

Mechanical Indexing: A Personal Remembrance

Herbert Ohlman

Abstract

Ideas leading to mechanical and automatic indexing go back 150 years. However, not until World War II, when appropriate tools became available, did these ideas become reality. Early products of these efforts were concordances to full texts, indexes to document collections, and auto-abstracts of documents. I was fortunate to be among those who created one of these systems: permutation indexing. My paper details the development of mechanical indexing and related systems during the 1950s.

Hypertext, which links all types of data, is now the dominant information retrieval technique. It is difficult to remember a time when the character repertoire of most computers was limited to all upper-case letters, ten digits, and a few special symbols. However, many researchers realized that these machines were symbol manipulators as well as number crunchers. Specifically, they could be used to process language material, using alphanumeric character sets. Suitable alphanumeric machines, including punched-card (tabulating) machines, punched paper-tape typewriters, and computers became available just before and during World War II. New specialties using these tools were created, notably “information retrieval,” “natural-language processing,” “speech processing,” and “mechanical translation.” My paper relates the early history of a subset of information retrieval, “mechanical indexing.”

My Work in Mechanical Indexing

My first job in information retrieval (1954–55) was as coordinator of technical information for Carrier Corporation, in Syracuse, New York. At Carrier a marginal punched-card system for technical reports was developed. At Battelle Memorial Institute, in Columbus, Ohio (1955–57), I worked on a team of information retrieval

specialists using an effective, but labor-intensive technique called “extracting,” developed by Ben-Ami Lipetz. This technique was applied to the documentation and analysis of the literature of titanium for the U.S. Department of Defense (Gibson & Lipetz, 1956). While at Battelle, my first paper was published in *American Documentation* (the original title of the *Journal of the American Society for Information Science*); the subject was the production of marginal punched cards on accounting (tabulating) machines and was based on work done at Carrier Corporation (Ohlman, 1957a).

In late 1957 I joined the System Development Division (SDD) of the RAND Corporation to work on the documentation of one of the largest postwar computing projects: the SAGE (Semi-Automated Ground Environment) air defense system. RAND spun off this division, which became the nonprofit System Development Corporation (SDC). SDD was located first in Lexington, Massachusetts, close to MIT’s Lincoln Laboratory, where Whirlwind, possibly the earliest real-time computer, was developed. The librarian at Lincoln Lab’s Division 6, Malcolm Ferguson, had installed a peek-a-boo (coordinate) indexing system to retrieve documents. This system used a large card for each index term, in which tiny holes were punched to show that a document used that term. Searches of the collection required superimposing two or more of these term cards over a light source using Boolean logic operations (Figure 1).

Noticing that peek-a-boo cards were sparsely punched, I felt they wasted most of the space (Figure 2). It occurred to me that isolating terms intended by the inherent structure of natural language to be used in context (e.g., words occur in phrases, sentences, or titles) was bound to reduce retrieval relevancy (Ohlman, 1957b).

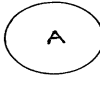
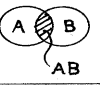
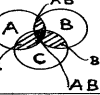
r (no. of coord.) \ n (no. of terms)	1	2	3	4	5	6	Σr
1 	1						1
2 	2	1					3
3 	3	3	1				7
4	4	6	4	1			15
5	5	10	10	5	1		31
6	6	15	20	15	6	1	63
n	${}^n C_1 = n$	${}^n C_2$	${}^n C_3$	${}^n C_4$	${}^n C_5$	${}^n C_6$	$\Sigma {}^n C_r = 2^n - 1$

Figure 1. Mathematically possible combinations of n terms.

To provide contextual retrieval, I devised a new method based on IBM punched cards and tabulating machines. Significant title words were keypunched, one card to a document title. Tabulating machines were “programmed” by Lewis Hart of SDC (actually, this was done by patch cords on control panels) to punch duplicate cards. In each duplicate card, title words were cycled to the left; the expanded deck was then sorted, and this final deck run through a printer to produce the index. I called the result a “permutation index” because the words went through a cyclic permutation process (Ohlman, 1957c).

