be assigned to a document unless the document contained considerable information related to that term.

This change was partially a result of the Information Center's use of computer searching to assist in answering user requests. When the information files were relatively small all documents with any information about a topic might be sent. As the information files grew those answering requests often read the abstracts and reduced the number sent to only those abstracts that contained considerable information about a requested topic.

To compare the effect of reducing the number of terms assigned to describe a document, identical searches were performed on Volume I and Volume II of ECEA. These searches were a part of the normal processing done to develop the selective bibliographies. The results of the searches were edited by the person in charge of producing the bibliographies to assure that only documents containing considerable information about the topic were retained. The person who wrote the question and did the editing was unaware that the results would be used in an analysis of the indexing procedures. An analysis for each search compared the precision for documents extracted from Volume I with the precision of documents extracted from Volume II. A sign test was used in examining 21 searches to determine if the precision of documents retrieved from Volume I versus Volume II was significantly different at the .01 level.

Summary

The purposes of the procedures described in this section were to:

1. Determine how effective various computerized indexing methods available for use by the CEC-ERIC Information Center were with:
   a. Staff members familiar with the indexing language used at the Center.
   b. Professional educators who are not familiar with the indexing language.
2. To examine the vocabulary used in the indexing terms assigned to Volume I of ECEA to determine if it was similar to that which was extracted by various means from the literature of special education.
3. To determine whether changes in indexing procedures between Volume I and II of ECEA had affected the precision of searches made on the two volumes.

The results of using these procedures is reported in the following chapter.

CHAPTER V: RESULTS OF THE EVALUATION AND ANALYSIS OF THE INFORMATION CENTER INDEXING METHODS

This chapter is divided into four major sections, each of which describes the results for a specific aspect of the total study. The procedures for the first section of this chapter are described in the first section of Chapter 4 while the procedures for the last three sections are discussed in the second section of Chapter 4.

The first section describes the results of an evaluation of the indexing method used in Volume I of ECEA and compares its effectiveness with two alternative methods. The measures of effectiveness of the three indexing methods are based on the computerized retrieval of randomly selected target documents. The computer search questions used to retrieve the documents were written by CEC-ERIC staff members familiar with the Center's indexing method, and professional educators who were unfamiliar with the Center's indexing method. The statistical measures used in the comparison are Average Macroprecision, Average Microprecision, and estimated average recall.

The second section describes the analysis of the indexing vocabulary based on a comparison of the vocabulary found in the ERIC terms used in indexing Volume I of ECEA with:

1. The vocabulary found in the collective titles of the document surrogates included in Volume I.
2. The vocabulary found in a thesaurus developed by Samuel

Price for indexing documents in special education.[220]

The third section describes a subjective analysis, done by indexers, of the ambiguity in assignment of similar ERIC descriptors to Volume I of ECEA. This analysis resulted in the development of a subset of the ERIC Thesaurus which has been used in indexing successive volumes of ECEA.[221]

The fourth section describes the preliminary results of an evaluation of the effect of applying refined indexing procedures to Volume II of ECEA. This evaluation was based upon the precision of 20 search questions which were written to retrieve documents from both Volume I and Volume II. The searches were part of the Center's normal processing done to identify abstracts to be included in selected bibliographies.

A Comparative Evaluation of Three Indexing Methods

A detailed description of the procedures used to obtain the results described in this and following sections is found in the previous chapter under appropriate subheadings. The procedures used in calculating Average Macroprecision, Average Microprecision, and estimated average recall are discussed on pages 63 through 70 of the review of related literature.

---

[220]Samuel Price, Thesaurus of Descriptors for an Information Retrieval System in the Subject Matter Area of Spedial Education, (Normal, Illinois: Illinois State University, Special Education Instructional Materials Laboratory, 1970), pp. 1-465.

[221]Thesaurus for Exceptional Child Education, (Arlington, Virginia: CEC-ERIC Information Center on Exceptional Children, 1971), 12 pp.

Questions Examined

The results reported in this section relate to three questions first stated on pages 8 and 9 of the introduction, and then later restated in an expanded form on page 133 of Chapter 4. The specific questions examined are:

Question 1 As measured by Average Macroprecision, Average Microprecision, and estimated average recall; how effective is the indexing method used by the Information Center for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?

b. Professional educators who are unfamiliar with the Information Center's indexing system?

Question 2 How effective is a computerized indexing method using terms extracted from the titles and abstracts for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?

b. Professional educators who are unfamiliar with the Information Center's indexing system?

Question 3 How effective is the indexing method used at the Information Center when combined with machine indexing of abstracts for:

a. CEC-ERIC staff who are familiar with the Information Center's indexing system?

b. Professional educators who are unfamiliar with the Information Center's indexing system?

Indexing Methods Compared

The results relating to these three questions are found in Table 5.1 with specific portions of the results presented graphically in Figures 5.1, 5.2 and 5.3. In the Figures and Tables indexing methods 1, 2, and 3 correspond to the indexing methods employed in questions 1, 2, and 3. Specifically, the indexing methods may be defined as follows:

Indexing Method 1 This method used terms manually assigned from the ERIC Thesaurus by the indexers and terms extracted by the computer from the titles of the document surrogates.

Indexing Method 2 This method used terms extracted by the computer from the titles and abstracts of the document surrogates.

Indexing Method 3 This method used terms manually assigned from the ERIC Thesaurus by the indexers and terms extracted by the computer from the titles and abstracts of the document surrogates.

Results of the Comparison of Indexing Methods

The examination of the three previously stated questions implies three corollary questions which consider whether or not there is a significant difference between the effectiveness of the three indexing methods when the search results of questions written by the CEC-ERIC staff members are compared with the search results for questions written by professional educators.

In examining these corollary questions a chi-square test of significance was used to test the following three null hypotheses:

TABLE 5.1

DESCRIPTIVE STATISTICS RESULTING FROM THE EVALUATION OF THREE INDEXING METHODS

	CEC-ERIC Staff	Professional Educators	Description of Data and Statistics
	209	143	Total No. of Documents Retrieved
METHOD 1 Terms from Titles and Descriptors	175	135	No. of Relevant Documents Retrieved
	.836	.945	Average Microprecision
	.960	.945	Average Macroprecision
	77	57	No. of Target Documents Retrieved
	.732	.54	Estimated Average Recall
	133	206	Total No. of Documents Retrieved
METHOD 2 Terms from Titles and Abstracts	102	191	No. of Relevant Documents Retrieved
	.767	.927	Average Microprecision
	.900	.973	Average Macroprecision
	38	81	No. of Target Documents Retrieved
	.362	.77	Estimated Average Recall
	305	296	Total No. of Documents Retrieved
METHOD 3 Terms from Titles Descriptors and Abstracts	227	265	No. of Relevant Documents Retrieved
	.743	.895	Average Microprecision
	.905	.974	Average Macroprecision
	85	84	No. of Target Documents Retrieved
	.81	.80	Estimated Average Recall

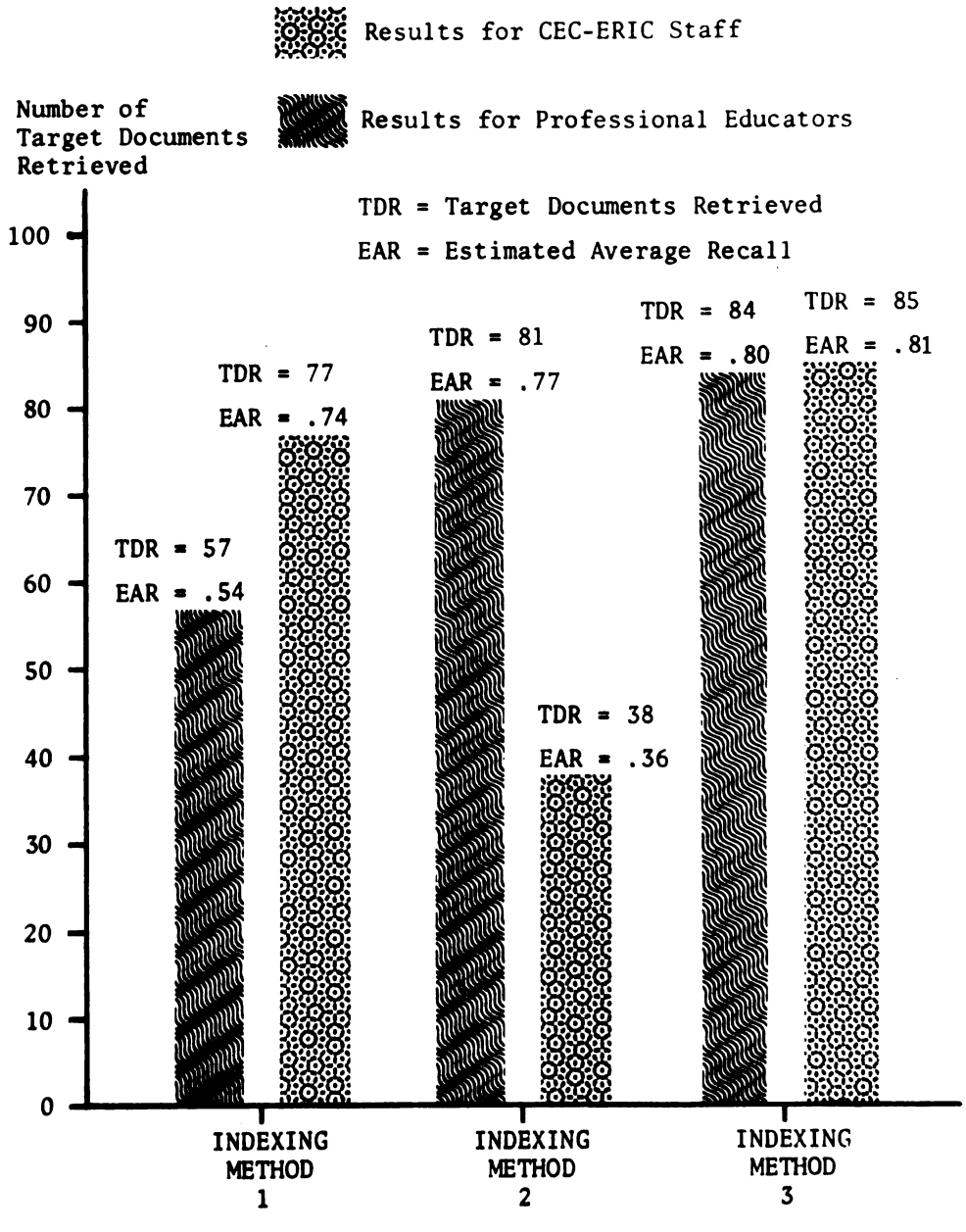


FIGURE 5.1

NUMBER OF TARGET DOCUMENTS RETRIEVED BY THREE INDEXING METHODS

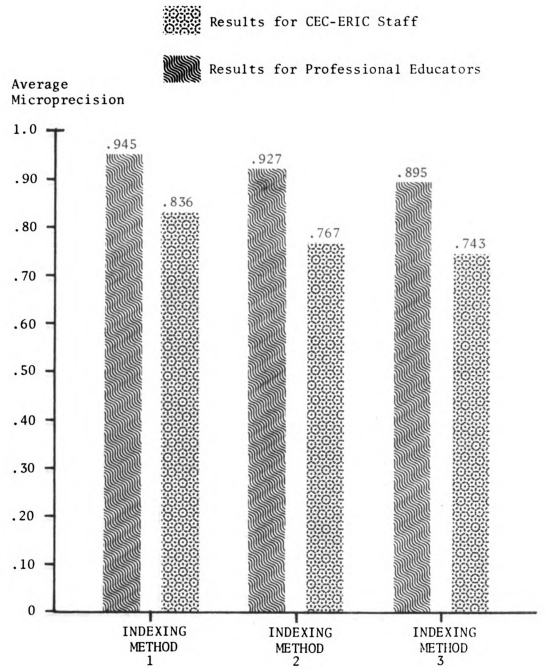


FIGURE 5.2

AVERAGE MICROPRECISION FOR THREE INDEXING METHODS

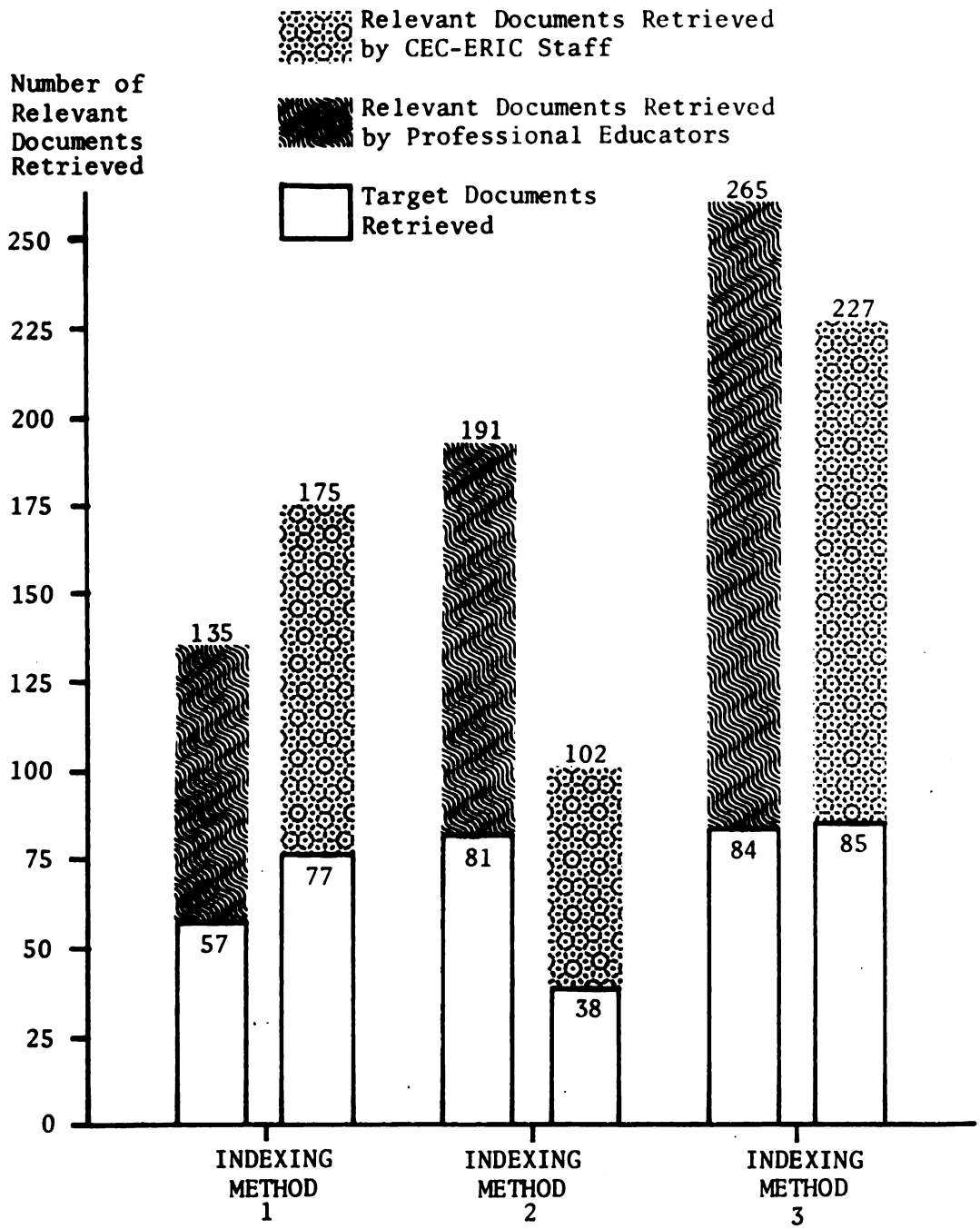


FIGURE 5.3

NUMBER OF RELEVANT DOCUMENTS RETRIEVED BY EACH INDEXING METHOD

Null Hypothesis 1 For indexing method 1 there is no significant difference at the .01 level between the observed and expected number of target documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

Null Hypothesis 2 For indexing method 2 there is no significant difference at the .01 level between the observed and expected number of target documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

Null Hypothesis 3 For indexing method 3 there is no significant difference at the .01 level between the observed and expected number of target documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

In Tables 5.2, 5.3, and 5.4 the upper portion of the cells having multiple data contains the observed values while the lower portion contains the expected values calculated by using the marginal totals. The value of chi-square necessary to reject the null hypothesis at the .01 level is 6.64. The values obtained for chi-square were such that the first two null hypotheses were rejected, while the third failed to be rejected.

A second set of corollary questions implied by the evaluation considered whether there was a statistically significant difference in the Average Microprecision resulting from search questions written by CEC-ERIC staff members as compared with search questions written by special educators. The three null hypotheses implied by this set of corollary questions are:

TABLE 5.2

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 1

Statement of Null Hypothesis 1 For indexing method 1 there is no significant difference at the .01 level between the observed and expected number of target documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

OBSERVED AND EXPECTED VALUES

	Number of Target Documents Retrieved	Number of Target Documents Not Retrieved	Totals
CEC-ERIC Staff	77 / 67	28 / 38	105
Professional Educators	57 / 67	48 / 38	105
Totals	134	76	210

Cell Contributions to Chi-Square

1.49	2.635
1.49	2.635

$X^2 = 8.25$

Since the value of chi-square is greater than 6.64 Null Hypothesis 1 is rejected at the .01 level.

TABLE 5.3

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 2

Statement of Null Hypothesis 2 For indexing method 2 there is no significant difference at the .01 level between the observed and expected number of target documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

OBSERVED AND EXPECTED VALUES

	Number of Target Documents Retrieved	Number of Target Documents Not Retrieved	Totals
CEC-ERIC Staff	38 / 59.5	67 / 45.5	105
Professional Educators	81 / 59.5	24 / 45.5	105
Totals	119	91	210

Cell Contributions to Chi-Square

7.8	10.2
7.8	10.2

$\chi^2 = 36.0$

Since the value of chi-square is greater than 6.64 Null Hypothesis 2 is rejected at the .01 level.

