

UNIV. OF MD. COLLEGE PARK



3 1430 01957098 7

ASLIB CRANFIELD RESEARCH PROJECT

REPORT ON THE TESTING AND ANALYSIS OF AN
INVESTIGATION INTO THE COMPARATIVE
EFFICIENCY OF INDEXING SYSTEMS

by
Cyril W. Cleverdon

An investigation supported by a grant from
The National Science Foundation
Washington

THOMAS A. SEBEOK
Roy/House, Indiana University,
Bloomington, Indiana

ASLIB CRANFIELD RESEARCH PROJECT

REPORT ON THE TESTING AND ANALYSIS OF
AN INVESTIGATION INTO THE COMPARATIVE
EFFICIENCY OF INDEXING SYSTEMS

- by -

Cyril W. Cleverdon

An investigation supported by a grant to ASLIB
by the National Science Foundation

Lib Sch

Z

645.92

.A8

Cranfield

October, 1962

3619889

PREFACE

This volume continues the account of the Aslib-Cranfield project as given in the "Final Report of the First Stage of an Investigation into the Comparative Efficiency of Indexing Systems". The major portion of the two years spent on this present stage has been involved with the analysis of the considerable amount of data which was obtained from the main test programme. A difficulty in this work was in deciding on the type of analysis which would be most likely to yield valuable information. In order to keep this volume within reasonable limits, it has been necessary to select from the analysis that was done, and even so in many cases only brief examples are given. The major emphasis has been placed on the reasons for failure to retrieve source documents, for this is considered to give some of the most interesting results of the project and has not, to our knowledge, been previously attempted. Of possible equal importance, but certainly more difficult to evaluate, is the reason for the retrieval of non-relevant references. This analysis has not been attempted within the present work, but will be one of the matters to be investigated in the continuation of the project.

The interest in this work has been widespread, and a large number of people either personally or in correspondence, have made many interesting and useful comments. To all these, as well as many others who have taken an active part in the work, I would acknowledge my debt. In particular, however, I would express my sincere thanks to the National Science Foundation for their support which alone made the project possible, and in particular to Dr. B. Adkinson and Mrs. Helen Brownson for their co-operation, advice and encouragement. I am also indebted to the Principal and Senate of The College of Aeronautics for their permission for the work to be undertaken at Cranfield and to Mr. L. Wilson, Director of Aslib, for coping so admirably with the administration of the project.

Cranfield, September 1962.

Cyril W. Cleverdon.

INDEX

	<u>Page</u>
Preface	
Chapter 1. Introduction	1
Chapter 2. Main Test Programme	7
Chapter 3. Results of Main Test	21
Chapter 4. Statistical Analysis	27
Chapter 5. Analysis of Failures	38
Chapter 6. Supplementary Test Programmes	51
Chapter 7. Testing of Existing Systems	61
Chapter 8. Supplementary Indexing	76
Chapter 9. Comments on the Results	82
Chapter 10. Basic Problems of Information Retrieval	95
References	107
Appendix 2A Part 1. Document Group 14001-14100	108
Part 2. Selection of documents used for compilation of search questions	114
Part 3. Questions resulting from selection	116
Part 4. Extract of letter sent to compilers of questions	117
Appendix 2B Programme Rules	118
Appendix 3A Statistical Analysis of Results of First Two Rounds of Aslib Cranfield Research Project	119
Appendix 4A 100 Questions based on Source Documents Listed in Appendix 4B	130
Appendix 4B Titles of Source Documents for Questions Listed in Appendix 4A	135
Appendix 4C Postings and Results for Documents P12001-P14000	141
Appendix 4D Correlation of Terms used in 300 Searches for Indexing and Searching	143
Appendix 4E Results for Documents P12001-P14000 Showing Number of Searches Required	174
Appendix 5A Analysis of Failures	176
Appendix 5B Examples of Analysis of Failures	217
Appendix 6A Letter to those Assisting in Compilation of Bibliographies	250
Appendix 7A Analysis of Results of Test on English Electric Facet Catalogue	252
Appendix 7B Analysis of Failures for Facet Index in W. R. U. Test	278
Appendix 8A Instructions for Supplementary Indexers	295
Figures 1 - 4	