The first permutation index was issued by SDC in 1957 as a subject guide to SAGE programming documents (Figure 3). It was based on the titles of 1,800 documents, two-thirds of which were from the Lincoln Laboratory (*Permutation Index No. 1*, 1953–mid-1957). A second edition was issued in 1958, which included 4,000 SDC and Lincoln Laboratory documents. In the introduction I suggested strategies to do quick searches, to broaden a search using “connection-of-ideas” (Whorf, 1956), to get an overview of the corpus of documents, and to find gaps in the corpus (*Permutation Index*, mid-1957–mid-1958).

In 1958 the National Science Foundation, the National Academy of Sciences–National Research Council, and the American Documentation Institute sponsored the first International Conference on Scientific

Information. My paper on superimposed coding was accepted for this conference (Ohlman, 1959), and I received preprints of all conference papers.

Here was a perfect way to demonstrate the speed and automation features of permutation indexing to information science and technology colleagues. At SDC, colleagues (Joan Citron and Lewis Hart) and I produced *A Permutation Index to the Preprints of the International Conference on Scientific Information* (Citron, Hart, & Ohlman, 1958). Entries were selected not just from titles; they included author names and affiliations, headings, captions, sentences, and even phrases selected for their significance as thought units. These excerpts provided an average of five permuted entries for every one of the 1,400 preprint pages. Tabulating machines produced the final index but with an improved appearance. Whereas the system used to index internal SDC documents had words truncated to fit into fixed fields, the new index let text flow naturally. The vital alphanumeric character making this possible was the space, which could be used on tabulating machine control panels to determine where to generate additional cards. Also instead of printing the index so that the look-up word appeared at the left margin, it was put in the center to provide context on both sides (Figure 4).

Shortly after the conference, Lewis Hart worked with G. R. Bach to use permutation indexing to analyze verbatim transcripts of psychiatric patients (Hart

Identification Code: *	1	7	6	5	4	3	2	Title starts in column:
5MRN-062A-2		COMPILER	SYSTEM				UNIVAC	2
51BMT0028		ABSTRACT CODING	SYSTEM	FACS		FLOATING	DECIMAL	3
6ACMJ3129		ACADEMY OF SCIENCES	USSR		ELECTRONIC	CALCULATIO	MACHINE	4
5ACMJ2095		ACCEPTANCE TESTS	USED ON	MIDAC			MAINTENANC	2
*Code: 7MITL5		ACCESSIONS LIST	MONTHLY				DIVISION 62	
1st numeral: 7RAND9		ACCESSIONS LIST	WEEKLY		RAND	DOCUMENT	CONTROL	4
year (in '50-59 decade) 7ACMJ4245		ACCOUNT IDENTIFICATION	AUTOMATIC	DATA PROC				1
31BMC-210		ACCOUNTING MACHINE					407	2
21BMC-035		ACCOUNTING MACHINES	24	CARD	PUNCH	26	PRINTING	1
4 letters: 31BMC-019		ACCOUNTING MACHINES	SORTERS	82	80	75		1
journal or organization 61BMC-015		ACCOUNTING MACHINES	CARD	INTERPRETE	550	551	552	1
4 numerals: 51BMT0253		ACCUMULATO	650	MDDPM			INDEXING	2
7ACMJ4005		ADDRESS				INAUGURAL	PRESIDENTI	3
7ACMJ4001		ADDRESS				RETIRING	PRESIDENTI	3
6ACMJ3169		ADDRESS	CALCULATIO				SORTING BY	2
7ACMJ4274		ADDRESS	COMPUTERS		PROGRAMMIN	TECHNIQUES	1-PLUS-1	4
6ACMJ3309		ADVANCED CODING	TECHNIQUE			USING THE	PACT I	3
51BMT0149		AIRCRAFT	650	MAGNETIC	DRUM	DATA PROC	MARQUARDT	6
6ACMJ3299		ALLOCATION DATA	STORAGE	PACT I			SEMIAUTOMA	2
91BMC-039		ALPHABETIC COLLATOR					89	2
41BMC-021		ALPHABETIC INTERPRETE					557	2
6ACMJ3175		ALPHAMERIC INFORMATION	701			GENERAL	SYSTEM FOR	3
5ACMJ2092		AMPLIFIERS				TESTING OF	OPERATIONA	3
6ACMJ3186		AMPLIFIERS	POTENTIOME	TRANSFER	FUNCTION		SIMULATION	5
5ACMJ2267		ANALOG DEVICES				CORRELATIO	COMPUTATIO	3
7ACMJ4012		ANALOG DIGITAL	COMPUTATIO				SYSTEM FOR	2
5ACMJ2083		ANALOG INTERPOLAT	AUTOMATIC	CONTROL				1
5ACMJ2028		ANALOG STUDY OF	ELECTRON	TRAJECTORI				1
6ACMJ3325		ANALOGUE SOLUTION	PARTIAL	DIFFERENTI	EQUATIONS	HIGHER ORD	DIFFERENCES	3
6ACMJ3085		ANALYSIS				BIBLIOGRAP	NUMERICAL	3
61BMT+085		ANALYSIS		ITERATIVE	TECHNIQUE	MULTIPLE	CORRELATIO	5
51BMT0229		ANALYSIS	650	FOURIER	SYNTHESIS	X-RAY	CRYSTAL	7
6ACMJ3360		ANALYSIS	DIGITAL	COMPUTING	NETS		STRUCTURE	7
51BMT0192		ANTENNA PATTERN	CALCULATIO	6D-307			ARITHMETIC	2
7MITL8		APPLICATIO LOGIC						1
7ACMJ4030		APPROXIMAT CONTINUOUS	FUNCTION	CLASS OF	FUNCTIONS		FUNDAMENTA	2
6ACMJ3355		APPROXIMAT CONTOURS					CHEBYCHEV	2
7ACMJ4024		APPROXIMATE EXPONENTIAL	FUNCTIONS				DIGITAL	2
7ACMJ4341		APPROXIMATE OVERDETERM	SYSTEM OF	LINEAR	EQUATIONS	BEST	RATIONAL	2
7ACMJ4018		APPROXIMATE PARABOLIC	EQUATION	STABLE	IMPLICIT	FINITE	LEAST QTH	3
41BMT8067		ARITHMETIC	650				DIFFERENCES	5
41BMT8060		ARITHMETIC	650			DOUBLE	COMPLEX	2
51BMT0108		ARITHMETIC	650		AUTOMATIC	FLOATING	PRECISION	3
							DECIMAL	4