TABLE 5.4

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 3

Statement of Null Hypothesis 3 For indexing method 3 there is no significant difference at the .01 level between the observed and expected number of target documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

OBSERVED AND EXPECTED VALUES

	Number of Target Documents Retrieved	Number of Target Documents Not Retrieved	Totals
CEC-ERIC Staff	85 / 84.5	20 / 20.5	105
Professional Educators	84 / 84.5	21 / 20.5	105
Totals	169	41	210

Cell Contributions to Chi-Square

.003	.00125
.003	.00125

$\chi^2 = .0085$

Since the value of chi-square is less than 6.64 Null Hypothesis 3 is accepted.

Null Hypothesis 4 For indexing method 1 there is no significant difference at the .01 level between the observed and expected number of relevant documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

Null Hypothesis 5 For indexing method 2 there is no significant difference at the .01 level between the observed and expected number of relevant documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

Null Hypothesis 6 For indexing method 3 there is no significant difference at the .01 level between the observed and expected number of relevant documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

The data used in testing null hypotheses 4 through 6 and the resulting values of Chi-square may be found in Table 5.5, 5.6, and 5.7. As may be noted all the values for chi-square are greater than 6.24. Thus, the null hypotheses 4 through 6 were rejected at the .01 level of significance.

Factors Important to the Analysis of Data Resulting from the Comparison of Indexing Methods

In analyzing the results there were three factors with their interactions which were considered. These factors were the differences between the two groups writing questions, the conditions under which the evaluation took place, and the vocabulary used in the questions written by each group. While all the differences in the comparisons would have to be attributed to the types of search questions which

TABLE 5.5

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 4

Statement of Null Hypothesis 4 For indexing method 1 there is no significant difference at the .01 level between the observed and expected number of relevant documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

OBSERVED AND EXPECTED VALUES

	Number of Target Documents Retrieved	Number of Target Documents Not Retrieved	Totals
CEC-ERIC Staff	175 / 184	34 / 25	209
Professional Educators	135 / 126	8 / 17	143
Totals	310	42	352

Cell Contributions to Chi-Square

.44	3.24
.64	4.78

$\chi^2 = 9.10$

Since the value of chi-square is greater than 6.64 Null Hypothesis 4 is rejected at the .01 level.

TABLE 5.6

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 5

Statement of Null Hypothesis 5 For indexing method 2 there is no significant difference at the .01 level between the observed and expected number of relevant documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

OBSERVED AND EXPECTED VALUES

	Number of Target Documents Retrieved	Number of Target Documents Not Retrieved	Totals
CEC-ERIC Staff	102 / 116	31 / 17	133
Professional Educators	191 / 177	15 / 29	206
Totals	293	46	339

Cell Contributions to Chi-Square

1.69	11.54
1.11	6.76

$\chi^2 = 21.10$

Since the value of chi-square is greater than 6.64 Null Hypothesis 5 is rejected at the .01 level.

TABLE 5.7

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 6

Statement of Null Hypothesis 6 For indexing method 3 there is no significant difference at the .01 level between the observed and expected number of relevant documents retrieved from questions written by CEC-ERIC staff members versus questions written by professional educators.

OBSERVED AND EXPECTED VALUES

	Number of Target Documents Retrieved	Number of Target Documents Not Retrieved	Totals
CEC-ERIC Staff	227 / 250	78 / 55	305
Professional Educators	265 / 242	31 / 54	296
Totals	492	109	601

Cell Contributions to Chi-Square

2.15	9.62
2.18	9.62

$X^2 = 23.75$

Since the value of chi-square is greater than 6.64 Null Hypothesis 6 is rejected at the .01 level.

were written, the specific questions which must be answered in attempting to interpret the results are these:

1. Are there any observable differences in the search questions written by the two groups?
2. Can these conditions be attributed to differences within the groups or the conditions under which evaluation took place?

Groups Writing Search Questions The two groups were selected and trained so that the major difference between groups would be that the CEC-ERIC staff was familiar with the ERIC Thesaurus and the indexing procedures used at the Information Center, while the professional educators were not familiar with the Thesaurus or the procedures.

Conditions under Which the Evaluation Took Place When the evaluation took place the following conditions existed:

1. Neither group was aware of the computer indexing method that would be used to make the documents computer searchable.
2. Both groups were given instructions to use each target document (a title and abstract) to write one question that would retrieve that document or similar documents.
3. Both the groups were aware that methods could be used which might generate indexing terms that would not be part of the title or abstract used as the basis for writing the search question.
4. Both groups understood that the evaluation was based upon the retrieval of documents which were similar to the descriptions

(titles and abstracts). They had no assurance that a specific target document would or would not be in each file searched.

Analysis of the Question Vocabulary

The reason for examining the question vocabulary of the two groups was to determine if there was evidence to support the position that a major factor affecting the results was the CEC-ERIC staff's knowledge of the ERIC Thesaurus and the procedures used to index ECEA. Two questions were considered in this analysis. The first was "What proportion of the terms used in the questions written by the CEC-ERIC staff and professional educators were phrases (more than one word)?" The second was "What proportion of the terms used in questions written by CEC-ERIC staff and professional educators were found in the ERIC Thesaurus?" A term was defined to be contained in the ERIC Thesaurus if it was identical to a term in the Thesaurus or it was contained in a term in the Thesaurus. For example, if an individual used MENTALLY HANDICAPPED and the term EDUCABLE MENTALLY HANDICAPPED was in the ERIC Thesaurus, the term MENTALLY HANDICAPPED would also be considered as contained in the Thesaurus. The two null hypotheses implied by these questions are:

Null Hypothesis 7 There is no significant difference at the .01 level between the observed and expected number of phrases used in questions written by CEC-ERIC staff versus professional educators.

Null Hypothesis 8 There is no significant difference at the .01 level between the observed and expected number of terms contained in the ERIC Thesaurus which were used in questions written by CEC-ERIC staff versus professional educators.

As indicated in Tables 5.8 and 5.9 both hypotheses were rejected at the .01 level. An examination of the questions and the data used in the calculation of the two null hypotheses indicates that:

1. Both groups used words and phrases as question terms in the search questions which they wrote.
2. The professional educators used a significantly higher proportion of single-word terms than did the CEC-ERIC staff.
3. The CEC-ERIC staff used a significantly higher ratio of terms contained in the ERIC Thesaurus. About 95% of the terms used by the CEC-ERIC staff were contained in the Thesaurus versus only 65% of those used by professional educators.

Analysis of the Indexing Vocabulary Used in Volume I of ECEA

The indexing terms assigned to describe abstracts contained in Volume I of ECEA were selected from the ERIC Thesaurus. Because the ERIC Thesaurus embraces many topical areas of education[222] there exists a question as to how well its vocabulary reflects the literature in any of the specific areas. The objective of this section is to examine the question, "To what extent do the terms selected from the ERIC Thesaurus to index Volume I of ECEA reflect the vocabulary used in the literature of special education?"

Two sources of vocabulary were used as a basis for examining the indexing terms assigned from the ERIC Thesaurus to describe the

---

[222]James L. Eller and Robert L. Panek, "Thesaurus Development for A Decentralized Information Network," American Documentation, July, 1968, pp. 213-220.

TABLE 5.8

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 7

Statement of Null Hypothesis 7 There is no significant difference at the .01 level between the observed and expected number of phrases used in questions written by CEC-ERIC staff versus professional educators.

OBSERVED AND EXPECTED VALUES

	Words--Terms of Word Length 1	Phrases--Terms of Word Length 2 or Greater	Totals
Questions by CEC-ERIC Staff	184 / 205	273 / 252	457
Questions by Professional Educators	203 / 182	203 / 224	406
Totals	387	476	863

Cell Contributions to Chi-Square

2.15	1.75
2.42	1.97

$\chi^2 = 8.29$

Since the value of chi-square is greater than 6.64 Null Hypothesis 7 is rejected at the .01 level.

TABLE 5.9

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 8

Statement of Null Hypothesis 8 There is no significant difference at the .01 level between the observed and expected number of terms contained in the ERIC Thesaurus which were used in questions written by CEC-ERIC staff versus professional educators.

OBSERVED AND EXPECTED VALUES

	Question Terms Found in ERIC Thesaurus	Question Terms Not Found in ERIC Thesaurus	Totals
CEC-ERIC Staff	435 / 371	22 / 86	457
Professional Educators	266 / 330	140 / 76	406
Totals	701	162	863

11.0	53.5
12.4	62.0

$\chi^2 = 138.9$

Since the value of chi-square is greater than 6.64 Null Hypothesis 8 is rejected at the .01 level.

abstracts in Volume I of ECEA. These sources were the vocabulary found in the collective titles of the abstracts in Volume I of ECEA and the terms contained in a thesaurus prepared by Samuel Price. This Thesaurus defined the words and language of special education by conducting a five-year retrospective search of the professional literature in that field.[223]

Notation

Part of the following notation has been defined in previous chapters, but is repeated here for the convenience of the reader.

Content Word = A word which is not found in an exclusion list containing articles, prepositions, conjunctions, and other words which have been subjectively determined not to have content.

A = The set of all content words found in one or more titles of the abstracts in Volume I of ECEA.

B = The set of all content words found in one or more terms of the Samuel Price Thesaurus.

C = The set of all content words found in one or more of the ERIC descriptors used to index abstracts in Volume I of ECEA.

$A \cup B$ = The union of sets A and B.

$A \cap B$ = The intersection of sets A and B.

$n(S)$ = Number of elements contained in the set S.

---

[223]Samuel T. Price, "The Development of a Thesaurus of Descriptors for an Information Retrieval System in Special Education," (unpublished doctoral dissertation, University of Pittsburgh, 1969), abstract.

RS = The set of all roots of the words which belong to the set S.

ϵ = "is an element of," or "is a member of."

: = "such that".

$\{\}$ = Symbols used to set apart (embrace) a set.

In addition to the above notation, AB will stand for a special set such that for a word to belong to this set it first must be found in either A or B and second must have a root which is found in those roots common to both A and B. Where w stands for any word the set AB may be more precisely defined as follows:

$$AB = \{w : w \epsilon (A \cup B) \text{ and } w \epsilon (RA \cap RB)\}$$

Results of Vocabulary Comparisons of Three Word Lists

Because the Thesaurus developed by Samuel Price was based on a retroactive search of five years' literature in special education,[224] it was used as a criterion vocabulary. The comparisons examined the proportion of words in two lists which had roots that were the same as words contained in Samuel Price's Thesaurus.

The specific problem was to find a basis for evaluating the results of comparisons of the words in the ERIC descriptors used to index Volume I of ECEA with words in the Samuel Price Thesaurus. To accomplish this a second list of words extracted from the literature of special education was used to estimate what proportion of words from two lists extracted from the literature might be expected to have similar roots.

---

[224] Ibid.

A set of words available and appropriate for this comparison were the words found in the titles in Volume I of ECEA.

In making the comparison and analyzing the results a chi-square test was used to examine the following null hypothesis:

Null Hypothesis 9 There is no significant difference at the .01 level between the observed and expected number of words in set A with roots in set RB versus the number of words in set C with roots in set RB.

The analysis of null hypothesis 9 found in Table 5.10 resulted in its rejection thus indicating that there was a significant difference in the number of words in set A having roots in the Samuel Price Thesaurus versus words in set B having roots in the Samuel Price Thesaurus. An examination of the observed and expected values used in calculating the total for chi-square indicated that there were more words from list C (words found in the terms selected from the ERIC Thesaurus to be used in indexing Volume I of ECEA) having roots in the Samuel Price Thesaurus than were expected. Other analyses and comparisons of the vocabulary of word lists A, B, and C provided the following results:

Result 1 $n(A) = 2411$

Result 2 $n(RA) = 1761$

Result 3 $\frac{n(RA)}{n(A)} = .73$

Result 4 $n(B) = 1660$

Result 5 $n(RB) = 1426$

Result 6 $\frac{n(RB)}{n(B)} = .87$

Result 7 $n(C) = 1750$

TABLE 5.10

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 9

Statement of Null Hypothesis 9 There is no significant difference at the .01 level between the observed and expected number of words in set A with roots in set RB versus the number of words in set C with roots in set RB.

OBSERVED AND EXPECTED VALUES

	Words With Roots in Set RB	Words With Roots Not In Set RB	Totals
Words in Set A	1111 1158	1300 1253	2411
Words in Set C	888 841	862 909	1750
Totals	1999	2162	4161

Cell Contributions to Chi-Square

1.90	1.76
2.62	2.43

$\chi^2 = 8.71$

Since the value of chi-square is greater than 6.64 Null Hypothesis 9 is rejected at the .01 level.

Result 8 $\frac{n(RC)}{n(C)} = 1404$

Result 9 $\frac{n(RC)}{n(C)} = .80$

Result 10 $\frac{n(A \cap B)}{n(A \cup B)} = \frac{715}{3365} = .212$

Result 11 $\frac{n(RA \cap RB)}{n(RA \cup RB)} = \frac{670}{2517} = .262$

Result 12 $\frac{n(AB)}{n(A \cup B)} = \frac{1247}{3365} = .372$

Result 13 $\frac{n(A \cap C)}{n(A \cup C)} = \frac{976}{3185} = .326$

Result 14 $\frac{n(RA \cap RC)}{n(RA \cup RC)} = \frac{869}{2296} = .379$

Result 15 $\frac{n(AC)}{n(A \cup C)} = \frac{1613}{3185} = .506$

Result 16 $\frac{n(B \cap C)}{n(A \cup C)} = \frac{646}{2764} = .234$

Result 17 $\frac{n(RB \cap RC)}{n(RB \cup RC)} = \frac{623}{2207} = .282$

Result 18 $\frac{n(BC)}{n(B \cup C)} = \frac{1009}{2764} = .365$

A Subjective Analysis of Terms Selected From the ERIC Thesaurus to Index Volume I of ECEA

Individuals doing computer and hand searches on the information contained in Volume I of ECEA became aware of a lack of precision resulting from the ambiguous assignment of similar indexing terms. Some of the indexers suggested that one reason for this ambiguity was that often the ERIC Thesaurus contained several terms which could be used for indexing the same concept. As a result of these observations, a decision was made to have the indexers subjectively examine the indexing terms assigned to Volume I of ECEA. The result of this examin-

ation was a subset of terms used in Volume I which is now serving as a basis for indexing successive volumes. The major objective of this section is to examine the vocabulary of the subset of terms which resulted from the indexing staff's subjective evaluation of the terms used in indexing Volume I of ECEA.

In the subjective analysis of the ERIC descriptors used in Volume I of ECEA, indexers compared each term with other terms which they felt had similar meanings. When the distinction between the concepts suggested by similar terms was unclear, the indexers examined abstracts to which the terms had been assigned. If it was still not apparent why one term should have preference in indexing a particular concept, the term retained was the one a professional staff member judged to be most common to the literature of special education. Occasionally when no decision could be made on this basis, it was decided to retain the term which had been used most frequently in indexing the documents of Volume I.

Results of Subjective Evaluation of the ERIC Descriptors Used in Volume I of ECEA

This subjective evaluation resulted in removing 1126 descriptors and leaving 1318 descriptors of the original list of 2444 descriptors. The original list of descriptors had 1750 words containing 1440 roots. The remaining list had 1061 unique words with 862 roots.

An analysis of the terms assigned by the indexers to index the abstracts in Volume I of ECEA is found in Table 5.11. This analysis shows the number of terms used once, twice, three times, etc. which were (a) left after the indexers examined the terms, (b) removed when the indexers examined the terms, and (c) the number of identifiers

TABLE 5.11

RESULTS OF INDEXERS' SUBJECTIVE ANALYSIS
OF TERMS USED TO INDEX VOLUME I OF ECEA

*Ratio = Number of ERIC descriptors retained divided by the number of ERIC descriptors assigned.

Number of Times a Term Was Assigned to Volume 1	Number of Identifiers	Number of ERIC Descriptors Retained	Number of ERIC Descriptors Removed	Total	*Ratio
1	403	218	463	1084	.32
2	56	139	184	379	.43
3	24	125	137	286	.48
4	9	76	75	160	.504
5	8	69	47	124	.59
6	5	51	36	92	.53
7	3	49	28	80	.64
8	4	41	19	64	.68
9	1	34	22	57	.60
10	1	38	12	51	.76
11	2	29	10	41	.74
12	3	25	9	37	.74
13	1	25	7	33	.78
14		21	8	29	.72
15		26	9	35	.81
16		20	7	27	.74
17		14	3	17	.82
18		16	2	18	.89
19		14	5	19	.74
20		13	2	15	.87
21	3	11	6	20	.64
22		7	3	10	.70
23		8	2	10	.80
24		11	2	13	.85
25		2		2	1.0
26		13	1	13	.93
27		4	1	5	.80
28		11	3	14	.79
29		12	1	13	.92
30		6	2	8	.75
31		4	1	5	.80
32		11		11	1.0
33		4		4	1.0
34		10	1	11	.91
35		8	2	10	.80
36		5	1	6	.83
37		3	2	5	.60
38		4	1	5	.80
39		3		3	1.0
40		3	1	4	.75

TABLE 5.11 (cont'd)

*Ratio - Number of ERIC descriptors retained divided by the number of ERIC descriptors assigned.

Number of Times a Term Was Assigned to Volume 1	Number of Identifiers	Number of ERIC Descriptors Retained	Number of ERIC Descriptors Removed	Total	*Ratio
41		5		5	1.0
42		6	2	8	.75
43		3		3	1.0
44		3		3	1.0
45		2		2	1.0
46		4		4	1.0
47		5	1	6	.83
48		3		3	1.0
49		4		4	1.0
50		2		2	1.0
51		4	1	5	.80
52		1	1	2	.50
53		2		2	1.0
54		2		2	1.0
55		3		3	1.0
56		0			--
57		2		2	1.0
58		2	1	3	.67
59		0			--
60		2		2	1.0
61-70		9	2	11	.82
71-80		17	1	19	.90
81-90		7	1	8	.88
91-100		7		7	1.0
101-110		7		7	1.0
111-120		5		5	1.0
121-130		6		6	1.0
131-140		2		2	1.0
141-150		2		2	1.0
151-160		2		2	1.0
161-170		1		1	1.0
171-180		5		5	1.0
181-190		1		1	1.0
191-200		0			--
201-300		6		6	1.0
301-400		2		2	1.0
401-700		0			--
701-710		1		1	1.0
Totals	523	1318	1126	2967	.534

assigned. Identifiers were special indexing terms which were not in the ERIC Thesaurus but assigned because of special meaning to a particular document. For an example, an identifier might be the name of a test or institution.

