

07 08

**THE COMPUTATION LABORATORY
OF
HARVARD UNIVERSITY**



Forskningsgruppen
KVAL

INFORMATION STORAGE AND RETRIEVAL

Scientific Report No. ISR-8

to

The National Science Foundation

**Cambridge, Massachusetts
December 1964**

**Gerard Salton
Project Director**

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-8
INFORMATION STORAGE AND RETRIEVAL
to
The National Science Foundation

Cambridge, Massachusetts
December 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
Donald Anderson
Lois Arkin
Thomas C. Bartee
Lynn Barth
Marie V. Bennett
William Bossert
Benjamin A. Bossin
Karen A. Brassil
Robert J. Burns
Mark Cane
Howard F. Coffin
Irene R. Collins
Evelyn J. Cone
Paul M. Conway
Isabel B. Corbató
Jean M. D'Agostino
Richard L. Delery
Arthur F. Dolan
James G. Donahue
Eva Doty
David E. Drew
Judith C. Eckian
Margaret Engel
Tom Evslin
Patrick C. Fischer
Leonard Gaetano
Robert Ginelli
Sheila A. Greibach
Claudine Harris
Inez B. Hazel
Elena Kirsch
Susumu Kuno
Marguerite E. Lemieux
Alan Lemmon

Michael Lesk
Thomas L. Lewis
Joseph I. Lewko
Irina Lynch
Mary Lynch
Edward L. Lyons
Michael A. McAnulty
Margaret McLean
Veronica E. McCloud
George McManus
Rita B. Mahony
Jeri L. Mignault
Sheila Nathanson
Anthony Oettinger
Richard L. Perault
Richard C. Pizzano
Richard J. Powers
Antonio Querido
Diane L. Redonnet
Joseph Rocchio
Gerard Salton
Jacquelin Sanborn Sill
George Shapiro
Carol Smith
Dorothy Thomson
Rodney W. Thorpe
Cynthia C. Tukis
John T. Van Bemmell
Laura Vermilyea
Hamilton Vreeland (III)
Hao Wang
Richard F. Whalen, Jr.
Thomas Wooten
Stephen P. Young
Norman Zachary

TABLE OF CONTENTS

	<u>page</u>
SUMMARY	xi

SECTION I

LESK, MICHAEL: "The SMART System -- Typical Processing Sequences"

Appendix	I-17
--------------------	------

SECTION II

LESK, MICHAEL: "The SMART Automatic Text Processing and Document Retrieval System"

1. Introduction	II-1
2. Machine Configuration	II-2
A. Operating System	II-2
B. Magnetic Tape Assignments	II-2
C. Printed Output	II-3
3. Program Loading	II-4
A. Chain Structure of SMART	II-4
B. SMART Program/Data Tape	II-5
C. Card Decks for SMART	II-6
4. Updating Program and Library Tapes	II-8
A. Updating the Program/Data Tape on B3	II-8
B. Updating the Library Tape on B5	II-10

TABLE OF CONTENTS (continued)

SECTION II (continued)

	<u>page</u>
5. Processing Options	II-31
A. Specifications	II-32
B. Names Table	II-41
C. Correlation Algorithms	II-42
6. Data	II-43
A. Control Cards which Introduce Documents	II-44
B. Control Cards which do <u>Not</u> Introduce Documents	II-47
C. Evaluation	II-48
8. Miscellaneous	II-54
A. Size Limits	II-54
B. Timing	II-54
C. THES	II-55
Appendix 1	II-56

SECTION III

ROCCHIO, JOSEPH: "Performance Indices for Document Retrieval Systems"

Summary	III-1
1. The Model	III-1
2. Evaluation Indices	III-5
3. Experimental Use	III-16

SECTION IV

	<u>page</u>
SALTON, GERARD: "The Evaluation of Automatic Retrieval Procedures -- Selected Test Results Using the SMART System"	
1. Introduction	IV-1
2. The SMART Retrieval System	IV-3
3. The Test Environment	IV-5
4. Evaluation Measures	IV-12
A. Recall and Precision	IV-12
B. The Generation of Relevance Judgments	IV-14
C. The Cut-off Problem	IV-15
D. Normalized Recall and Normalized Precision	IV-17
5. Test Results	IV-21
A. Output Formats	IV-21
B. Results Derived from the Normalized Measures	IV-24
C. Results Using the Standard Measures	IV-30
6. Conclusions	IV-34

SECTION V

SALTON, GERARD: "Automatic Phrase Matching"	
1. Introduction	V-1
2. The Content Analysis Problem	V-2

SECTION V (continued)

	<u>page</u>
3. Language Structure	V-4
4. The Processing of Unrestricted Text	V-9
5. Syntactic Phrase Matching	V-16

SECTION VI

LEMMON, ALAN: "A Compact Format for Criterion Tree Specifications"

1. Introduction	VI-1
2. Input Format	VI-4
A. Index Field	VI-4
B. Output Concept Number Field	VI-5
C. Relations Field	VI-5
D. Specifications Field	VI-6
3. Continuation Cards	VI-7
4. Internal Processing	VI-8
5. Error Conditions	VI-9
6. Use of a TRECND Program	VI-10
Appendix 1	VI-16
Appendix 2	VI-18

SECTION VII

	<u>page</u>
HARRIS, CLAUDINE: "Dictionary Construction and Updating"	
1. Introduction	VII-1
2. Formation of the First Dictionaries	VII-2
3. The Null Dictionary	VII-3
4. Updating the Dictionaries	VII-6
A. The Null Alphabetic List	VII-6
B. The Null Frequency List	VII-6
C. "Use" and "Discuss"	VII-9
5. Refining the Numerical Dictionary by Concept Concordance	VII-11
A. Analysis of 101TAG	VII-12
B. "Point" and "Pointed out"	VII-12
C. Analysis of 168HIE	VII-13
D. Analysis of 80MAKE	VII-14
6. The Phrase Dictionaries	VII-15

SECTION VIII

EVSLIN, TOM: "The Extended SMART System"	
1. Introduction	VIII-1
2. Limitations Removed	VIII-1
3. New Options	VIII-2

SECTION VIII (continued)

	<u>page</u>
4. Logical Changes	VIII-3
5. New Support Programs	VIII-5

SECTION IX

CANE, MARK: "Adapting SMART to the M.I.T. Computible
Time-Sharing System"

1. Introduction	IX-1
2. Modifications of the SMART System	IX-2
3. Implementation - Executive Program Description . . .	IX-5
4. Progress Report	IX-10
Appendix	IX-16

SECTION X

ROCCHIO, JOSEPH AND ENGEL, MARGARET: "Test Design and Detailed
Retrieval Results"

Summary	X-1
1. Current Status of the SMART System	X-1
2. Experimental Retrieval Requests	X-5
3. Retrieval Experiments	X-9