Figure 3. Fixed-field permutation index.

commercial venture called Permutation Indexing, Inc. The first project was PILOT (Permutation Indexing of the Literature of Technology). The inaugural issue was to be published monthly starting in January 1959. Unfortunately, despite almost one hundred subscriptions, the company was undercapitalized and did not survive long enough to distribute the first issue. Other organizations, however, soon brought out publications using KWIC, notably *Chemical Abstracts* and other chemical publications. Also Eugene Garfield at the Institute of Scientific Information developed Permuterm for the automatic indexing of *Current Contents*.

References

Bernier, C. L. (1957). Correlative indexes: I. Alphabetical correlative indexes. *American Documentation*, 7(4), 293-298.

Bibliography and index: Literature on information retrieval and machine translation (1958). New York: The Service Bureau Corporation (This bibliography was based on one by Charles Bourne, Stanford Research Institute).

Borko, H., & Bernier, C. L. (1978). *Indexing concepts and methods*. New York: Academic Press.

Busa, R. (1951). *Varia specimina concordantiarum*. Milano: Fratelli Bocca-Editori.

Carlsen, R. D., Garner, W. H., Jr., & Marshall, H. S. (1958). *Information control*. Canoga Park, CA: Rocketdyne, a Division of North American Aviation.

Citron, J., Hart, L., & Ohlman, H. (1958). *A permutation index to the pre-prints of the International Conference on Scientific Information*. Santa Monica, CA: System Development Corporation. SP-44; Rev. ed., 1959.

Crestadoro, A. (1856). *The art of making catalogues of libraries*. London: The Literary, Scientific & Artistic Reference Office (facsimile reprint OP 14401, Ann Arbor, MI: University Microfilms).

Doyle, L. B. (1966). *The story of language processing and retrieval at SDC*. Santa Monica, CA: System Development Corporation, N-23059.

Doyle, L. B. (1975). *Information retrieval and processing*. Los Angeles: Melville Publishing Company.