The results of this analysis indicated that if a term had been used one, two, or three times, the probability of it being removed in the subjective evaluation by the indexers was greater than the chance of its remaining. All terms which had been used four or more times had a greater probability of remaining than of being removed.

The results shown in Table 5.11 have some similarity to Luhn's suggestion that terms used very frequently and terms used very infrequently have little indexing value. In this case the indexers subjective judgment concerning the value of indexing terms supports the part of Luhn's suggestion that terms used infrequently have little indexing value.[225]

The Effect of Indexing Procedure Changes in Volume II of ECEA

Two major changes were made during the indexing of Volume II of ECEA. In Volume I the indexers had been told to assign a descriptor to describe an abstract even if an abstract contained very little information related to the descriptor. This resulted in individuals retrieving abstracts which had very little information related to their question. As a result of this observation, the decision was made that in future volumes a descriptor would not be assigned to an abstract

---

[225]H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, II (1958), 159-165.

unless the document contained considerable information related to the concept described by the indexing term.

A second change initiated part way through Volume II was the use of an authority list[226] developed as a result of the indexers' subjective evaluation of the ERIC descriptors used in Volume I. The reason the list was not used sooner was because the subjective evaluation took place during the time that the first two issues of Volume II were being indexed.

Results of Indexing Procedure Changes

In an initial attempt to evaluate the effect of the change in indexing procedures a set of 20 questions was used in searching both Volume I and II with separate values of Average Microprecision and Average Macroprecision being calculated for each volume.

The questions were written by one of the staff members to retrieve documents for use in a bibliography series. For each question the staff member determined which of the documents retrieved had considerable information about the question and which did not. The documents retained then served as a basis for developing special bibliographies about that topical area. At the time that the evaluation was made the staff member was unaware that the results would be used in the evaluation.

Retrieval results for Volume I resulted in Average Microprecision of .54 and Average Macroprecision of .60. For Volume II the Average Microprecision was .60 and the Average Macroprecision .68. A chi-square

---

[226]Thesaurus for Exception Child Education, (Arlington, Virginia: CEC Information Center on Exceptional Children, 1970), pp. 1-10.

test of statistical significance was used to examine the following null hypothesis:

Null Hypothesis 10 There is no significant difference between the observed and expected number of relevant documents retrieved by 20 search questions from Volume I versus the number retrieved by the same questions from Volume II.

The data and calculations used in examining this hypothesis are found in Table 5.12. As may be observed, the hypothesis was rejected at the .01 level. An examination of the data reveals that more relevant documents than expected were retrieved from Volume II.

The data resulting from each of the 20 questions is found in Table 5.13. As revealed by the data in this table the precision was greater for documents retrieved from Volume II than from Volume I for every question except Number 7 and in this question the precision was almost equal--.97 for Volume I and .966 for Volume II. In other words in 19 out of 20 questions there was a greater precision for documents retrieved from Volume II than from Volume I. The probability of this happening by accident is less than one in fifty thousand.

Summary

An examination of the data indicated the following results:

1. CEC-ERIC staff had significantly better retrieval of target documents than professional educators when the indexing method used terms from titles and ERIC descriptors.
2. The professional educators had significantly better retrieval of target documents when the indexing method used terms from titles and abstracts.

TABLE 5.12

DATA AND CALCULATIONS USED IN TESTING NULL HYPOTHESIS 10

Statement of Null Hypothesis 10 There is no significant difference between the observed and expected number of relevant documents retrieved by twenty search questions from Volume I versus the number retrieved by the same questions from Volume II.

OBSERVED AND EXPECTED VALUES

	Relevant	Not Relevant	Total
Volume I	1449 / 1546	1226 / 1129	2675
Volume II	2243 / 2146	1469 / 1566	3712
Totals	3692	2695	6387

Cell Contributions to Chi-Square

6.08	8.34
6.00	4.38

$X^2 = 24.80$

Since the value of chi-square is greater than 6.64 Null Hypothesis 10 is rejected at the .01 level.

TABLE 5.13

SEARCH RESULTS OF TWENTY QUESTIONS
USED ON VOLUME I AND VOLUME II
OF ECEA

Question No.	RESULTS FOR VOLUME I			RESULTS FOR VOLUME II		
	Documents Retrieved	Documents Relevant	Precision	Documents Retrieved	Documents Relevant	Precision
1	10	9	.9	18	17	.94
2	171	122	.71	301	240	.798
3	13	12	.92	13	13	1.
4	14	20	.83	29	27	.93
5	189	92	.487	300	179	.596
6	256	112	.438	287	144	.502
7	108	105	.97	115	111	.966
8	144	48	.333	264	108	.41
9	154	91	.592	155	97	.623
10	126	61	.484	145	81	.56
11	186	40	.216	219	59	.27
12	232	90	.388	388	174	.516
13	37	9	.244	55	34	.63
14	88	55	.625	115	87	.756
15	191	111	.582	278	199	.718
16	129	84	.65	169	122	.722
17	247	133	.538	360	140	.388
18	114	100	.876	219	194	.885
19	154	76	.494	191	104	.545
20	102	79	.774	141	113	.801
Totals	2675	1449	12.051	3712	2243	13.556

3. There was no significant difference in the retrieval of target documents when the indexing method used terms from titles, abstracts and ERIC descriptors.
4. Professional educators had significantly better Average Micro-precision for all three indexing methods than did the CEC-ERIC staff.
5. Professional educators used significantly more single word terms in questions than did CEC-ERIC staff.
6. CEC-ERIC staff used significantly more terms found in the ERIC Thesaurus than did the professional educators.
7. The subjective analysis by the CEC-ERIC indexers of the ERIC descriptors used in indexing Volume I of ECEA resulted in
   a. The reduction of 2440 ERIC descriptors having 1750 words and 1440 roots to 1313 ERIC descriptors having 1061 words and 862 roots.
   b. More terms which were used 1, 2, and 3 times in Volume I were removed from the list than retained.
   c. If a term had been used to index 4 or more documents in Volume I it was more likely that it be retained than removed.
8. The changes in indexing procedures between Volume I and Volume II resulted in significantly greater precision in Volume II.

The following chapter--Summary, Conclusions, Recommendations, and Implications--analyzes the procedures and results of this study, states possible conclusions implied by the results, makes specific recommendations concerning possible changes in the Information Center and examines the implications of this study that may pertain to future studies.

CHAPTER VI: SUMMARY, CONCLUSIONS, RECOMMENDATIONS, AND IMPLICATIONS

As indicated by the title, this chapter is divided into four major sections. The first summarizes the procedures and results of this study, the second discusses the conclusions implied by the results of testing ten null hypotheses, the third makes specific recommendations concerning possible changes in operations at the Information Center, and the fourth examines the implications of this study for future studies and for improving communication in the field of education.

Summary

The increased publishing in the area of education has made it important to find better ways to store, organize, and disseminate this information. In the area of special education the Council for Exceptional Children has cooperated with the ERIC system in developing the CEC-ERIC Information Center.

Procedures Used at the CEC-ERIC Center

The initiation of computer processing at the Center was made possible by the use of the Basic Indexing and Retrieval System (BIRS) and resulted in an automated abstract journal, Exceptional Child Education Abstracts. Abstracts stored on computerized information files serve as the basis for computer-controlled typesetting for ECEA, computer-generated indexes for ECEA, computerized searching to aid in answering user

requests, and the development of special annotated bibliographies. The bibliographies are also printed using computer controlled typsetting.

A Comparison of Three Indexing Methods

The effectiveness of three indexing methods was compared for two groups of individuals. The first group was ten CEC-ERIC staff members familiar with the Center's indexing procedures while the second group was seven professional educators not familiar with the Center's indexing procedures. The three indexing methods used in the comparisons were the following:

Indexing Method 1 (the method normally used at the Center) used ERIC descriptors assigned by indexers and computer-extracted terms from the titles of the documents.

Indexing Method 2 used computer-extracted terms from titles and abstracts of documents.

Indexing Method 3 combined Methods 1 and 2 using computer-extracted terms from titles and abstracts as well as ERIC descriptors.

One hundred five target documents were randomly selected from the file of 2100 abstracts used for Volume I of ECEA and a set of 105 questions (one question to retrieve each target document) was written by each group. Training and technical assistance was provided so that each group demonstrated similar technical competence in writing computer search questions.

The results indicated that the CEC-ERIC staff retrieved significantly more target documents than professional educators when Indexing Method 1 was used. This result was reversed for Method 2, while in Method 3 there was no significant difference in the number of target documents retrieved by either group. The professional educators retrieved

191 relevant documents including 81 target documents by Method 2 while the CEC-ERIC staff retrieved only 175 relevant documents including 77 target documents by Method 1 (the method used at the center). The professional educators had significantly better Average Microprecision for all three methods than the CEC-ERIC staff and retrieved more relevant documents (target and non-target documents) by Method 2 and 3 than did the CEC-ERIC staff. This data tends to suggest that the need for carefully controlled indexing languages is minimized in the field of education when sophisticated computer searching algorithms are available.

Comparison of Vocabulary of Three Word Lists

The words in a thesaurus developed by Samuel Price through a retrospective analysis of five years' special education literature[227] were used as a criterion in comparison with: (1) words extracted from the titles of the 2100 abstracts in Volume I of ECEA and (2) words in the ERIC descriptors used to index these abstracts. The results indicated that the words in the ERIC descriptors had as much or more in common with the vocabulary used by Samuel Price as did the words in the collective titles.

Changes in Indexing Procedures

A subjective analysis by indexers of the ERIC descriptors assigned to Volume I of ECEA resulted in a reduced list of descriptors for use in indexing subsequent volumes of ECEA. It also resulted in a change

---

[227]Samuel Price, Thesaurus of Descriptors for an Information Retrieval System in the Subject Matter Area of Special Education, (Normal Illinois: Illinois State University, Special Education Instructional Materials Laboratory, 1970), pp. 1-465.

of indexing procedures which required an abstract to have more information about a subject before an ERIC descriptor would be assigned. A preliminary evaluation indicated that the change in indexing procedures between Volume I and Volume II resulted in increased precision for computerized search questions.

Conclusions

In evaluating the indexing methods used at the CEC-ERIC Information Center and analyzing the vocabulary of the ERIC descriptors assigned to abstracts of Volume I of ECEA ten null hypotheses were examined. In each case a chi-square test of statistical significance with one degree of freedom was used to determine if the null hypothesis should be accepted or rejected at the .01 level. Under these conditions any value of chi-square greater than 6.64 would cause the null hypothesis to be rejected.

The first eight null hypotheses were used to examine retrieval effectiveness of three indexing methods for CEC-ERIC staff versus professional educators. The ninth null hypothesis was used to examine if words in the ERIC descriptors used to index Volume I of ECEA were typical of those in the literature of special education. The tenth was used in to examine if changes in indexing procedures between Volume I and Volume II had affected the Average Microprecision.

Results of Testing Null Hypothesis 1

The value of chi-square resulting from the data used to test Null Hypothesis 1 was 8.25, thus the hypothesis was rejected at the .01 level. The rejection of the hypothesis and the examination of the data resulted in the conclusion that the CEC-ERIC Staff retrieved significantly more

target documents when Indexing Method 1 was used than did the professional educators.

Results of Testing Null Hypothesis 2

The value of chi-square resulting from the data used to test Null Hypothesis 2 was 36.0, thus the hypothesis was rejected at the .01 level. The rejection of the null hypothesis and the examination of the data resulted in the conclusion that the professional educators retrieved significantly more target documents when Indexing Method 2 was used than did the CEC-ERIC Staff.

Results of Testing Null Hypothesis 3

The value of chi-square resulting from the data used to test Null Hypothesis 3 was .0085, thus the hypothesis was accepted. The conclusion resulting from this acceptance was that there was no significant difference in the number of target documents retrieved by professional educators versus CEC-ERIC staff when Indexing Method 3 was used.

Results of Testing Null Hypotheses 4, 5, and 6

The values of chi-square for Null Hypotheses 4, 5, and 6 respectively were 9.10, 21.10 and 23.75, thus these hypotheses were all rejected at the .01 level. The rejection of these hypotheses and the examination of the data resulted in the conclusion that for all three indexing methods the professional educators had significantly greater Average Microprecision than did the CEC-ERIC staff.

Results of Testing Null Hypothesis 7

The value of chi-square resulting from the data used to test Null

Hypothesis 7 was 8.29, thus the hypothesis was rejected at the .01 level. The rejection of the hypothesis and the examination of the data resulted in the conclusion that the professional educators used a significantly greater proportion of single-word terms in their questions than did the CEC-ERIC staff. Another way of stating this conclusion would be that the CEC-ERIC Staff used a significantly greater proportion of phrases (terms of two or more words) in their questions than did the professional educators.

A possible explanation for the results and conclusions of Null Hypothesis 7 was that this difference in the question vocabulary related to the CEC-ERIC staff's knowledge and use of the ERIC Thesaurus where most terms are multiple-word descriptors.

Results of Testing Null Hypothesis 8

The value of chi-square resulting from the data used to test Null Hypothesis 8 was 138.9, thus the hypothesis was rejected at the .01 level. The rejection of the hypothesis and the examination of the data resulted in the conclusion that the CEC-ERIC staff used more terms found in the ERIC Thesaurus than did the professional educators. The results and conclusions related to this null hypothesis were interpreted to indicate that a major difference between the two groups was their knowledge of the ERIC Thesaurus.

Results of Testing Null Hypothesis 9

The value of chi-square resulting from the data used to test Null Hypothesis 9 was 8.71, thus the hypothesis was rejected at the .01 level. The rejection of the hypothesis and the examination of the data resulted

in the conclusion that the vocabulary (words) found in the ERIC descriptors used to index Volume I of ECEA had significantly greater similarity to the vocabulary of the Samuel Price Thesaurus[228] than did the vocabulary found in the titles of the abstracts contained in Volume I of ECEA. This was interpreted to mean that the vocabulary of the ERIC descriptors used in indexing Volume I of ECEA was representative of the field of special education.

Results of Testing Null Hypothesis 10

The value of chi-square resulting from the data used to test Null Hypothesis 10 was 24.80, thus the hypothesis was rejected at the .01 level. The rejection of the hypothesis and the examination of the data resulted in the conclusion that in this comparison the Average Micropreprecision for Volume II was significantly greater than the Average Microprecision for Volume I. This was interpreted to mean that the change in indexing procedures between Volume I and Volume II had resulted in an increase in Average Microprecision.

In addition to the data used in testing Null Hypothesis 10 the precision of each question for Volume I and Volume II was compared. In 19 out of 20 questions the precision was higher in Volume II than in Volume I. In the single instance where this was reversed the precision was extremely high for both volumes (.97 for Volume I and .966 for Volume II). The chance that the precision should be higher in one volume than the other in 19 out of 20 cases is less than one in fifty thousand.

---

[228] Ibid.

Interpretation of the Results of the Comparison of Three Indexing Methods

The question vocabulary used by the two groups and the apparent advantage of the CEC-ERIC staff in retrieving target documents when Indexing Method 1 was used would tend to support the assertion that the major difference between the two groups was their knowledge of the CEC indexing procedures. Of particular interest to the interpretation of the results is the fact that this apparent advantage for Method 1 gained by a knowledge of the CEC indexing procedures appeared to be a disadvantage when other methods were used.

With Indexing Method 2 which used terms from titles and abstracts the professional educators did significantly better in retrieving target documents. Not only did they do better, but surprisingly they retrieved more target documents using Method 2 (81) than the CEC-ERIC staff retrieved using Method 1 (77)--the method to which they were accustomed. Still more remarkable is the fact that the professional educators had significantly better Average Microprecision for all three indexing methods than did the CEC-ERIC staff.

The comparisons between Methods 1 and 2 raise the question, "What is the advantage in using ERIC descriptors when those trained to use them retrieved fewer relevant documents by Indexing Method 1 (using terms from titles and ERIC descriptors) than did professional educators with computer extracted indexing terms from titles and abstracts?" This question becomes even more germane when it is remembered that professional educators also had greater Average Microprecision for both methods. More specifically it would tend to suggest that the use of controlled indexing vocabularies needs to be reconsidered in light of the computer searching methods now available.

To examine this question in perspective it is necessary to consider the results from Indexing Method 3. In this method there was not a significant difference in the number of target documents retrieved by either group. However, the CEC-ERIC staff retrieved 8 additional target documents as compared to the method to which they were accustomed (Method 1) while the professional educators retrieved only three additional target documents as compared to Method 2. This in itself might not be a sufficient reason to consider using the combination method--the method which used terms from titles, abstracts, and ERIC descriptors, but when the total number of relevant documents retrieved (both target and non-target documents) is considered, the advantage becomes more apparent. The CEC-ERIC staff was able to retrieve 227 relevant documents by this method as compared to 175 by Method 2, an increase of 52 relevant documents. Professional educators were able to retrieve 269 documents by this method as compared to 191 by Method 2, an increase of 72 documents.

As expected from other studies the recall obtained by professional educators showed an inverse relationship to Average Microprecision.[229] In other words as the number of relevant documents retrieved increased from 135 to 191 to 265, the Average Microprecision decreased from .945 to .927 to .895.

Before assuming that the combination method (Method 3) should be adopted by the Information Center it is important to consider that the

---

[229]F. W. Lancaster and J. Mills, "Testing Indexes and Index Language Devices: The ASLIB Cranfield Project," American Documentation, January, 1964, p. 9; and G. Salton, E. M. Keen, and M. Lesk, "Design Experiments in Automatic Information Retrieval," The Growth of Knowledge, Manfred Kochen, editor, (New York: John Wiley & Sons, Inc., 1967), pp. 344-346.

cost of the computer indexing and the computer searching is about fifty per cent more than with Method 1.

There remains the disturbing problem of attempting to explain why a knowledge of the CEC indexing procedures which use the ERIC Thesaurus would result in less Average Microprecision for all methods and the retrieval of fewer relevant documents in every method except the one specifically used in the Information Center. There appear to be no certain answers, nevertheless it is worth noting that those not having the knowledge of the ERIC Thesaurus used significantly more single word terms in their questions and significantly less terms found in the ERIC Thesaurus than the CEC-ERIC staff.

