

T H E C O M P U T A T I O N L A B O R A T O R Y

Harvard University
Cambridge, Massachusetts

Scientific Report No. ISR-7
INFORMATION STORAGE AND RETRIEVAL
to
The National Science Foundation

Cambridge, Massachusetts
June 1964

Gerard Salton
Project Director



Copyright, 1964

By the President and Fellows of Harvard College

Use, reproduction, or publication, in whole or in part, is permitted for any purpose of the United States Government.

STAFF OF THE COMPUTATION LABORATORY AND OF THE COMPUTING CENTER

Lynda Addison
Deborah D. Aulenback
Thomas C. Bartee
Marie V. Bennett
William Bossert
Benjamin Bossin
Walter Broderick
Robert J. Burns
Karen Brassil
Mark Cane
Irene Collins
Evelyn J. Cone
Paul Conway
Isabel Corbató
Jean M. D'Agostino
Richard Delery
Arthur F. Dolan
James Donahue
David Drew
Judith C. Eckian
Frank Engel, Jr.
Tom Evslin
Patrick Fischer
Annette Fluendy
Leonard Gaetano
Salvador Gallegos
Robert Ginelli
Sheila Greibach
Claudine Harris
Inez Hazel
Eric Johansson
Elena Kirsch
Susumu Kuno
Beverley Lee
Alan Lemmon

A. Richard LeSchack
Michael Lesk
Thomas L. Lewis
Joseph I. Lewko
Irina Lynch
Mary Lynch
Edward Lyons
Rita Mahony
Michael McAnulty
Pernell McFarlane
Veronica E. McLoud
George McManus
Jeri Mignault
Sharon O'Brien
Anthony G. Oettinger
Arthur O'Neill
Marguerite E. Pandt
Richard Perault
Richard Pizzano
Warren Plath
Richard Powers
Diane L. Redonnet
Joseph Rocchio
Gerard Salton
Jacquelin Sanborn
George Shapiro
Carol Smith
Valerie Smith
Edward H. Sussenguth, Jr.
Rodney Thorpe
C. Cynthia Tukis
John T. van Bemmell
Hao Wang
Richard Whalen
Stephen P. Young

TABLE OF CONTENTS

	<u>page</u>
SUMMARY	xiii

SECTION I

SALTON, GERARD: "The SMART System — An Introduction"

1. Introduction	I-1
2. Characteristics of the SMART System	I-2
3. SMART Documentation and Principal Systems Features	I-5

SECTION II

LESK, MICHAEL: "The SMART System — General Program Description"

1. Introduction	II-1
2. The SMART Library Tape	II-5
3. The CHIEF Supervisor	II-9
A. PHASE I: Initialization	II-10
B. PHASE II: Document Absorbtion	II-12
C. PHASE III: Request Answering	II-20
4. Text Read In and Lookup	II-22
5. Operational Details	II-30

TABLE OF CONTENTS (continued)

page

SECTION III

HARRIS, CLAUDINE: "Dictionary and Hierarchy Construction"

1. The Document Collection	III-1
2. Nature of the Language Data	III-2
3. The Dictionaries	III-2
A. Stem and Suffix Dictionaries	III-2
B. The Semantic Concept Hierarchy	III-5
C. Phrase Dictionaries -- Criterion Trees and Word Pairs	III-7
D. The Vacuous Dictionary	III-15

SECTION IV

CANE, MARK: "Dictionary Lookup and Updating Procedures"

1. Introduction	IV-1
2. Input Deck, Control Card, and Data Card Formats	IV-2
3. Implementation of Setup and Updating	IV-4
4. The Tree Programs -- SUFTR, SUFAD, TREET, and TRADD	IV-7

TABLE OF CONTENTS (continued)

page

SECTION IV (continued)

5. General Description of Thesaurus Lookup	IV-16
6. Implementation of Lookup	IV-18
7. Spelling Rules Incorporated into Lookup	IV-21
8. Processing of Words Not Found by the Thesaurus Lookup	IV-24
Appendix	IV-28