Ellison, J. W. (1957). *Nelson's complete concordance to the Revised Standard Version Bible*. New York: Thomas Nelson & Sons.

	INSTIT	PAGE
117 RECOMMENDATION FOR AN INTERNATIONAL	INSTITUTE FOR SCIENTIFIC INFORMATION *	
7122 ANNUAL BUDGET FOR AN INTERNATIONAL	INSTITUTE FOR SCIENTIFIC INFORMATION *	
7119 THE CHARACTER OF THE INTERNATIONAL	INSTITUTE /	
TY , LONDON / 7085 D . J . FOSKETT THE	INSTITUTE OF EDUCATION , LONDON UNIVERS	
ING CHESTER BEATTY RESEARCH INSTITUTE ,	INSTITUTE - 3007 DENNIS A . BRUN	
IO / 1087 GROUP - CASE	INSTITUTE OF TECHNOLOGY , CLEVELAND , O	
ERVICE / 2202	INSTITUTES PHOTO AND MICROFILM COPYING	
67 B M CROWTHER SCIENCE ABSTRACTS , THE	INSTITUTION OF ELECTRICAL -	2
NED FROM THE LITERATURE AND THE TYPE OF	INSTITUTION / 1037 INFORMATION	OBT
THEMATICAL PLAN OF THE ESTABLISHMENT OR	INSTITUTION / 1179 THE	
ON - 1176 APPENDIX 1 ORGANIZATIONS AND	INSTITUTIONS COOPERATING IN THE COLLECT	
1039 LITERATURE - AND	INSTITUTIONAL AFFILIATION /	
2110 APPENDIX 2 DATA COLLECTION	INSTRUCTIONS FOR PROJECT 957 /	
5172 CONCEPTUAL AND	INSTRUMENTAL FORMS /	
/ 3075 MAREK CIGANIK CABLES AND	INSULATING MATERIALS RESEARCH INSTITUTE	
NTATIVE / 2094 RANDOM SAMPLE TO	INSURE THAT PAPERS SELECTED WERE REPRESE	
IDERATIONS IN THE DESIGN - 5003	INTELLECTUAL PROBLEMS AND EQUIPMENT CON	
OF TECHNOLOGISTS FOUND TO BE POSITIVELY	INTERCORRELATED / 1238 CHARACTERISTICS	
2179 THE SECOND REASON , IS VESTED	INTEREST /	
INTERNATIONAL COMPARISON WOULD BE VERY	INTERESTING /	118
	5297 INTERFIX *	
	5259 INTERFIXING /	
	5266 INTERFIXING *	
	5275 INTERFIXING OF SETS /	
ES . 5201	INTERLINGUAL COMMUNICATION IN THE SCIENC	
MENT / 5204	INTERLINGUAL COMMUNICATION DAILY MORE UR	
5107 A SUGGESTED	INTERLINGUAL ANALYSIS OF SYNTAX *	
5091 DOES NOT REQUIRE AN	INTERMEDIATE CODE BOOK /	
INTERESTING /	1187 INTERNATIONAL COMPARISON WOULD BE VERY I	
TRACTING .	2167 INTERNATIONAL COOPERATION IN PHYSICS ABS	
BUILDING , AN APPRAISAL .	2177 INTERNATIONAL COOPERATIVE ABSTRACTING ON	
	3055 INTERNATIONAL AGENCIES *	
OGRAMS *	7032 INTERNATIONAL ORGANIZATIONS AND THEIR PR	
	7061 INTERNATIONAL ORGANIZATIONS /	
	7093 INTERNATIONAL - COOPERATION *	
FORMATION . 7113 AN	INTERNATIONAL INSTITUTE FOR SCIENTIFIC I	
POSAL FOR THE ESTABLISHMENT OF A SINGLE	INTERNATIONAL JOURNAL / 7 94 PRO	
NS - 7093 G . - A . BOUTRY SECRETARY ,	INTERNATIONAL COUNCIL OF SCIENTIFIC UNIO	
CTING BOARD , THE STORY OF A VENTURE IN	INTERNATIONAL - 7093 THE ICSU ABSTRA	
	INSTITUTE /	
FORMATION * 7119 THE CHARACTER OF THE	INTERNATIONAL INSTITUTE FOR SCIENTIFIC I	
MATION . 7122 ANNUAL BUDGET FOR AN	INTERNATIONAL INSTITUTE FOR SCIENTIFIC I	
	7107 CREATION OF AN INTERNATIONAL CENTER OF SCIENTIFIC INFOR	
7005 DIFFERENCES IN NATIONAL AND	INTERNATIONAL ARRANGEMENTS FOR FINANCE /	
FORMATION * 7117 RECOMMENDATION FOR AN	INTERNATIONAL INSTITUTE FOR SCIENTIFIC I	
SUPPORT - 7025 DIFFERENCES IN	INTERNATIONAL ARRANGEMENTS FOR FINANCIAL	