It should be noted that the results might have been significantly different if it had not been for the sophisticated searching algorithms of the BIRS system which tend to assist individuals having knowledge of the content area but little knowledge of computer searching techniques. While difficult if not impossible to prove, it would appear that the capability of these algorithms to compare a word or portion of a phrase to a total phrase and to reduce words to their root form may have played an important role. The implications of the term reduction algorithms will be further seen in a later section where the results of the vocabulary analysis of three word lists related to special education are examined.

Reflections on Methodology Used in Comparing Indexing Methods

Swanson raised a question about using target documents as a basis for writing search questions, indicating that there may be some type of relationship between the target document and the question that would not exist in an ordinary information request. The Cranfield project

searches (about which this criticism was first made) were made to get measures of recall for use in evaluating the effectiveness of four different indexing methods.[230] While Cleverdon conceded that there might be unnatural relationships between the questions and the target document, he did not concede that this relationship was sufficient to rule out using this method in all other studies.[231]

In this study the objective was not only to compare different indexing methods, but to analyze the affect of these methods on two different groups of individuals. The only way to effectively do this was to provide each group of individuals with identical descriptions of target documents. In this manner both groups perceived that they were looking for the same information.

The implication of these questions could be that the results from the indexing methods using titles and abstracts were positively affected because people were presented with titles and abstracts.

Response to Questions It seems reasonable that the manner in which information is presented will influence the type of questions written and that the type of questions written will influence the search results for different indexing methods.

To make the comparison required for this study some method of presenting information had to be chosen. The method used for giving information to both groups was chosen because it was felt that it was a natural way of presenting information (information about documents is

---

[230]Don R. Swanson, "The Evidence Underlying the Cranfield Results," The Library Quarterly, January, 1965, p. 1-20.

[231]Cyril Cleverdon, "The Cranfield Hypotheses," The Library Quarterly, April, 1965, pp. 121-124.

commonly communicated through titles and abstracts) and because this is the way information is presented to users of ECEA.

Response to Implications If this procedure did make one method look better, it apparently did so only for those who had not previously used ERIC descriptors. While these questions provide for interesting speculation, it was not the objective of this study to consider them.

A more pragmatic implication from the above two questions would be: If being presented with titles and abstracts did affect the results in a positive manner and since ECEA users are normally presented with titles and abstracts (ECEA is an abstract journal), then why not use an indexing method which extracts terms from the titles and abstracts--the method that is most congruent with the format of ECEA.

Interpretation of the Vocabulary Comparisons

The analysis of Hypothesis 9 resulted in its rejection thus indicating that there was a significant difference in the number of words in set A having roots in the Samuel Price Thesaurus versus words in set B having roots in the Samuel Price Thesaurus. An examination of the observed and expected values used in calculating the total for chi-square indicated that there were more words from list C (words found in the terms selected from the ERIC Thesaurus to be used in indexing Volume I of ECEA) having roots in the Samuel Price Thesaurus than were expected. Because the words found in the titles of the abstracts were directly extracted from the literature of special education, it is assumed that words from a list which have as much or more in common with the Samuel Price Thesaurus as these are representative of the vocabulary used in special education.

An examination of results 1 through 18 found on pages 169 and 171 provides the basis for at least three interesting observations and possible interpretations. Included in these are:

1. The number of different words having the same root was less in the controlled vocabularies than in the free vocabularies. Specifically there were 87 percent as many roots as words in the Samuel Price Thesaurus; 80 percent as many in the ERIC terms used for indexing Volume I of ECEA; and 73 percent as many in the titles of Volume I. A possible interpretation of this may be that when a thesaurus is developed by a single individual there is more consistency in the word forms used than when a thesaurus is developed by a group, and that still less consistency results when words come from many authors.
2. Results 10, 11, and 12 for the intersection and unions of A and B; results 13, 14, and 15 for the intersections A and C; and the results 16, 17, and 18 for the intersections of B and C illustrate the importance of reducing words to roots to assist in information retrieval. As is noted in each set of three comparisons, the ratio of the number of elements in the intersection divided by the number of elements in the unions increase as the reduction of words to roots plays a more important role.
3. The largest ratios involving the comparison of sets A to B, A to C, and B to C resulted when the words of the titles of Volume I were compared to the words found in the ERIC descriptors used to index Volume I of ECEA (Sets A to C). The most likely explanation for this would be the fact that sets A and

B describe the same set of documents, whereas the words in the Thesaurus of Samuel Price were extracted from a different set of documents.

Interpretation of the Effect of Changing Indexing Procedures

The rejection of Null Hypothesis 10 and the data found in Table 5.13 leave little doubt that changes in indexing procedures between Volume I and Volume II resulted in an increase in precision for searches done on Volume II.

When the preliminary examination of the effect of a change in indexing procedures was done there was no practical way to determine or estimate recall. However, data was examined to determine if the expected number of relevant documents retrieved by the 20 questions was proportional to the number of abstracts in Volume I and Volume II. In Volume I, 1449 relevant documents were retrieved from a file of 2100 abstracts. For this proportion to be maintained in Volume II which contained 3615 abstracts, about 2500 relevant documents would need to be retrieved. Approximately 90% or 2143 relevant documents were actually retrieved. If the number of relevant documents compared to the file size were used to estimate changes in recall this data would suggest that whatever the recall in Volume I it would be only 90% of that figure for Volume II. If the assumptions used in these calculations are accepted, it could be concluded that an increase in precision did, in this case, result in an apparent decrease in recall.

A second alternative that should be considered is that the content of Volume I and Volume II is not similar. It should be noted that there were more historical documents in Volume I, thus the acquisition policies were not identical. On the other hand, both the historical

and recent documents obtained for Volume I were acquired under the same philosophy and acquisition criteria as documents acquired for Volume II.

There is no simple way to determine whether the number of relevant documents retrieved in Volume II represents a drop in recall or a difference in file content between Volume I and Volume II. However, based upon results of other research and knowledge of the file content, the author feels that the most probable interpretation is that there was a reduction in recall.

Recommendations

Two specific types of recommendations will be made. The first directly relates to the data and results of this study while the second relates to observations made during the study.

Data Related to Recommendation 1

The results of changing the indexing procedures between Volume I and Volume II improved the precision of search results with some apparent loss of recall. When the size of information files increases this is usually a desirable result. However, in some instances, it may be important to sacrifice precision for recall, such as when a search is being made on a topic about which the file has little information. It is possible by using the facilities of BIRS to have improved precision when desired and in other cases to sacrifice precision for improved recall. Recommendation 1 relates to procedures which could be implemented to meet this objective without major changes in the indexing processes or cost.

Recommendation 1

It is recommended that the CEC-ERIC Information Center give consideration to assigning ERIC descriptors to three separate fields, employing three levels of indexing. In the first field or level at most two ERIC descriptors would indicate what the document is primarily about. In the second field or level ERIC descriptors would indicate topics about which the document has considerable information. In the third field or level ERIC descriptors would indicate topics about which the document contained only marginal amounts of information.

Depending upon the desired retrieval results any or all of the above levels could be searched. If high precision were important, no more than the first two indexing levels would be used. If recall were the most important factor all three levels would be used. It might also be desirable to include in ECEA first, second, and third level indexes.

Data Related to Recommendation 2

The data indicated that additional terms extracted from abstracts improved the recall of both the CEC-ERIC staff and the professional educators. It also indicated that professional educators using terms from titles and abstracts were able to retrieve documents with greater Average Microprecision and recall than the CEC-ERIC staff did by using the method to which they were accustomed. Because the BIRS system makes it possible to search on any combination of fields which have been indexed and included on the description file, the addition of terms from abstracts, would not have to change any of the techniques presently used.

Recommendation 2

Because of the apparent advantages for both the CEC-ERIC staff and professional educators, it is recommended that consideration be given to including terms extracted from the abstracts as part of the description file.

Observations Related to Recommendation 3

Each user obtaining data from the Center is sent one or two simple questionnaires which attempt to gather data about the user and how well he was served. If part of the information given to the user is a bibliography the questionnaire attempts to gather information about how well this bibliography met the user's need. The other questionnaire attempts to evaluate how well other types of materials answered the user's specific questions. About ten percent of these questionnaires are returned with the cumulative results presenting a positive picture. The question which is unanswered is "Would those who have not returned the questionnaires respond in the same way as those who have returned the questionnaires?"

Recommendation 3

Considering the low percentage of the questionnaires returned it is recommended that consideration be given to using one or two possible techniques to gather information about users. The first technique would offer (perhaps for a limited time) those who return the questionnaires a bonus publication with the hope that this would increase returns to 70 percent or more. The second technique would use telephone interviews with a random sample of users to obtain more representative data.

Observations Related to Recommendation 4

Regardless of how many users are served by an information center certain costs remain fixed. Because of this the cost per information request can be reduced by increasing the number of users.

While some data is presently collected concerning the characteristics of those using the Information Center, little if any data has been collected concerning the proportion of potential users who know about the Center or who use the Center.

Recommendation 4

To aid in making decisions that may help improve cost effectiveness, and aid in better serving users it is recommended that consideration be given to conducting a study, a part of which would include:

1. Defining the characteristics of individuals who are considered to be potential users of the Information Center.
2. Estimating the number and location of individuals who have the defined characteristics.
3. Collecting from a random sample of these individuals information including the following:
    a. Whether or not they are aware of the services provided by the Information Center.
    b. What their information needs are in the area of special education.
    c. Whether they have ever used the Information Center.
    d. If they have used the Information Center, how effectively it served their needs.

e. If they have used the Information Center, how they first found out about the Center.

f. If they have not used the Center, where they obtain the type of information provided by the Center.

The information collected would be used to determine if there are appropriate modifications in procedure that might help potential users to become aware of the Center, improve the Center's services to users, or make acquisition policies more congruent with users' information needs.

Observations Related to Recommendation 5

A subjective analysis was made of the vocabulary used in the Samuel Price Thesaurus[232] which was not found in the ERIC descriptors used in indexing Volume I of ECEA. This analysis suggested that a major category of terms not included in the ERIC descriptors were terms of a technical nature--especially those related to medical literature.

Recommendation 5

It is recommended that the Information Center give careful consideration to its acquisition procedures and policies including the type of medical literature that might be of value to those working in special education. If the study suggested in Recommendation 4 is carried out it might aid in the examination of the acquisition procedures.

---

[232] Ibid.

Implications

The results of this study and the procedures developed at the CEC-ERIC Information Center appear to have two potentially important implications related to improving communication within the field of education and in other areas. The first implication relates to the use of controlled indexing vocabularies when powerful computer searching algorithms are available and the second relates to the use of searchable information files in the publication of selected materials.

The Use of Controlled Indexing Vocabularies

The rationale which is sometimes used to support various types of controlled indexing languages is that the restricted vocabulary provides the communication linkage between the indexer and the user. The following results from this study related to this assumption:

1. CEC-ERIC staff retrieved 175 relevant documents (target and non-target) and 77 target documents by Indexing Method 1. This was the indexing method that they were accustomed to and the method which depended most on the controlled vocabulary of the ERIC descriptors.
2. Professional educators retrieved 191 relevant documents (target and non-target documents) including 81 target documents by Indexing Method 2 (the method which extracted terms from titles and abstracts without the use of a controlled vocabulary).
3. Professional educators not familiar with the controlled vocabulary (the ERIC descriptors) had significantly better precision on all three indexing methods than the CEC-ERIC staff and

retrieved more relevant documents on Methods 2 and 3 than did the staff.

In other words, the controlled vocabulary appeared to benefit the CEC-ERIC staff only when they were using Indexing Method 1. When using other methods their knowledge of the ERIC descriptors appeared to be of no advantage and was perhaps a disadvantage. Those not familiar with the ERIC descriptors were able to retrieve more relevant documents with greater precision by Methods 2 and 3 than could the ERIC staff by any method including the method to which they were accustomed.

The analysis of these results cannot help but raise serious doubts concerning the value of using a controlled vocabulary when flexible and powerful searching algorithms are available. Specifically, it raises the question "What is the advantage of using ERIC descriptors (a controlled vocabulary) when those trained to use them retrieved fewer relevant documents by the indexing method designed to utilize the ERIC descriptors, than professional educators retrieved by an indexing method which uses terms extracted from titles and abstracts?"

One answer is that the descriptors are used for generating printed indexes for both ERIC and CEC-ERIC publications. These indexes would be much more cumbersome if a controlled vocabulary was not used. A second answer is that the ERIC descriptors when added to terms extracted from titles and abstracts did increase the total number of relevant documents retrieved by professional educatiors from 191 to 265.

A third answer to this question is that if the searching algorithm available in BIRS had not been used an entirely different result might have been obtained. These algorithms minimize the need for controlled vocabulary by reducing words to root forms and by matching words or

small phrases with larger phrases. The results of this study would tend to support the position that when such algorithms are available the value of a controlled vocabulary is minimized.

These results not only have implications for those establishing new information systems, but also imply the need for further studies. Specifically studies are needed to compare the interaction between controlled indexing vocabularies and the type of searching algorithms used in this study versus algorithms which make only exact matches. It is possible to design such a study that will build on the data and questions used in this study.

An Evolving Thesaurus

A problem common to computerized information retrieval systems which use controlled thesauri is the dependency of those writing questions on the thesauri. Because of this dependency, a human interface has often been used between the actual user and the information system. When this is done, an additional subjective judgment is added which may reduce the chance of the user getting the information desired.

In this study the indexing and searching algorithms of the BIRS system made it possible for individuals not familiar with a thesaurus to get results comparable to or better than those obtained by individuals familiar with a thesaurus. This would tend to support the position that systems can be developed which eliminate the need for a human interface. In analyzing the algorithms to determine how they facilitated this result, the most apparent reason was that it was not necessary for terms to match exactly. For example, the terms MENTALLY RETARDED and MENTAL RETARDATION would be considered a match, similarly MENTALLY RETARDED and EDUCABLE MENTALLY RETARDED.

An analysis of the results of questions written by those familiar with the ERIC Thesaurus established that one failure resulted from individuals using the term MENTALLY RETARDED when the thesaurus contained the term MENTALLY HANDICAPPED. Because handicapped and retarded do not have similar roots, it is necessary for a human decision to be made before the computer can equate such terms.

One reason for the use of thesauri is to show relationships between terms. For example, if a term is not in the thesaurus, it may suggest an alternative term or it may identify for a given term corresponding broad terms, narrow terms, and related terms. If such relations were permanently stored in a computer and used by appropriate indexing and searching algorithms, it might be possible to further improve search results.

One means of giving an empirical base to these relationships would be to analyze user requests. If a user wrote a question which contained a term that was not indexed in any document, this would be noted by the computer and the term stored on a separate file for later analysis. Individuals familiar with the indexing terms and the information files could then examine lists of such terms to decide if there were indexing terms to which terms on the list could be equated. Such a system could be refined by continuing analysis of user search results and would also provide valuable empirical data concerning the terminology actually used by users.

The techniques described are applicable both to interactive and batch processing systems. An advantage of an interactive system would be its ability to aid the user in learning to write search questions more effectively. However, as demonstrated in this study, individuals

can be trained with a minimum of effort to write good computer questions for batch processing. Additional studies examining this approach both on systems using batch processing and interactive terminals might lead to systems which are more user oriented, thus eliminating intermediary personnel used to translate user requests into computer search questions.

Selective Publication from Information Files

The problem of finding better ways to deal with the rapidly expanding amount of printed information is not solely one of finding better ways to store and retrieve information. Also important to this problem is finding better ways to disseminate organized information. The publication of Exceptional Child Education Abstracts and the related activities using the computerized abstract files provides a model which may aid in dealing with the rapid expansion of knowledge.

By having abstracted information ECEA shares with similar journals the advantage of presenting readers minimal data to help them determine if they should read a specific document. Because the abstracts in ECEA are on a computer file, other alternatives for coping with the rapid expansion of information are available. Included in the procedures made available and used by the CEC-ERIC Information Center are:

1. The use of the computer to generate indexes for inclusion in the abstract journal or in other selected publications.
2. The use of computers to control typesetting of both the abstracts and the computer-generated indexes.
3. The use of computer searching to assist in answering information requests of users.
4. The use of the computer to retrieve data for, and control typesetting of, selected publications.

By using these procedures there are a number of advantages that may not be immediately apparent. First by having the abstracts published in a journal rather than making them available only through computer searches, the information becomes more accessible. Second, by having more comprehensive indexes available through computer indexing, it is less likely that those using the journal will need specific computer searches. Third, by being able to organize, index, and publish selected bibliographies it is possible to answer many user requests without a specific computer search, thus again minimizing the number of specific searches. (At present more than half of all user requests are answered by one or more of the Center's 59 computer-generated bibliographies.) Fourth, a single keying operation provides both for generating computer searchable files and input for computer controlled typesetting.

Not used but also available is the capability of using computer controlled equipment to generate microfilm images of files. This could be done at a minimal cost and would reduce the amount of space required to store abstract journals or special bibliographies.

The variety of ways with which the Center uses a single information file to organize and disseminate knowledge about special education provides a model which may suggest some possible answers to the problem of the rapid growth of knowledge.

With some imagination it is possible to envision procedures whereby Wells' "universal encyclopedia"[233] might become a reality. If one

---

[233]H. G. Wells, "World Encyclopedia," World Brain (Garden City, New York: Doubleday, Doran & Co., Inc., 1938), pp. 3-35. Paper read at the Royal Institution of Great Britain Weekly Evening Meeting, Friday, November 20, 1936.

information file can be used as flexibly as the ECEA file this can be done with other files. If single documents can be indexed and abstracted, information files may also be indexed and abstracted. Thus, by searching one file it would be possible to identify other information files that would be most likely to contain the types of data needed.

BIBLIOGRAPHY

SELECTED BIBLIOGRAPHY

A. BOOKS

Artandi, Susan. An Introduction to Computers in Information Science. Metuchen, N. J.: Scarecrow Press, 1968. 145 pp.

Borko, Harold. Automated Language Processing. New York: John Wiley and Sons, Inc., 1967. 386 pp.

Chorafas, Dimitris N. Systems and Simulation. New York: Academic Press, 1965. 487 pp.

Cleverdon, C. W. Identification of Criteria for Evaluation of Operational Information Retrieval Systems. Cranfield, Bedford, England: Cranfield College of Aeronautics, November, 1964.