LIST OF TABLES

		<u>Page</u>
1.1.	Number of Headings or Notational Elements used during the Indexing Programme	3
1.2.	Total Postings for different Time Allowances during Final Sub-Programme	4
1.3.	Average Postings per Document for Reports and Journal Articles	5
1.4.	Average Postings by Series and Journals	6
3.1.	Combined Results of all Searches by Project Staff (s.e. 2.6% approx.)	22
3.2.	Retrieval for Various Indexing Times for Searches by Project Staff in Final Sub-Programme (Documents 12001 - 18000) (s.e. 5% - 6%)	22
3.3.	Percentage Retrieval According to Indexer for Searches by Project Staff in Final Sub-Programme (s.e. 4% approx.)	23
3.4.	Percentage Retrieval According to Subject for Searches by Project Staff in Final Sub-Programme (s.e. 3.5% approx.)	23
3.5.	Percentage Retrieval According to Indexing Sub-Programme for Searches by College Staff (s.e. 3% - 8%)	24
3.6.	Percentage Retrieval for Searches by Project Staff in the Three Rounds of Testing (s.e. 5% approx.)	24
3.7.	Percentage Retrieval by Searching for Project Staff in First two Rounds of Testing (s.e. 4% - 6%)	25
3.8.	Results of all Searches by Technical Staff (s.e. 3.2% - 4.1%)	25
3.9.	Percentage Retrieval for Various Indexing Times for Searches by Technical Staff in Final Sub-Programme (s.e. 6.2% - 11.6%)	26
3.10.	Percentage Retrieval for Indexers for Searches by Technical Staff in Final Sub-Programme (s.e. 5.1% - 9%)	26
3.11.	Percentage Retrieval According to Subject for all Searches by Technical Staff (s.e. 4.3% - 5.4%)	26

List of Tables (Continued)

	<u>Page</u>
4.1. Results of Searches Grouped According to Correlation of Language of Title and Question	28
4.2. Percentage Success Rate for Searches with Varying Degrees of Correlation between Title and Question	31
4.3. Percentage Retrieval Efficiency for Searches by Project Staff after Eliminating Questions where Source Documents were Retrieved by all Systems	32
4.4. Percentage Retrieval Efficiency by Indexing Times for Searches by Project Staff after Eliminating Questions where Source Documents were Retrieved by all systems.	33
4.5. Comparison of Average Postings by Indexing Times for Documents P12001 to P14000 for Successful and Failed Searches by Project Staff	34
4.6. Use of Terms in Indexing 200 Source Documents and Making 200 Searches	35
4.7. Possibly Redundant Terms Showing Number of Times used in Indexing and Searching	35
4.8. Percentage of Number of Searches Required in Second Round of Testing	36
4.9. Percentage of Number of Searches Required in Third Round of Testing	37
5.1. Reasons for Failures in Second Round of Tests by Project Staff	39
5.2. Reasons for Failure in Third Round of Tests by Project Staff	40
5.3. Total Reasons for Failures in Second and Third Rounds of Tests by Project Staff	41
5.4. Reasons for Failures (Project Staff)	49
5.5. Breakdown of Reasons for Failures (Project Staff)	49
5.6. Reasons for Failures (Technical Staff)	50
5.7. Breakdown of Reasons for Failures (Technical Staff)	50
6.1. Results of Searches for Relevant Documents from Bibliographies	53

List of Tables (Continued)

	<u>Page</u>
6.2. Summary of Results of Searches for Relevant Documents from Bibliographies	54
6.3. Total Documents Retrieved in Sample of 79 Searches	55
6.4. Documents Retrieved in Sample of 67 Searches	56
6.5. Relevance Ratio of Documents Retrieved in 79 Searches	56
6.6. Non-Source Relevant Documents Retrieved in 79 Searches	58
7.1. Search Results in English Electric Catalogue	63
7.2. Reasons for Failures in English Electric Catalogue	64
7.3. Results of searches in W. R. U. and Cranfield indexes	69
7.4. Recall as affected by number of entries	70
7.5. Recall and Relevance Ratio in Facet Catalogue of W. R. U. Test at Varying Index Entries	72
7.6. Performance of Facet Catalogue of W. R. U. Test	72
7.7. Reasons for Failure in Facet Catalogue of W. R. U. Test	73
8.1. Data on Supplementary Indexing	76
8.2. Comparison of Results of Project Staff and Supplementary Indexers	78
8.3. Supplementary Indexing Results by Country	78
8.4. Supplementary Indexing Results of Individual Organisations	79
8.5. Analysis of Reasons for Failures with Supplementary Indexing	80
8.6. Reasons for Failures in Supplementary Indexing	81
8.7. Efficiency at Varying Indexing Times for Supplementary Indexing	81
9.1. Efficiency by Time for all Systems	83
9.2. Breakdown of Reasons for Failure on Documents Indexed in First Sub-Programme (Documents 1 - 6000)	86

List of Tables (Continued)

	<u>Page</u>
9.3. Breakdown of Reasons for Failure on Documents Indexed in Second Sub-Programme (Documents 6001 - 12000)	86
9.4. Probable Operating Area of I. R. Systems	89
10.1. Translation of Concepts into Descriptor Languages	97
10.2. Question Generality Figures	101
10.3. Hypothetical Performance of Four Descriptor Languages Operating at Different Levels	105