SECTION V

SHAPIRO, GEORGE: "Processing of the Concept Hierarchy"

1. Introduction	V-1
2. Structure in Core	V-2
3. Instructions for Using Subroutine GEORGE	V-6
4. Hierarchy Setup and Update Processing Examples	V-9
5. Description of Programs for Setting Up and Updating the Hierarchy	V-12
6. The Hierarchy Lookup Program	V-16
Appendix A	V-19

TABLE OF CONTENTS (continued)

page

SECTION VI

LEMMON, ALAN: "Syntax and Criterion Procedures"

1. Introduction	VI-1
2. The Input Editing Problem	VI-2
3. Internal Processing — The BTOKUN Routine	VI-7
4. Organization and Operation of the Criterion Routine (CRITER)	VI-11
5. Internal Processing of the Criterion Routine	VI-23

SECTION VII

SUSSENGUTH, EDWARD H. Jr: "The Sentence Matching Program — GRAPH"

1. Introduction	VII-1
2. The GRAPH Structure-Matching Program	VII-6
Appendix A. Programming Details for GRAPH	VII-10
Appendix B. Subroutines Used by GRAPH	VII-12

TABLE OF CONTENTS (continued)

page

SECTION VIII

EVSLIN, TOM and LEWIS, THOMAS: "Setup and Updating of the Criterion Tree File"

1. Introduction	VIII-1
2. Updating of the Criterion Tree File	VIII-1
3. Construction of New Criterion Trees	VIII-6
A. Format of the Input Cards	VIII-6
B. Format of the Criterion Tree File	VIII-9

SECTION IX

LESK, MICHAEL and EVSLIN, TOM: "Statistical Phrase Processing"

1. Introduction	IX-1
2. Updating of the Statistical Phrase File	IX-1
3. Statistical Phrase Counting	IX-6
4. Output of Phrase Counts	IX-10

TABLE OF CONTENTS (continued)

page

SECTION X

LESK, MICHAEL: "Procedures for Statistical Processing and
Request Alteration"

1. Introduction	X-1
2. Statistical Programs	X-2
3. Request Processing	X-10
Appendix	X-17

SECTION XI

LESK, MICHAEL and EVSLIN, TOM: "Housekeeping Routines"

1. SMART System Routines	XI-1
2. Memory Clean-up Procedures	XI-4
3. Formation of the Vacuous Thesaurus	XI-6

SECTION XII

ROCCHIO, JOSEPH J: "Possible Time-sharing Organization For a
SMART Retrieval System"

1. Introduction	XII-1
---------------------------	-------

TABLE OF CONTENTS (continued)

page

SECTION XII (continued)

2. Time-sharing and Retrieval Systems	XII-1
3. Operational Aspects	XII-3
4. Implementation	XII-6
Appendix A.	XII-13

SECTION XIII

LeSCHACK, A. RICHARD: "A Note on Measures of Similarity"

1. Introduction	XIII-1
2. Measures of Similarity — General Properties and Definitions	XIII-2
3. Relation Between the Cosine and Correlation Measures .	XIII-6

SECTION XIV

LeSCHACK, A. RICHARD: "The Determination of Clusters by
Matrix Analysis"

Abstract	XIV-1
1. Introduction	XIV-1

TABLE OF CONTENTS (continued)

page

SECTION XIV (continued)

2. A Model for Strong Clustering	XIV-3
3. Spectral Analysis and Factor Analysis	XIV-10
4. Organization of Programs for Finding Eigenvalues and Eigenvectors	XIV-17
5. Program Test	XIV-19
6. Spectral Analysis to Determine Clusters	XIV-20
A. The 13-variable Case (Psychological Tests) . . .	XIV-21
B. The 25-variable Case (Word Correlations) . . .	XIV-25
7. Evaluation and Conclusions	XIV-37