_____, Jack Mills, and Micahel Keen. ASLIB Cranfield Research Project, Factors Determining the Performance of Indexing Systems, Vol. 1. Design, Part 1. Text, Part 2. Appendices. Cranfield, Bedford, England: College of Aeronautics, 1966, 337 pp.

Cuadra, Carlos A., Robert V. Katter, Emory H. Holmes, and Everett M. Wallace. Experimental Studies of Relevance Judgments: Final Report 3 vols. Santa Monica, California: System Development Corporation, June, 1967.

Deutsch, Ralph. System Analysis Techniques. Englewood Cliffs: Prentice-Hall, Inc., 1969. 464 pp.

Directory of Educational Information Centers. U. S. Government Printing Office, 1969. Document No. FS5.212:12042.

Directory of Federally Supported Information Centers. Clearinghouse for Federal, Scientific, and Technical Information, April, 1968. PB 477050.

Fairthorne, R. A. Towards Information Retrieval. London: Butterworths, 1961. 211 pp.

Hayes, R. M., and Joseph Becker. Handbook of Data Processing for Libraries. New York: Riley-Hayes-Becker Publications, a subsidiary of John Wiley & Sons, Inc., 1970.

Lancaster, F. Wilfrid. Information Retrieval Systems. New York: John Wiley & Sons, Inc., 1968. 217 pp.

Meadow, Charles T. The Analysis of Information Systems. New York: John Wiley & Sons, Inc., 1967. 301 pp.

Salton, Gerard. Automatic Information Organization and Retrieval. New York: McGraw-Hill Book Company, 1968.

Vickery, B. C. Faceted Classification Schemes. Vol. V of Systems for The Intellectual Organization of Information. Edited by Susan Artandi. New Brunswick, N. J.: The Rutgers University Press, 1966. 108 pp.

_____. On Retrieval System Theory. London: Butterworths, 1965. 183 pp.

Wells, H. G. World Brain. Garden City, New York: Doubleday, Doran & Co., Inc., 1938.

White, Harry J., and Selmo Tauber. Systems Analysis. Philadelphia: W. B. Saunders Company, 1969. 492 pp.

B. ARTICLES AND PERIODICALS

"All About ERIC," Journal of Educational Data Processing, VII (April, 1970), 51-129.

Artandi, Susan. "Computer Indexing of Medical Articles," Journal of Documentation, XXV (September, 1969), 185-282.

_____. "Document Description and Representation," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. Chicago: William Benton, 1970. V, 143-168.

_____, and Edward H. Wolf. "The Effectiveness of Automatically Generated Weights and Links in Mechanical Indexing," American Documentation, July, 1969, pp. 198-202.

Baxendale, P. B. "'Autoindexing' and Indexing by Automatic Processes," Special Libraries, LVI (December, 1965), 715-719.

_____. "Content Analysis, Specification and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. New York: John Wiley & Sons, 1966. I, 71-106.

Borko, Harold. "Design of Information Systems and Services," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. New York: John Wiley & Sons, 1967. II, 35-62.

Bourne, Charles P. "Evaluation of Indexing Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. New York: Interscience Publishers, 1966. I, 171-190.

Burchinal, Lee G. "The Educational Resources Information Center: An Emergent National System," Journal of Educational Data Processing, VII (April, 1970), 55-67.

Clearinghouse on Exceptional Children," Exceptional Children, Summer, 1967, p. 693-694.

Clearinghouse on Exceptional Children, March, 1967.

Cleverdon, Cyril W. "The Cranfield Hypotheses," Library Quarterly, XXXV (April, 1965), 121-124.

_____, F. W. Lancaster, and J. Mills. "Uncovering Some Facts of Life in Information Retrieval," Special Libraries, LV (February, 1964), 86-91.

Commission and the Council for the International Union of Chemistry. "Definitive Report of the Commission on the Reform of Nomenclature of Organic Chemistry," Journal of American Chemical Society, LX (1933), 3905-25.

Cooper, William S. "Is Interindexer Consistency a Hobgoblin?" American Documentation, July, 1969, pp. 268-278.

Cuadra, Carlos A. (ed.). Annual Review of Information Science and Technology, 5 vols. New York: John Wiley & Sons, 1966-69.

Dale, A. G. and N. Dale. "Some Clumping Experiments for Associative Document Retrieval," American Documentation, January, 1965, pp. 5-9.

deSolla Price, Derek J. "Network of Scientific Papers," Science, CXLIX (July 30, 1965), 510-515.

Doyle, Lauren B. "Indexing and Abstracting by Association," American Documentation, October, 1962, pp. 378-390.

Eller, James L. and Robert L. Panek. "Thesaurus Development for a Decentralized Information Network," American Documentation, July, 1968, pp. 213-220.

"ERIC Excerpt," Exceptional Children, October, 1967, pp. 143-148; April, 1968.

Exceptional Child Education Abstracts, II (November, 1970).

Fischer, Marguerite. "The KWIC Index Concept: A Retrospective View," American Documentation, XVII (April, 1966), 57-70.

Gibson, R. E. "A Systems Approach to Research Management," Part I, Research Management, V (1962), 215 pp.

Gull, C. D. "Seven Years of Work on the Organization of Materials in the Special Library," American Documentation, VII (October, 1956), 320-329.

Hall, A. and R. Fagan. "Definition of a System," General Systems, Vol. I of Yearbook of the Society for General Systems. 1956.

Hyslop, Marjorie R. "Sharing Vocabulary Control," Special Libraries, LVI (December, 1965), 708-714.

Jordan, June B. "CEC-ERIC-IMC" A Program Partnership in Information Dissemination," Exceptional Children, XXXV (December, 1968), 311-313.

King, Donald W. "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor, Chicago: Encyclopedia Britannica, 1968. III, 61-104.

Kochen, Manfred. "Systems Technology for Information Retrieval," The Growth of Knowledge, Manfred Kochen, editor. New York: John Wiley and Sons, 1967. pp. 352-372.

Lancaster, F. W. "Evaluating the Performance of a Large Operating Retrieval System." Electronic Handling of Information. Allen Kent, Orrin E. Taulbee, Jack Belzer, and Gordon D. Goldstein, editors. Washington, D. C.: Thompson Book Company, 1967. pp. 199-216.

_____. "MEDLARS: Report on the Evaluation of its Operating Efficiency," American Documentation, April, 1969. pp. 119-142.

_____, and Constantine J. Gillespie. "Design and Evaluation of Information Systems," Annual Review of Information Science and Technology, Carlos A. Caudra, editor. Chicago: Encyclopedia Britannica, Inc., 1967. V, 33-70.

_____, and J. Mills. "Testing Indexing and Index Language Devices: The ASLIB Cranfield Project," American Documentation, XV (January, 1964, 4-13.

Lesk, M. E. and G. Salton. "Relevance Assessments and Retrieval System Evaluation," Information Storage and Retrieval, December, 1968, pp. 343-359.

Luhn, H. P. "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, II (1958), 159-165.

_____. "Keyword-In-Context Index for Technical Literature," American Documentation, XI (1960), 288-295.

Montgomery, Christine, and D. R. Swanson. "Machine Like Indexing by People," American Documentation, XIII (October, 1962), 359-66.

Moon, R. D., and J. F. Vinsonhaler. "The Title-Generated Thesaurus: A Practical Method for Automated Indexing," Proceedings of the Sixth Annual National Colloquium on Information Retrieval - The Information Bazaar. Philadelphia: The Medical Documentation Service of the College of Physicians, 1969.

O'Connor, John. "Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems," American Documentation, XV (1964), 96-104.

Perry, Peter. "Combined Grouping for Coordinate Indexes. American Documentation, XIX (April, 1968), 142-145.

Price, Nancy, and Samuel Schiminovich. "A Clustering Experiment: First Step Towards a Computer-Generated Classification Scheme," Information Storage and Retrieval, IV (August, 1968), 271-280.

Research in Education, July, 1967.

Salton, Gerard. "A Comparison Between Manual and Automatic Indexing Methods," American Documentation, January, 1969, pp. 61-71.

_____. "The Evaluation of Automatic Retrieval Procedures--Selected Test Results Using the SMART System," American Documentation, July, 1965, pp. 209-222.

_____, E. M. Keen, and M. Lesk. "Design Experiments in Automatic Information Retrieval," The Growth of Knowledge, Manfred Kochen, editor. New York: John Wiley & Sons, Inc., 1967. pp. 336-351.

Schultz, Claire K. "Do-It-Yourself Retrieval System Design," Special Libraries, LVI (December, 1965), 720-723.

Sharp, John R. "Content Analysis, Specification and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. New York: John Wiley & Sons, 1967. II, 87-122.

Simmons, Robert F., Sheldon Klein, and Keren McConlogue. "Indexing and Dependency Logic for Answering English Questions," American Documentation, July, 1964, pp. 196-204.

_____, and Keren L. McConlogue. "Maximum Depth Indexing for Computer Retrieval of English Language Data," American Documentation, January, 1963, pp. 68-73.

Sparck, Jones, Karen and Roger M. Needham. "Automatic Term Classifications and Retrieval," Information Storage and Retrieval, IV (June, 1968), 91-100. (Presented at the First Cranfield International Conference on Mechanized Information Storage and Retrieval Systems, College of Aeronautics, Cranfield, England, 29-31 August, 1967.

Swanson, Don R. "The Evidence Underlying the Cranfield Results," The Library Quarterly, XXXV (January, 1965), 1-20.

Swets, John A. "Information-Retrieval Systems," Science, CXLI (July 19, 1963), 245-250.

Tate, F. A. "Handling Chemical Compounds in Information Systems," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. New York: Interscience Publishers, 1967. II, 285-310.

Taulbee, Orrin E. "Content Analysis, Specification, and Control," Annual Review of Information Science and Technology, Carlos A. Cuadra, editor. Chicago: William Benton, 1968. III, 105-136.

Wyllys, Ronald E. "Extracting and Abstracting by Computer," Automated Language Processing, Harold Borko, editor. New York: John Wiley & Sons, Inc., 1967. pp. 127-180.

Zunde, Pranas, and Margaret E. Dexter. Indexing Consistency and Quality," American Documentation, XX (July, 1969), 259-267.

C. REPORTS, TECHNICAL MANUALS, AND UNPUBLISHED MATERIAL

Burchinal, Lee G. Development of ERIC Through December, 1968. Division of Information Technology and Dissemination, Bureau of Research, U. S. Office of Education, Department of Health, Education and Welfare, Office of Education/ Office of Information Dissemination. First Printed in August 1969, Revised February, 1970.

_____. Evaluation of ERIC, June, 1968. Report from U.S. Department of Health, Education and Welfare, Office of Education, Bureau of Research. Available as ED 020449. Bethesda, Maryland: ERIC Document Reproduction Service, 1968.

CEC-ERIC Information Center. "Processing Costs & Formulas," an unpublished summary prepared under the direction of Carl Oldsen. September, 1970.

Jordan, June B. "Handicapped Children and Youth ERIC Clearringhouse on Research Dissemination," a proposal submitted to the U.S. Department of Health, Education and Welfare, Bureau of the Handicapped, 1966, 8 pp.

Oldsen, Carl, ECEA editor. Unpublished statistical information based on an analysis of 5,715 acquisitions in Volumes I and II of ECEA.

_____. Unpublished statistical information on user requests.

Price, Samuel T. "The Development of a Thesaurus of Descriptors for an Information Retrieval System in Special Education," Unpublished doctoral dissertation, University of Pittsburgh, 1969, abstract.

_____, (comp.). Thesaurus of Descriptors for an Information Retrieval System in the Subject Matter Area of Special Education, Normal, Illinois: Special Education Instructional Materials Laboratory, Illinois State University, January, 1970.

Rees, Allan, and Douglas G. Schultz, Principal investigators. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. Final Report to the National Science Foundation. Cleveland: Center for Documentation and Communication, School of Library Science, Case Western Reserve University, October, 1967, I, 287 pp.; II, Appendices A-Q.

Stevens, M. E. Automatic Indexing: A State of the Art Report. NBS Monograph 91. Washington: National Bureau of Standards, March 1965.

Thesaurus of ERIC Descriptors: Working Copy Descriptor Listing, ERIC Processing and Reference Facility. Bethesda: Maryland: Leasco Systems and Research Corporation, August, 1971. 224 pp.

Thesaurus for Exceptional Child Education. Arlington, Virginia: CEC-ERIC Information Center on Exceptional Children, 1971. 12 pp.

Trester, Delmer J, System Coordinator, Department of HEW, Office of Education. Statistics on ERIC accompanied by cover letter to Carl Oldsen, CEC-ERIC, February 16, 1971.

Vinsonhaler, John F. The Information Systems Laboratory: A Progress Report for 1969. ISL Report No. 10. East Lansing: Michigan State University, January, 1970.

_____(ed.). Technical Manual Basic Indexing and Retrieval System BIRS 2.0. East Lansing: Educational Publications Services, College of Education, Michigan State University, January, 1968.

_____, John M. Hafterson, and Stuart W. Thomas, Jr. (editors). Basic Indexing Retrieval System Technical Manual. 10 vols. East Lansing, Michigan: Information Systems Laboratory, College of Education, Michigan State University, 1970.

_____, and John M. Hafterson (editors). Technical Manual for Basic Indexing and Retrieval System, BIRS 2.5, Appendix I. East Lansing: Educational Publications Services, College of Education, Michigan State University, January, 1969.

Weinberg, Alvin. Science, Government, and Information: The Responsibilities of the Technical Community and the Government in the Transfer of Information. President's Science Advisory Council. Washington: Government Printing Office, 1963.

APPENDIX A

APPENDIX A

A Description of the Operating Procedures
Used by the CEC-ERIC Information Center

The procedures described in this Appendix are an expansion upon the overview of the operating procedures found in Chapter 3. There have been some duplications of the material in Chapter 3 so that the Appendix may be read in an entirety without referring to other sections of the text.

Legend and Nomenclature

The symbols used in the following diagrammatic representations of CEC-ERIC's processing are those commonly used on computer program and systems flowcharting. Occasional liberties are taken by using a single symbol to imply more than is commonly done in computer programming; however, in these cases the operations represented by the symbol will be described verbally. The descriptions of the symbols found in Figure 1A are those given on the cover of an IBM flowcharting template, Form X 20-8020.

In addition to the symbols in Figure 1A the following alphabetic legend will be used to identify specific symbols in various Figures:

1. C (n) stands for Connection number N where N may be any number and the connection may be between flowcharts or within the same flowchart.

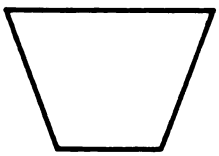
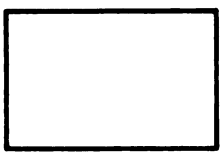
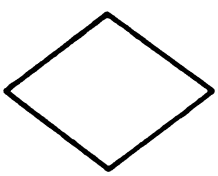
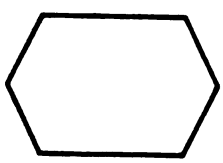
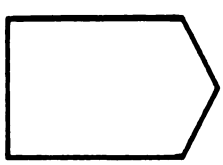
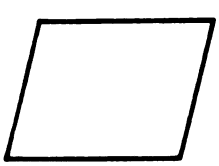
SYMBOL	REPRESENTS
	INPUT/OUTPUT Any function of an input/output device (making information available for processing, recording processing information, tape positioning, etc.)
	PROCESSING A group of program instructions which perform a processing function of the program.
	DECISION The decision function used to document points in the program where a branch to alternate paths is possible based upon variable conditions.
	PREDEFINED PROCESS A group of operations not detailed in the particular set of flowcharts.
	PROGRAM MODIFICATION An instruction or group of instructions which changes the program.
	CLERICAL OPERATION A manual offline operation not requiring mechanical aid.

FIGURE 1A

FLOWCHARTING SYMBOLS

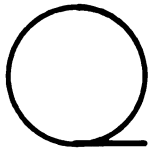
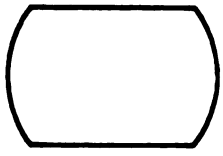
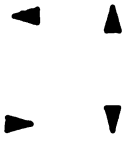
SYMBOL	REPRESENTS
	DOCUMENT Paper documents and reports of all varieties.
	MAGNETIC TAPE
	PUNCHED CARD All varieties of punched cards including stubs.
	KEYING OPERATION An operation utilizing a key-driven device.
	FLOW DIRECTION The direction of processing or data flow.
	CONNECTOR An entry from, or an exit to, another part of the program flowchart.
	OFFPAGE CONNECTOR A connector used instead of the connector symbol to designate entry to or exit from a page.

FIGURE 1A (cont'd)

2. D (N) stands for Decision number N.
3. IP (N) stands for Input number N.
4. OP (N) stands for Output number N.
5. PP (N) stands for Predefined Process N.
6. S (N) stands for Step (N).
7. SB (N) stands for Symbol N of a given Figure. This will be used when it is necessary to identify a given symbol for discussion which is not identified in another way.

Sometimes the number N may be a decimal number such as 1.1 or 1.12. This is used to tie various closely related operations together. For example, S1 would stand for Step 1, S1.1 would stand for a small step in the major step 1. Additional points past the decimal will be used to indicate further refinement of steps. For example, Steps S7.11, S7.12, and S7.13 would all have functions in common to those represented by the S7.1. This legend allows the reader to identify similar steps appearing in various portions of the flowchart.

Overview of the Information Center's Major Activities

Figure 2A provides a simplified overview of the Information Center processing by dividing the processing into six major activities; document acquisition, document management, file maintenance, file processing, information processing, and evaluation with system modification. The core of activities found in Figure 2A are presented with greater detail in the later diagrams Figures 3A, 4A, and 5A. The activities described are found in most information centers utilizing computer processing; however, the specific steps and products resulting in these broadly defined activities will vary considerably from center to center.

