UNITED STATES DEPARTMENT OF COMMERCE • Maurice H. Stans, *Secretary*

NATIONAL BUREAU OF STANDARDS • Lewis M. Branscomb, *Director*

Automatic Indexing: A State-of-the-Art Report

Mary Elizabeth Stevens

Center for Computer Sciences and Technology
National Bureau of Standards
Washington, D. C. 20234

National Bureau of Standards Monograph 91

Issued March 30, 1965

Reissued with Additions and Corrections (See Preface), February 1970

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D. C. 20402 (Order by SD Catalog No. C13.44:91), Price $2.25

Foreword
(1970 Edition)

Widespread interest in the use of computers in automatic indexing created a demand for this publication that led to a recent exhaustion of all stock. While updating and revision would have been desirable, other demands have prescribed reissuance with additional material added as appendices. These are a paper updating the field through September 1966 (Appendix B), and bibliographic citations, pertinent to the subjects in the original text, through August 1969 (Appendix C).

Lewis M. Branscomb
Director

Library of Congress Catalog Card Number: 65-60023

Foreword
(1965 Edition)

The Research Information Center and Advisory Service on Information Processing, (RICASIP), which is jointly supported by the National Science Foundation and the National Bureau of Standards, is engaged in a continuing program to collect information and maintain current awareness about research and development activities in the field of information processing and retrieval. An important responsibility of RICASIP is the preparation of state-of-the-art reviews on topics of current interest in various areas of this broad field.

This report is one of a series intended as contributions toward improved interchange of information among those engaged in research and development in this field. The report considers new uses of machines and automatic data processing procedures for the compilation and generation of indexes to the scientific and technical literature.

A.V. Astin, Director

Contents

Page

Abstract 1

1. Introduction 1

1.1 Definitions and background 2
1.2 Scope of this study 10
1.3 Derivative vs. assignment indexing 13

2. Indexes compiled by machine 14

2.1 Concordances and complete text processing 15
2.2 Card catalogs, book catalogs, bibliographies and subject index listings prepared by machine 19
2.3 Tabledex and other special purpose indexes 25
2.4 Citation indexes 27
2.5 Machine conversion from one index set to another 38

3. Indexes generated by machine - automatic derivative indexing 40

3.1 KWIC indexes 40

3.1.1 Applications of KWIC indexing techniques 41
3.1.2 Advantages, disadvantages and operational problems of KWIC indexing 55

3.2 Modified derivative indexing 68

3.2.1 Title augmentation 68
3.2.2 Book indexing by computer 71
3.2.3 Modified derivative indexing - Baxendale's experiments 73

3.3 Derivative indexing from automatic abstracting techniques 75

3.3.1 Auto-condensation and auto-encoding techniques of H. P. Luhn 75
3.3.2 Frequencies of word n-tuples - Oswald and others 79
3.3.3 Relative frequency techniques - Edmundson and Wyllys, and others 81
3.3.4 Significant word distances 83
3.3.5 Uses of special clues for selection 84
3.3.6 Recent examples of mixed systems experimentation 86

3.4 Quality of modified derivative indexing by machine 89

4. Automatic assignment indexing techniques 91

4.1 Swanson and later work at Thompson Ramo Wooldridge 91
4.2 Maron's automatic indexing experiments 93
4.3 Automatic indexing investigations of Borko and Bernick 94
4.4 Williams' discriminant analysis method 97
4.5 SADSACT 98
4.6 Assignment indexing from citation data 99
4.7 Similarities and distinctions among assignment indexing experiments 100
4.8 Other assignment indexing proposals 105

Page

5. Automatic classification and categorization 106

5.1 Factor analysis 108
5.2 The theory of clumps 110
5.3 Latent class analysis 113
5.4 Examples of other proposed classificatory techniques 113

6. Other potentially related research 114

6.1 Thesaurus construction, use and up-dating 114
6.2 Statistical association techniques 118

6.2.1 Devices to display associations: EDIAC 119
6.2.2 Statistical association factors - Stiles 119
6.2.3 The association map - Doyle and related work at SDC 122
6.2.4 Work of Giuliano and associates, the ACORN devices 124
6.2.5 Spiegel and others at Mitre Corporation 126

6.3 Clues to index-term selection from automatic syntactic analysis 127
6.4 Probabilistic indexing and natural language text searching 132

6.4.1 Probabilistic indexing - Maron, Kuhns and Ray 133
6.4.2 Natural language text searching - Swanson 134
6.4.3 Full text searching - legal literature 135

6.5 Other examples of related research in linguistic data processing 136
6.6 Machine assistance in translations of subject content indications to special search and retrieval language 140
6.7 Example of a proposed indexing-system utilizing related research techniques 142

7. Problems of evaluation 143

7.1 Core problems 145
7.2 Bases and criteria for evaluation of automatic indexing procedures 149

7.2.1 The Cranfield project 150
7.2.2 O'Connor investigations 151
7.2.3 Questions of comparative costs 153
7.2.4 Summary: potential advantages as bases for evaluation 156

7.3 Findings with respect to inter-indexer and intra-indexer consistency 157
7.4 Special factors and other suggested bases for evaluation 160

8. Operational considerations 164

8.1 Questions of input 164
8.2 Examples of processing considerations 168
8.3 Output considerations 171

9. Conclusion: Appraisal of the state of the art in automatic indexing 173

	Page
Acknowledgments	182
Appendix A: List of references cited and selected bibliography	183
Appendix B: Progress and prospects in mechanized indexing	223
Appendix C: Selective bibliography of additional references	237