Figure 4. Sample page from "A permutation index to the preprints of the International Conference on Scientific Information."

- Gibson, R. W., Jr., & Lipetz, B. (1956). New look in manual methods. *Special Libraries*, 47, 106–113.
- Hart, L. D., & Bach, G. R. (1959). *Natural language indexing by means of data-processing machines: Observation of the growth of perception protocol*. Paper presented at the Western Psychological Association meeting, San Diego, CA. Santa Monica, CA: System Development Corporation, SP-78.
- Heumann, K. F., & Dale, E. (1957). Statistical survey of chemical structure. In G. L. Peakes, A. Kent, & J. W. Perry (Eds.), *Progress report in chemical literature retrieval* (pp. 201–214). New York: Interscience Publishers, Inc.
- Luhn, H. P. (1959). *Keyword-in-context index for technical literature (KWIC index)*. Yorktown Heights, NY: IBM Advanced System Development Division, RC-127.
- Metcalfe, J. (1957). *Information indexing and subject cataloging: Alphabetical, classified, co-ordinate, mechanical*. New York: Scarecrow Press.
- Netherwood, D. B., (1958). Logical machine design: A selected bibliography. *IRE Transactions on Electronic Computers*, 7(2), 155–178.
- Ohlman, H. (1957a). Low-cost production of marginal punched cards on accounting machines. *American Documentation*, 8(2), 123–126.
- Ohlman, H. (1957b). *Limitations of peek-a-boo and a new retrieval system: Permutation indexing*. Lexington, MA: RAND Corporation System Development Division.
- Ohlman, H. (1957c). *Permutation indexing: Multiple-entry listing by electronic accounting machines*. Lexington, MA: RAND Corporation System Development Division.
- Ohlman, H. (1959). Subject-word letter frequencies with applications to superimposed coding. *Proceedings of the International Conference on Scientific Information*, 2, 16–21. Washington, DC: National Academy of Sciences–National Research Council.
- Ohlman, H. (1960). *Mechanical indexing: Historical development, techniques and critique*. Paper presented at 23rd annual meeting of the American Documentation Institute, Berkeley, CA.
- Permutation index no. 1: A subject guide to SAGE programming documents* (1953–mid-1957). Santa Monica, CA: System Development Corporation.
- Permutation index: A document finder for FN, SD, SP, TM, 66D, 6D & 6M series*. Experimental issue no. 2 (mid-1957–mid-1958). Santa Monica, CA: System Development Corporation, SD-2845.
- Proceedings of the International Conference on Scientific Information* (1958, November). Washington, DC: National Academy of Sciences–National Research Council.
- Shera, J. (1966). *Documentation and the organization of knowledge*. Hamden, CT: Archon Books.
- Veilleux, M. (1962). Permuted title word indexing procedures for a man/machine system. In *Machine indexing: Progress and problems* (Papers presented at the Third Institute on Information Storage and Retrieval, February 13-17, 1961, pp. 77–111). Washington, DC: American University Center for Technology and Administration.
- Whorf, B. L. (1956). On the connection of ideas. In J. B. Carroll (Ed.), *Language, thought and reality: Selected writings of Benjamin Lee Whorf*. Cambridge, MA, and New York: Technology Press of MIT and John Wiley & Sons, Inc.
- Yardley, H. O. (1931). *The American black chamber*. Indianapolis, IN: Bobbs-Merrill Co.