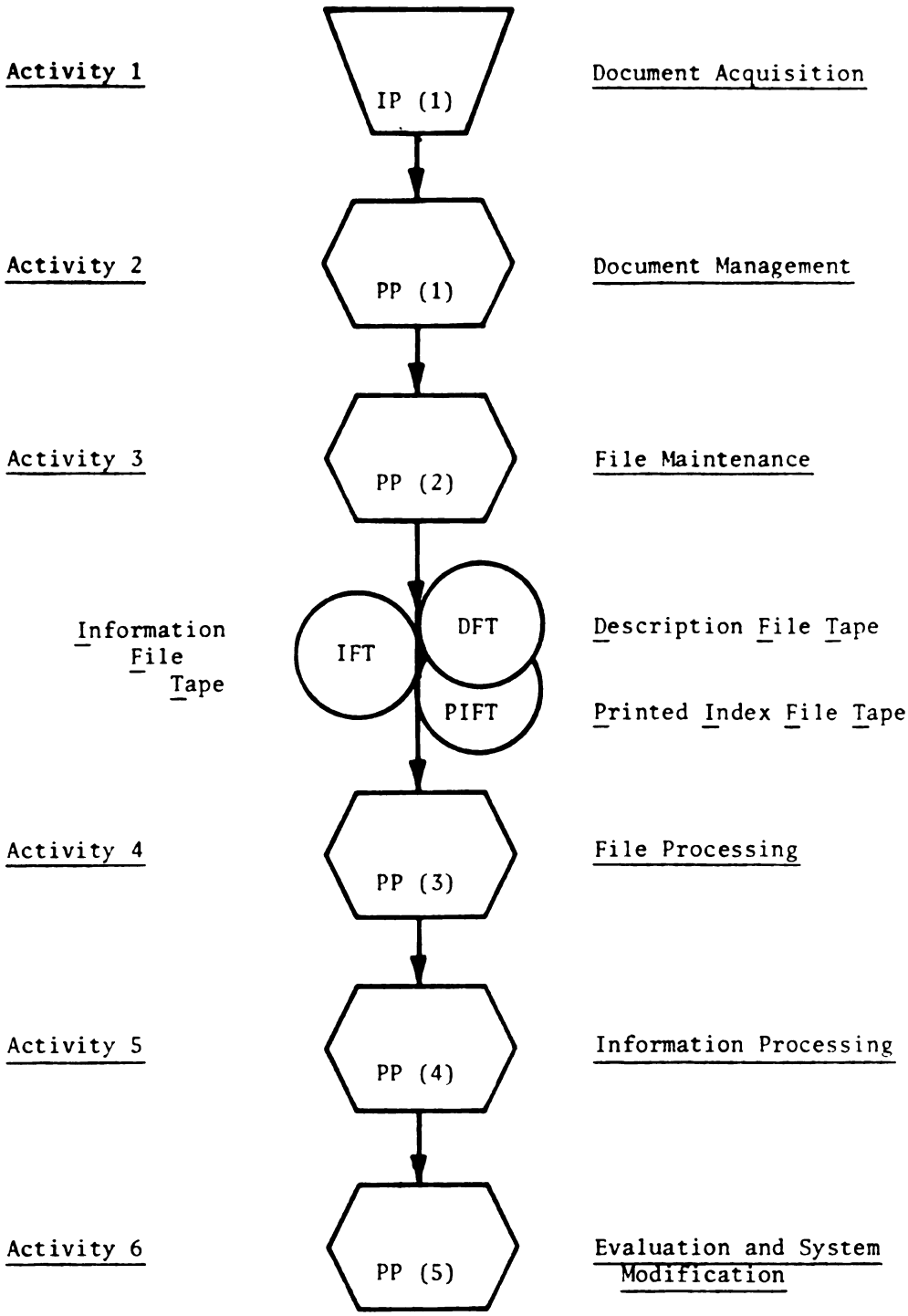


FIGURE 2A

OVERVIEW OF INFORMATION CENTER MAJOR ACTIVITIES

Briefly these activities can be described in the following manner:

Activity 1 - Document Acquisition This activity includes the selection of documents which will be bought or acquired by other methods so that they may be examined to determine if they are appropriate for inclusion in Information Center holdings.

Activity 2 - Document Management This activity includes examining documents to determine if they should be included in the Information Center data bank, the abstracting of documents, the indexing of documents, and the cataloging of documents.

Activity 3 - File Maintenance This activity includes keypunching document surrogates, storing the document surrogates on a computerized information file and preparing computerized description files and printed index files.

Activity 4 - File Processing This activity includes computer processing of files to organize the information in a form that will be more useful and easier to disseminate.

Activity 5 - Information Processing This activity involves processing user requests, providing users with information, publishing new documents from information contained on the computer files, and providing information to be used in evaluation of the system. The activities in this section are primarily manual activities but they may initiate computer file processing (Activity 4) as one of the several steps in the procedure.

Activity 6 - Evaluation This activity involves examining the procedures used by the Center and, if appropriate, modifying these pro-

cedures to make the total operation of the Information Center more effective.

Overview of Major Input and Output

Figure 3A provides an overview of the major input to the Information Center and the output generated as a result of processing the input. Documents are acquired (IP (1)) and processed in the document management activities (PP (1)) to generate copy for Research in Education (OP (1)) and Current Index to Journals in Education (OP (2)). All documents which will eventually become part of the Information Center holdings, including those which are processed for RIE and CIJE, are then put in the form used on the Center's information files and passed to file maintenance processing (PP (2)). In the file maintenance activity the documents are put in computer readable form and various computer files are generated. These computer files provide input for the file processing (PP (3)) and output for selected publication (OP (4)). This output is in a form that allows for computer typesetting, and computer-generated indexes, and printing with a minimum effort.

ECEA and the selected publications in turn become input Information Processing (PP (4)). These with input from CEC publications, user requests, and additional file processing, are utilized in providing information to users (OP (5)) and in assisting staff members in generating new documents (OP (6)).

Overview of Evaluation and Processing Modifications

Figure 4A provides an overview of the continuing evaluation which is used to monitor and if appropriate modify the processing of the In-

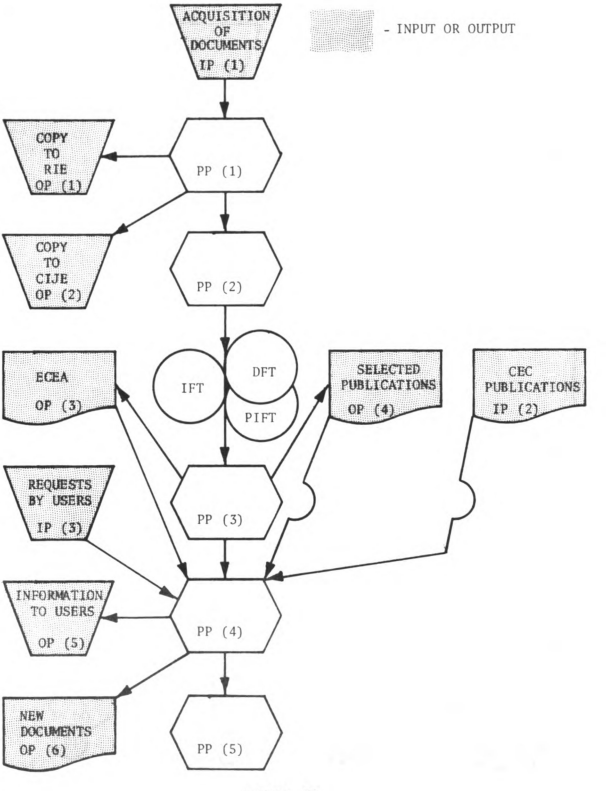


FIGURE 3A

OVERVIEW OF MAJOR INPUT AND OUTPUT

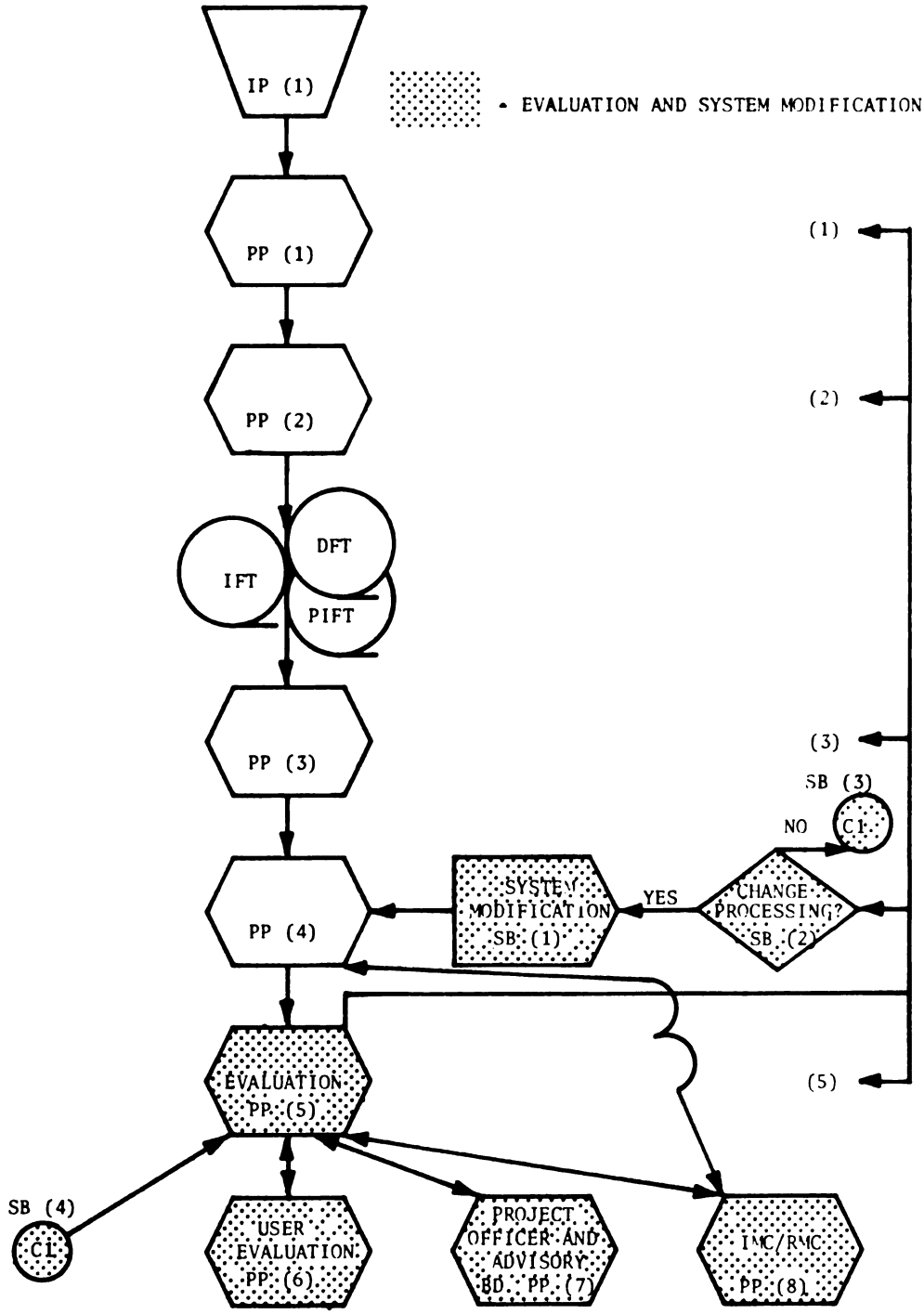


FIGURE 4A

AN OVERVIEW OF THE INFORMATION CENTER'S
EVALUATION AND SYSTEMS MODIFICATION COMPONENTS

formation Center. Input to the evaluation component is provided from information processing (PP (4)), user evaluation (PP (6)), the project officer and advisory board (PP (7)), and the IMC/RMC Network (Instructional Materials Center/Regional Media Center Network) (PP (8)). The arrows going in both directions indicate that there is an interaction between evaluation and other components. The input from the various sources is processed by the evaluation component to determine if there are system modifications which should be made. The flow of decisions is illustrated by symbols SB (1), SB (2) and in the modification occuring to (PP (4)). The numbers 1, 2, 3, and 5 appearing within parentheses opposite arrows indicate that the same series of symbols; namely, SB (1), SB (2), and SB (3) would appear at these points and be connected to the predefined processes PP (1), PP (2), PP (3), and PP (5) as done in later diagram Figure 5A. If no change is made this serves as feedback to the evaluation procedures as indicated by the connection C (1) to PP (5).

Overview and Model of the Information Center's Operation

Figure 5A provides an overview of the Information Center's processing. In this overview six major activities can be seen in the center of the flowchart. The input and output operation shown in Figure 3A is present as well as the evaluation procedures indicated in Figure 4A. The model as presented indicates a continual flow of input, output, evaluation, and appropriate systems modification to improve the operation procedures. Figure 5A and the more simplified Figures 2A, 3A, and 4A can be used as a reference as the more detailed

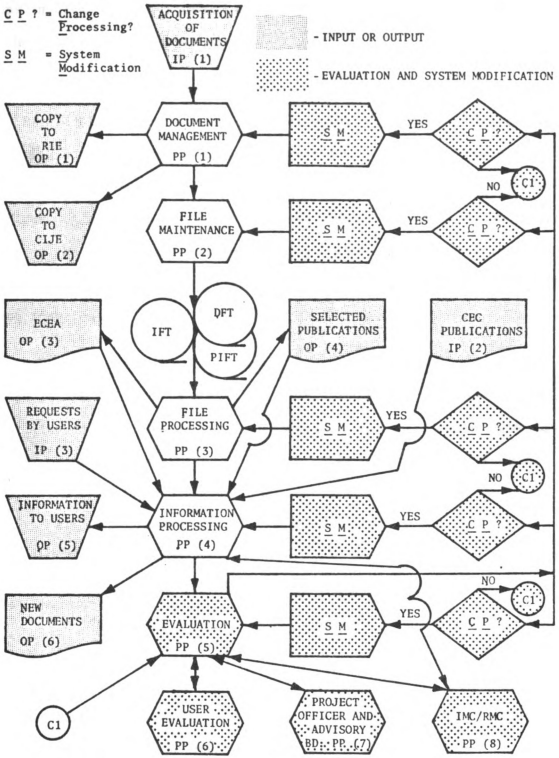


FIGURE 5A

AN OVERVIEW AND MODEL OF THE INFORMATION CENTER'S OPERATIONS

steps involved in the Information Center's operations are discussed in the following sections.

Acquisition Control and Document Management

Figure 6A provides a detailed description of the steps involved in acquiring documents and preparing document surrogates to be placed on the Center's information file. The four sources of input are identified in Figure 6A as IP (1.1) through IP (1.4) and are an expansion of IP (1) found in the overview of Figure 5A. Step 2 through Step 5 relate to the predefined process PP (1) of Figure 5A with the output OP (1) and OP (2) being identical to that on Figure 5A.

Step 1 - Acquisition of Documents Documents which become part of the Information Center holdings are obtained from four major sources, identified in Figure 6A as IP (1.1) through IP (1.4). These sources are:

1. Journals containing articles which will be processed for use in CIJE. These journals are divided into two categories:
   a. Journals where all articles are automatically processed for use in CIJE.
   b. Journals containing articles which are examined to determine if they are relevant for processing in CIJE.
2. Journals which contain articles that are not considered for CIJE, but are considered for ECEA.
3. Documents ordered from various publishers as a result of publishers' announcements and literature reviews.
4. Documents which are donated or suggested by various sources to the Information Center. Two major sources of contribution are:

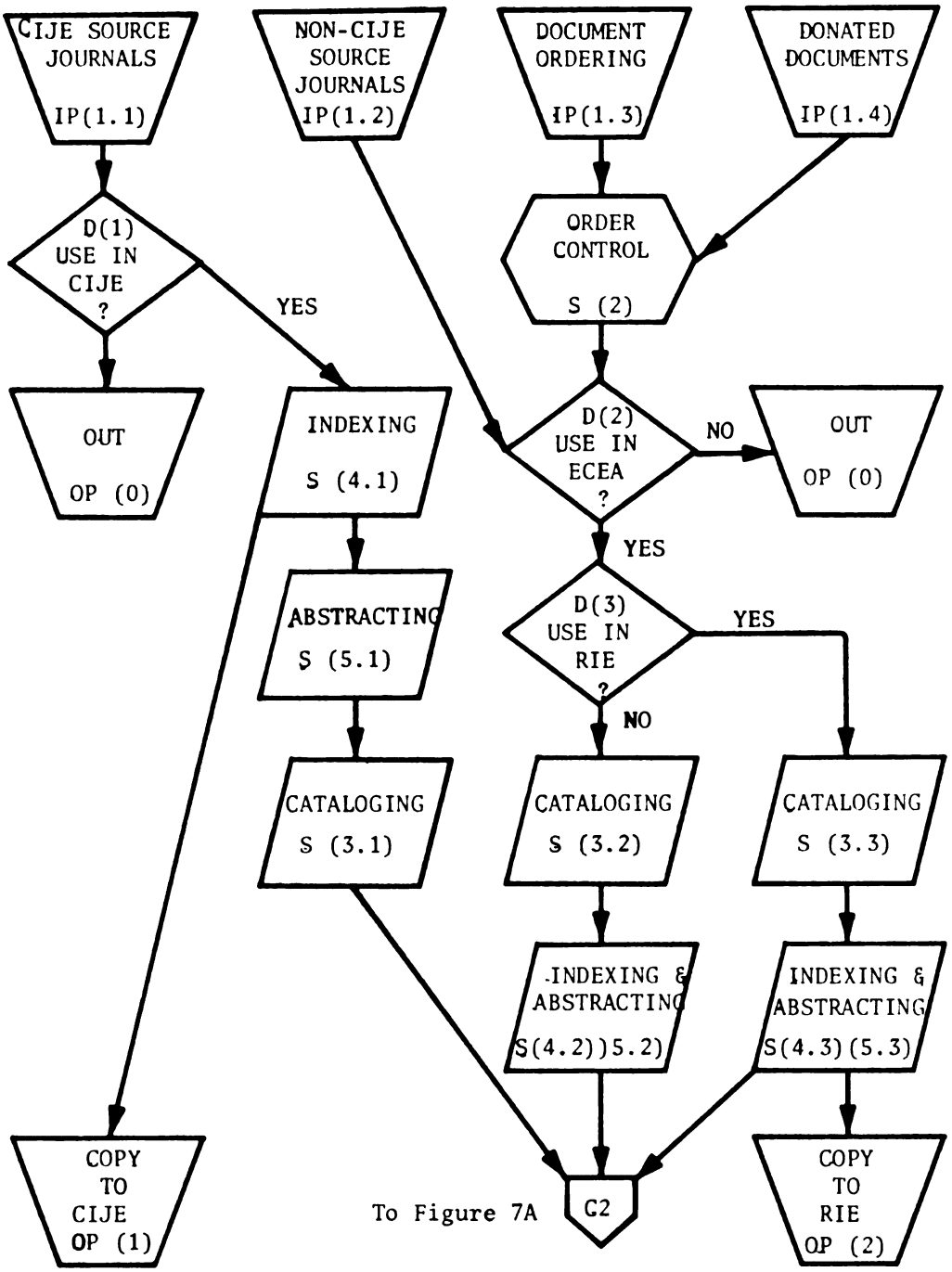


FIGURE 6A

ACQUISITION CONTROL AND DOCUMENT MANAGEMENT

a. Documents relating to research projects contributed by the U. S. Department of Health, Education and Welfare.

b. Documents relating to instructional procedures and media recommended by the IMC/RMC Network.

Step 2 - Order Control The second step is to process orders in a way that will prevent a document from appearing more than once on an information file. To prevent this, it is necessary to determine that documents ordered are not already part of the Information Center's holdings and that documents coming to the Center through orders and donations do not contain duplicates. This is done by filling out a form on all documents which are obtained from either commercial sources or which are donated to the Center. Included on the form is information about the author, publisher, title, and number of pages. This information is keypunched and placed on a computerized file which is sorted to generate separate listings by title order, author-publisher order, and by author-title-publisher order. These listings and similar sorted lists of documents already processed are used to prevent duplication.

Step 3 - Cataloging The third step for documents not used in CIJE is cataloging. This includes placing on a processing form the title, author, source where the document may be obtained, pulication date, number of pages, and assigning an EC number. The EC number is a six-digit number where the first two digits refer to a volume number of Exceptional Child Education Abstracts and the last four digits to an abstract number. For example, if the document had the number EC 03 1234, this would indicate that the document surrogate is abstract 1234 in Volume III of Exceptional Child Education Abstracts.

Step 4 - Indexing In the fourth step indexers assign terms from the ERIC Thesaurus[1A] to describe the document. All documents which are processed for use in CIJE are sent to Central ERIC after these indexing terms have been assigned, as is indicated by S (4.1) and OP (1) of Figure 6A. Since the beginning of Volume II, the Information Center has used a subset of the ERIC Thesaurus to prevent the unnecessary proliferation of terms with similar meanings.[2A]

Step 5 - Abstracting In Step 5 an abstract is written for each document or, where permission has been granted from specific journals, the author abstract is used. In all documents except journal articles processed for CIJE the indexing and abstracting are done simultaneously. The documents processed for inclusion in CIJE also differ in processing order in that the cataloging is not done until the document has been indexed and abstracted. The reason for this is that the copy sent to CIJE contains only indexing and bibliographic information and does not contain a summary (abstract). Regardless of the original acquisition source, all document surrogates which become a part of the Information Center files or are published in ECEA are processed according to the same criteria. At the point in processing indicated by C (2) all document surrogates contain the information which will be keypunched for

---

[1A]Thesaurus of ERIC Descriptors, ERIC Processing and Reference Facility, Operated for U. S. Office of Education by Leasco Systems & Research Corporation, 4833 Rugby Avenue, Bethesda, Maryland, 1970, p. 82.

[2A]Thesaurus for Exceptional Child Education, (Arlington, Virginia: Council for Exceptional Children, Information Center on Exceptional Children, 1970), p. 1-10.

inclusion on the Center's information files. Copies of document surrogates which are to become part of Research in Education are sent to Central ERIC as indicated by OP (2).

File Maintenance

Figure 7A provides a detailed description of the steps involved in preparing computer readable copy and placing this copy on the Center's information files. The steps described correspond to predefined process PP (2) which generates the IFT (Information File Tape), DFT, (Description File Tape) and PIFT (Printed Index File Tape) of the overview presented in Figure 5A.

Step 6 - Preparation of Computer Readable Copy Step 6 is a sequence of small tasks involving repeated keypunching, computer processing, and proofreading. Step 6 involves keypunching the document surrogates which include the information resulting from the cataloging, indexing, and abstracting steps. The documents are initially punched in as free a format as possible with distinct types of information (fields) designated with an equals sign followed by a letter. Each document surrogate is separated by an *$ABSTRACT card followed by the last four digits of the EC number assigned in the cataloging step.

In Step 6.21 the keypunched cards are read into Preprocessor Program I which adds a sequence number to each line. For example if there are a total of 3,500 cards the sequencing would be from 1 to 3,500. The program then provides a listing, printing one abstract per page with the abstract number and the sequence information, and punches a new deck of cards including the sequencing information on the right side of each card.

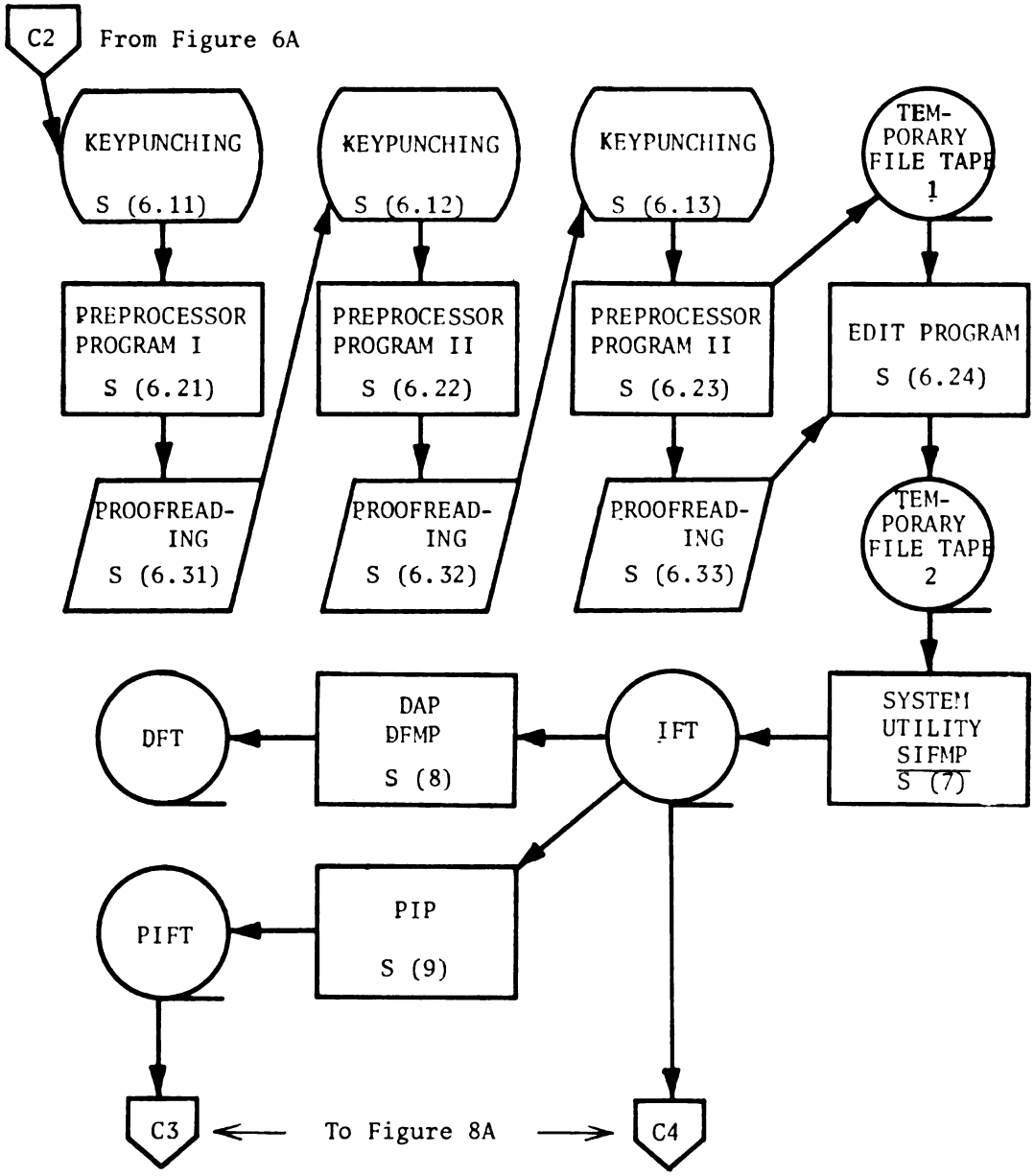


FIGURE 7A

FILE MAINTENANCE

The listing is proofread (Step 6.31) and corrections are sent for keypunching (Step 6.12) where corrections are punched and inserted into the new deck that was punched by Preprocessing Program I. The sequencing information on the right side of the card is used to locate and insert the corrections.

In Step 6.22 the corrected deck is used as input into Preprocessing Program II, which:

1. Converts the two letter field code into full field names;
2. Inserts control codes to be used in computer typesetting;
3. Sequences the document surrogates placing the abstract number followed by the card within that abstract in the right portion of the line;
4. Checks each EC number to see that it corresponds with the abstract number;
5. Prints the abstract number on a new page
6. Prints an error message for each field code or EC number inconsistent with what is expected;
7. Prints the correct abstract number or (corrected abstract number) on the right of each abstract record;
8. Punches a new deck containing the complete field name, code for computer typesetting, and sequencing information; and
9. Provides a listing to be used in proofreading.

In Step 6.32 the new listing is proofread and corrections sent to keypunchers. In Step 6.13 the corrections are keypunched and inserted by a data processor who then uses the corrected deck as input to Preprocessing Program II (Step 6.23) which generates a file tape (Temporary File Tape 1) and a listing indicating the corrections made in Step 6.13.

In Step 6.33 the listing is proofread and corrections if any keypunched for use in the editing program, S (6.24), which can replace entire abstracts or change lines within the abstracts as needed to generate a corrected file tape (Temporary File Tape 2).

In the first two keypunching/preprocessing/ proofreading operations abstracts are handled in batches of 50. In the third sequence of keypunching/preprocessing/proofreading (S (6.13), S (6.23, and S (6.33)), ten of the previous batches are grouped together to create a batch of 500 abstracts. The number of the first abstract in the batch is indicated and used as a parameter by the preprocessing program to check the following sequential abstracts to determine if they have the correct abstract and EC number. If there is disagreement with what is expected this is printed out as part of an error message which is examined in the proofreading Step 6.33.

Step 7 - The Creation of an Information File Tape In Step 7 a systems utility program adds the 500 abstracts to the information file tape containing the previous abstracts for the current volume of ECEA. This utility program is designated as SIFMP, standing for Surrogate Information File Maintenance Program. As was previously mentioned, BIRS (Basic Information and Retrieval System) is designed in modular format so that portions of programs or entire programs can be replaced by more economic special purpose programs. Step 7 is an example where a systems utility program was used to replace an operation which might have been done at more cost by the BIRS Information File Maintenance Program (IFMP).

The Information File Tape (IFT) created in Step 7 is in a line image format containing all the information that will appear in the

published abstracts in ECEA and the information necessary to control the computer typesetting in a later step. This information file tape with information file tapes from previous volumes is the data base for a variety of Information Center activities. Because of the importance of these tapes and the work required in their preparation, multiple backups of the tapes are kept in separate locations.

When the BIRS system was first developed, random access disk storage did not have the wide use and lower cost it does today. Many of the operations which relate to the Information File could now be more efficiently done using disk storage. Work is presently being done to develop an Information File Maintenance module which will take advantage of disk storage. Because of BIRS modular structure this can be done in a way that will not change the user's procedures, thus, again illustrating the advantage of a modular design.

Step 8 - The Creation of a Description File Tape In Step 8 the BIRS program Descriptive Analysis Program, DAP is used to extract descriptive terms from designated fields of the information file. In the processing at the CEC-ERIC Information Center the terms are extracted from the title field, the author field, the descriptors field, the date of publication field and a categories field. The information selected by DAP is put on a temporary file which is processed by the BIRS Descriptive File Maintenance Program (DFMP) to generate a description file tape or add new information to an existing Description File Tape, (DFT).

There is a one-to-one correspondence between the information on the DFT and IFT; i.e., for each abstract on the IFT there is a descrip-

tion on the DFT such that description 1 on the DFT corresponds to abstract 1 on the IFT. The purpose of the DFT is to have a subset of the information contained on the IFT which will be useful in finding documents in computerized search. By using a tape with less information (the DFT) for searching, the speed of computer searching is increased. The DFT serves as the basis for all computerized searches done by the BIRS programs and as with the IFT, multiple backups of this tape are kept in separate locations.

Step 9 - Preparation of Printed Indexes In Step 9 the BIRS Printed Indexing Program, (PIP) is used to select information from the information file tape and order this information to create a printed indexing file tape (PIFT). As with the Descriptive Analysis Program, information from selected fields can be extracted and used to create from the information file any number of different indexes.[3A] The Information Center primarily uses the program to provide indexes of the author field, the title field, and the descriptors field, all of which are published as part of Exceptional Child Education Abstracts.

File Processing for ECEA

Once information files have been established there are many ways that these files can be processed to organize the information and generate new products. Figure 8A illustrates the steps taken in file

---

[3A]John F. Vinsonhaler, John M. Hafterson, Stuart W. Thomas, Jr. (editors), Basic Information Retrieval System Technical Manual (East Lansing, Michigan: Information Systems Laboratory, College of Education Michigan State University, 1970), V, 901-956.

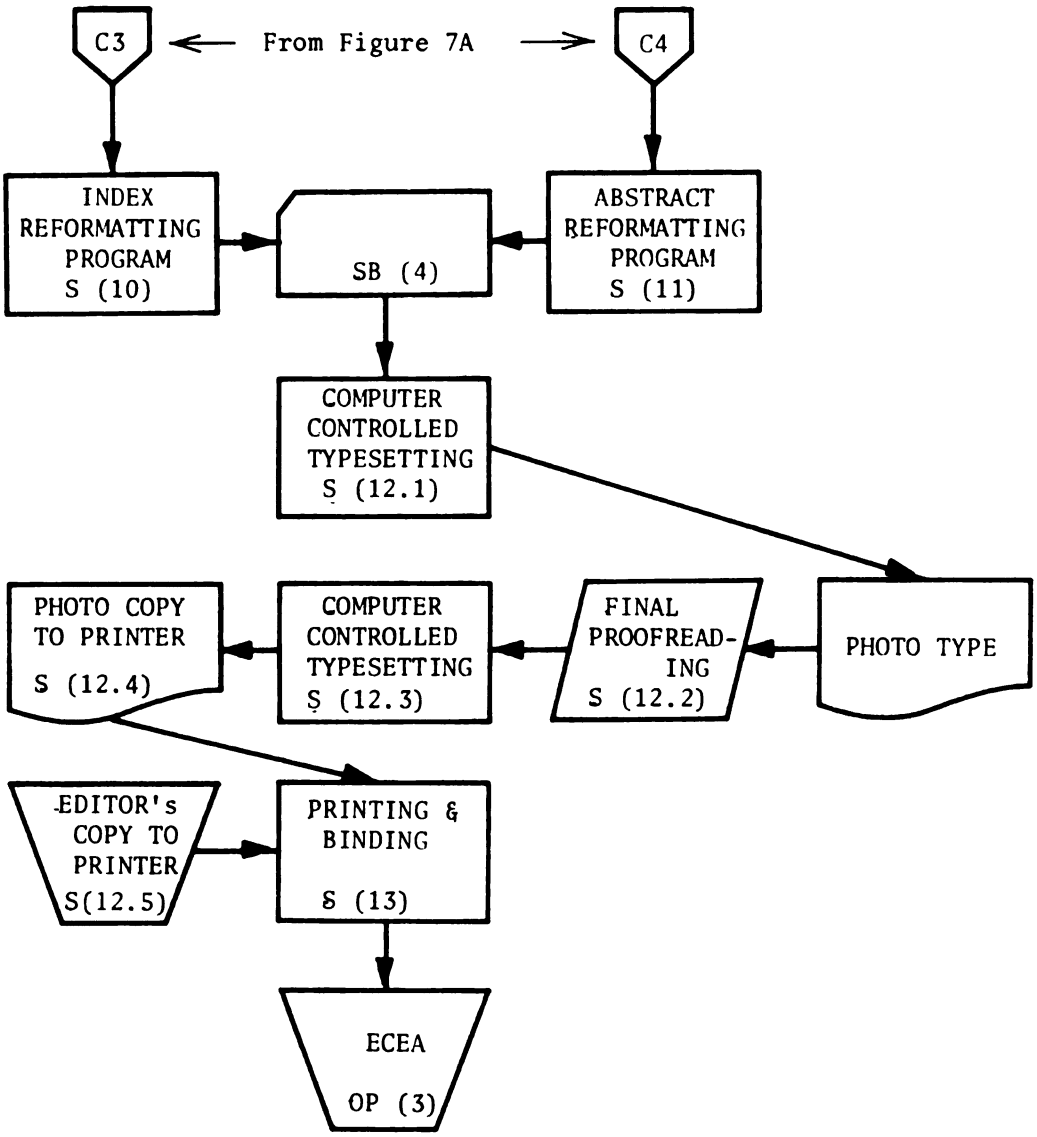


FIGURE 8A

FILE PROCESSING FOR EXCEPTIONAL CHILD EDUCATION ABSTRACTS

processing to generate the journal Exceptional Child Education Abstracts. In the overview of the Information Center's major activities this would occur as part of the predefined process PP (3) labeled "file processing." With slight modification the steps illustrated in Figure 8A could be used to generate a variety of products some of which will be discussed in a later section.

Step 10 and Step 11 - Preparation of Input for Computer Typesetting In Step 10 an index reformatting program reads the printed indexed file tape (PIFT), adds coded information to be used in computer typesetting, and punches out a deck which will be utilized in computer typesetting. In Step 11 an abstract reformatting program reads the IFT, selects all or portions of specified abstracts from the tape and provides a deck of punched cards in the form that will be used in computer typesetting. The program used in this step does not change the abstract in any manner except to delete portions which are not to be printed; however, it would be possible to have a program restructure the abstract to fit a new format if needed.

Step 12 - Preparation of Copy for Printers The input of punched cards from steps 10 and 11 are used by a computer program run on an IBM 1130 to punch a paper tape which is input for a phototypesetter. The paper tape contains information concerning the various type faces that are to be used, the width of the line, and how the line is to be set. The camera-ready offset copy generated in Step 12 is given a final proofreading in Step 12.2, corrections needed are reset by the phototypesetter in Step 12.3, and the corrected copy is prepared in page format and sent to the printer in Step 12.4.

In Step 12.5 copy which is not computer generated (ads, instructions about how to use the journal, etc.) is prepared by the editorial staff and sent to the printers to be merged with the computer-generated copy.

Step 13 - Printing and Binding In Step 13 the camera-ready pages provided by the computer-controlled phototypesetting and the additional copy provided by the editor are used to prepare the offset plates. ECEA is then printed, bound and sent to the Council for Exceptional Children for distribution.

Processing Information Requests

The major objective of any information center is to disseminate information to its users in a form that will be most effective. An examination of the overview of the Center's activities found in Figure 3A shows how the various inputs and outputs of the Center revolve around this activity designated PP (4). Even the evaluation activity described in Figure 4A obtains information from this step and exists solely to determine how the total procedures may be modified to more effectively carry out the information processing and dissemination procedures.

The Information Center processes two major categories or requests, (1) requests made by CEC-ERIC staff members involved in generating new documents or organizing information so that it may be more effectively used by those outside the Center; and (2) requests made by a variety of individuals outside the Center. Included in the various types of users that are not on the CEC-ERIC staff are: (1) educational administrators and decision makers, (2) federal and public agencies, (3) parents,

(4) psychologists, (5) public officials, (6) research and development specialists, (7) social workers, (8) special education supervisors and consultants, (9) staff members of professional organizations, (10) students, (11) teacher educators, and (12) teachers.

Often the requests from users outside the Center can be answered by documents which have been prepared by the CEC-ERIC staff or reprints of CEC publications; however, when needed the Center has a set of powerful computer search programs to aid in answering difficult questions. Figure 9A illustrates a predefined operation used at the Center for computer searches which will later be referred to as Predefined Process PP (9). In this Figure an information request IP (3) is first translated into a computer searchable question by one of the CEC-ERIC staff members processing user requests, SB (5). Next the question is read by the Description File Search Program (DFSP), which uses information stored on the Description File Tape (DFT) to determine which documents are most closely related to the question. The results of the search with instructions concerning format of output desired (a list of access numbers or a computer printout of the total document surrogate) is placed on a question file tape or disk which is read by the Information File Retrieval Program (IFRP). If the full text of the document surrogate is requested, IFRP obtains this information from the Information File Tape (IFT). If not, the access numbers and original questions are output to a line printer or a temporary storage device OP (6). The total search operation excluding the information request IP (3) and the output generated OP (6) will hereafter be referred to as PP (9). Detailed information concerning the various searching alternatives avail-

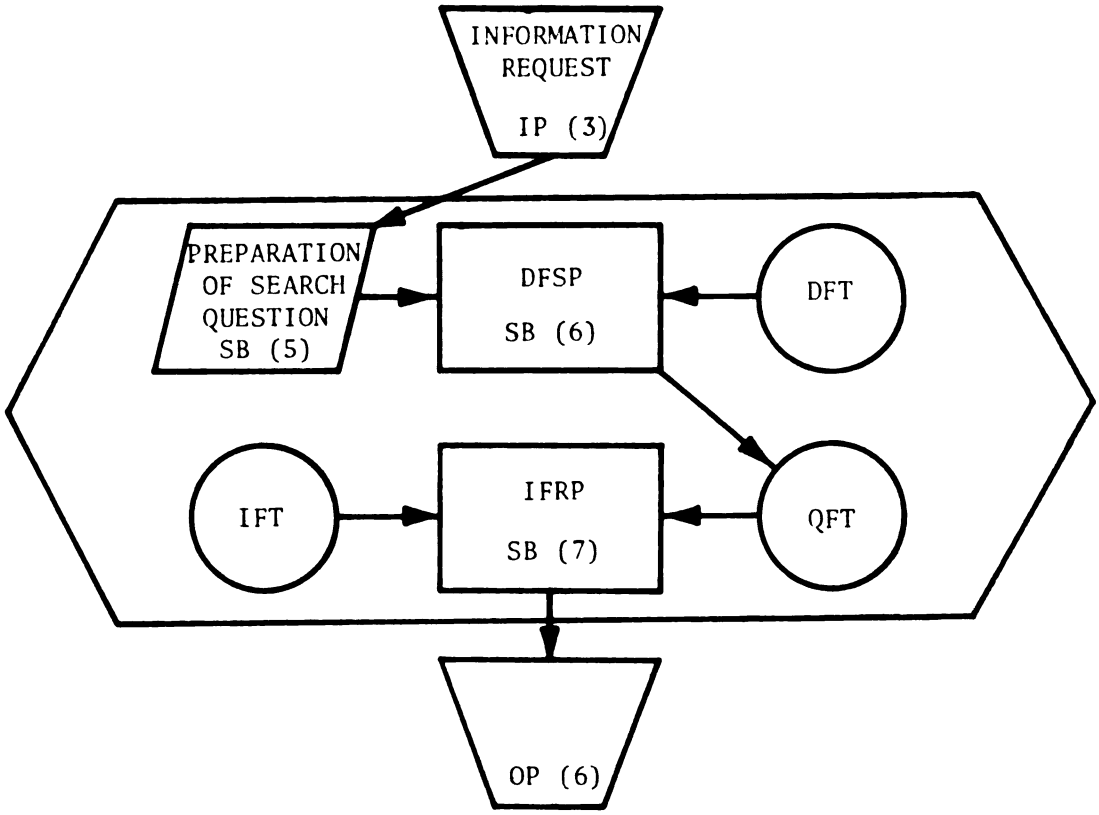


FIGURE 9A

A COMPUTER SEARCH - PREDEFINED PROCESS 9

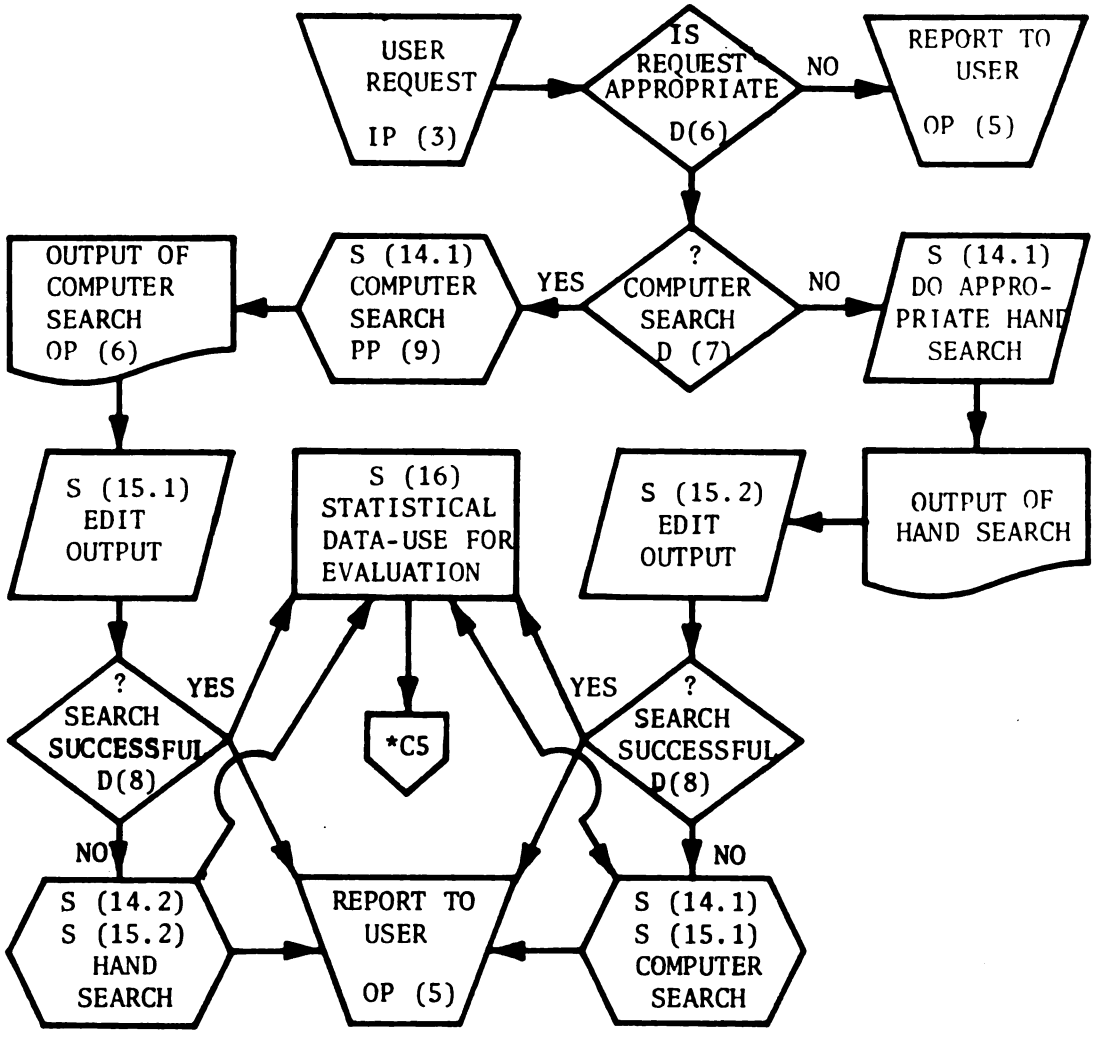
able as a part of DFSP can be obtained from the BIRS documentation.[4A]

The overall procedures used in answering an information request are outlined in Figure 10A. When a request is received IP (3) it is examined to determine if it is relevant to the information stored at the Center (D6) and if not a letter is sent to the person requesting information stating why the Center cannot process the request. If the request concerns the information contained at the Center, the decision (D7) is made whether it can be best processed by a computer search or by a hand search.

If a computer search is indicated it follows the procedures indicated in PP (9). (See Figure 9A for detail.) The output from the computer search OP (6) is edited S (15.1) to determine if it is meaningful and depending on the nature of the request, items that are inappropriate may be removed. If after editing it was determined that the search was successful a report of the search results is sent to the user, and statistical information about (1) the type of question, (2) the type of individual requesting information, and (3) the type of information sent is transmitted for processing S (16). This information is used to assist in providing quarterly reports to ERIC and in determining how users may be served better.

If Decision 7 indicated that a hand search was best, it is determined if there are obvious documents available that can be sent, if a hand search of ECEA indexes is most appropriate, or if there is a selected bibliography with an index reference that could be sent (Step 14.2).

[4A]John F. Vinsonhaler, John M. Hafterson, Stuart W. Thomas, Jr. (editors), Basic Information Retrieval System Technical Manual (East Lansing, Michigan: Information Systems Laboratory, College of Education Michigan State University, 1970), I-XII.



*C5 also connects to evaluation procedures PP (5) of Figures 5A and 6A. The connector is not shown at the figures.

FIGURE 10A

INFORMATION REQUEST PROCESSING

The information gathered from the hand search is edited to determine which information is most appropriate S (15.2) and a decision is made D (8) as to whether or not the search was successful. If it is determined that the search was successful, the report and statistical data are processed in the same manner as in the computer search.

If it is indicated at D (8) in either a hand search or a computer search that there is not sufficient relevant data to warrant sending a report to the user, an alternate search method is attempted--a hand search if a computer search was first done, or a computer search if a hand search was first attempted. The results of the alternative search methods are edited, the statistical information processed and a report is sent to the user. If the second search was no more successful than the first, the report to the user may not contain documents but merely state that the Center was unable to find information relevant to the user's question.

The statistical data collected and processed as part of Step 16 is used by staff members to help determine what types of new documents would be most valuable to users and, if appropriate, these are developed by Center staff or commissioned to experts outside the Center.

Selective Publication

The manner in which the files are prepared for the CEC-ERIC Information Center not only makes it possible to create new subfiles, but to publish these subfiles. Thus if there are a number of requests that could be answered by using the same documents it is possible to directly publish these documents using computer typesetting and a very inexpensive offset process. As of August 1971 the Council for Exceptional Children has 59 separate bibliographies which have been published in this manner.

Step 17 - Selection of Topical Subfiles Figure 11A illustrates the procedure used for creating the selected bibliographies. In Step (17.1) statistical information which has been gathered from user information requests and processed in Step (16) of Figure 10A is analyzed to determine what bibliographies will be of greatest use. Step 17.2 involves a computerized search which generates a list of access numbers that are used to locate abstracts in ECEA so they may be examined (Step 17.3) to determine their relevance to the selected topic.

All documents that are relevant to the type of information desired are selected for the Special Information File Tape (SIFT) generated in A (17.4) and the Special Printed Index File Tape (SPIFT) generated in S (17.5). As the Center's holdings increase the search question is rerun on the new holdings and the edited results added to the SIFT and the SPIFT. These two tapes are used as the basis for continued updating of the printed bibliography in this topical area.

Step 18 - Selection of Abstracts to be Printed While it is possible to publish all of the relevant document surrogates on a special topic, as the file grows larger, even the subfiles contain more documents than can be published and distributed at a reasonable cost. For this reason it has been determined that bibliographies which are given away as answers to requests for information will contain no more than 100 abstracts. The criteria by which 100 or less abstracts are selected from sometimes a thousand or more relevant abstracts are:

1. Availability
2. Recency
3. Information value

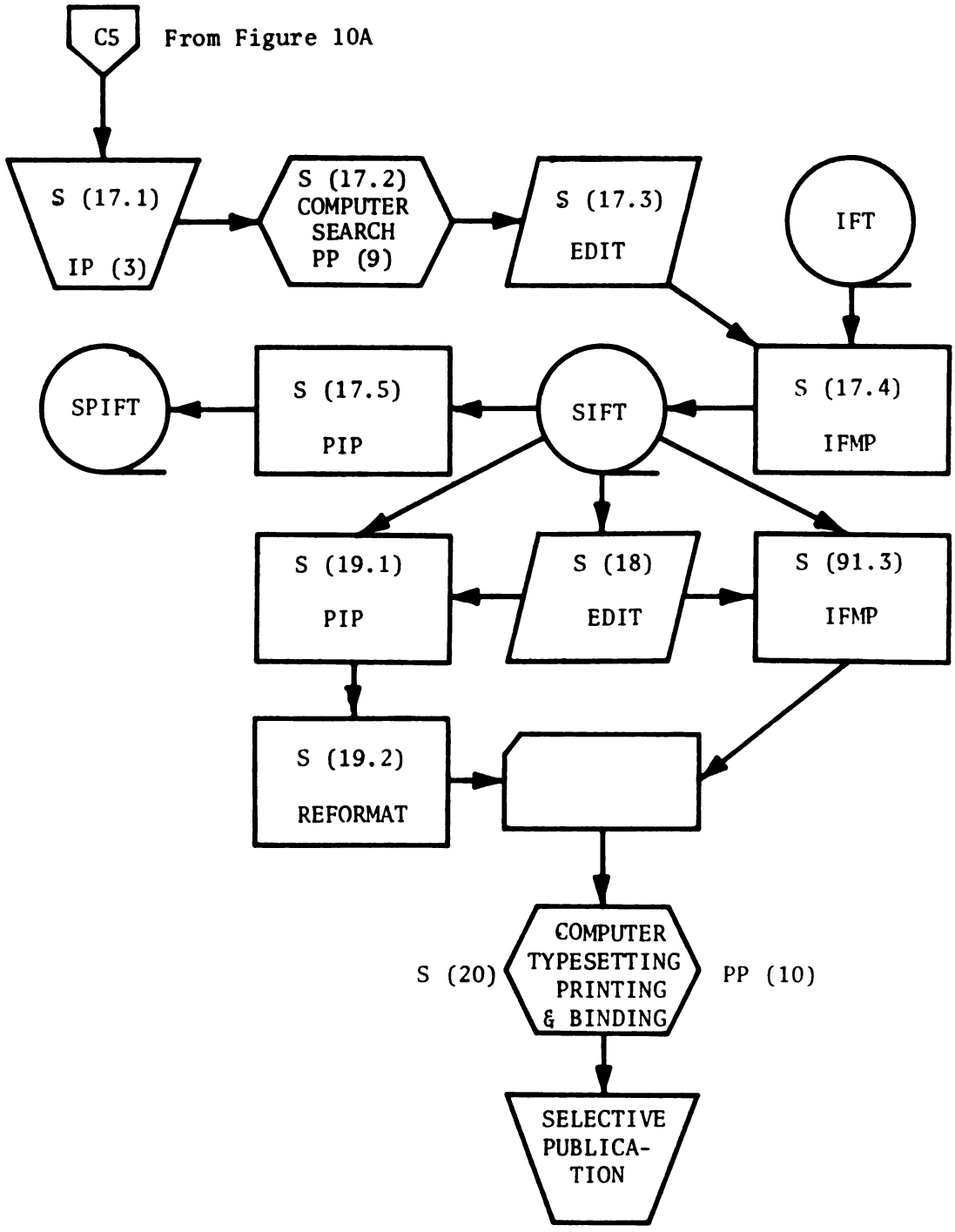


FIGURE 11A

PROCEDURES FOR PROCESSING SELECTIVE PUBLICATIONS

4. Author's reputation

5. Classical content[5A]

Step 19 - Preparation of Input for Computer Typesetting In Step (19.1) the abstracts selected in S (18) are indexed by EC numbers; in Step S (19.2) cards of the index are punched for use in phototypesetting, and in Step S (19.3) cards of the selected abstracts are punched for use in phototypesetting.

Step 20 - Computer-Controlled Typesetting Step 20 consists of computer-controlled phototypesetting, the preparation of offset plates, printing, and binding. This step is indicated as PP (10) and is almost identical to the procedures used for the printing of ECEA.

The bibliographies and their indexes provide a powerful tool that is used in answering about 55% of all information requests. By having a selected topic with an index to that topic it is often possible to answer search requests by using a single item in the index of a special bibliography. The bibliography can be sent with the particular index term or terms circled and a covering letter indicating that the user should look at the abstracts specified by the circled indexing terms.

If the 100 abstracts in the printed bibliography contain too few to answer the specific search request, the special printed index generated in S (17.5) can be used by the person answering user requests to find additional items from the total subfile. Often when using the index of a selected bibliography an individual can obtain in less time the same results as a computer search.

[5A]CEC Information Center, Educational Resources Information Center, Newsletter, June 23, 1971, p. 2.