OPTICAL PAGE READING DEVICES

ROBERT A. WILSON
Consultant in Information Retrieval Systems
Dallas, Texas

New York
REINHOLD PUBLISHING CORPORATION
Chapman & Hall, Ltd., London

Library of Congress Catalog Card Number: 66-20810
Printed in the United States of America

Preface

THIS BOOK has two primary aims. The first is to introduce and explain, in a way as non-technical as possible, one of the most complex mechanisms yet devised by man. The second is to suggest some of the ways in which that mechanism, when teamed with the digital computer, can be used to channel, and eventually control, the flood of printed information which now engulfs us.

The most obvious and most immediate function of optical reading machines is to provide a new and very fast way of converting printed pages into computer language, freeing us from the constraints imposed by present keystroking methods.

The breaking of the input bottleneck will make possible a host of new and exciting applications for computers. The writing of programs for engineering, scientific, and mathematical research will be simplified, and business data processing users will find the entering of their data much easier.

But there is much more to the story than this. Disciplines whose literature has heretofore been considered too bulky to be processed by computer because of the expense of keypunching the source documents will be able for the first time to utilize the machines to their benefit, opening up entire new areas of historical, literary and linguistic research. The diffuse materials of economics, political science, sociology, psychology, and the other social sciences can be more effectively brought to bear on the solving of modern problems. The literatures of medicine, the law, engineering, and even science itself can be made more amenable to research than ever before.

What the future holds beyond these, no one can tell. For the

present, it is enough to know that the advent of the optical page-reading machine is here, and that it gives promise of providing us with a tool powerful enough to help in satisfying one of mankind's great basic hungers: *the need to know.*

Robert A. Wilson

Dallas, Texas
April, 1966

Contents

CHAPTER I

Glossary of Terms

"It is one of the maxims . . . that definitions are hazardous."

Samuel Johnson (1709-1784) in *"The Rambler."*

"When I use a word," Humpty Dumpty said, "it means just what I choose it to mean—neither more nor less."

Lewis Carroll (1832-1898) in *"Alice's Adventures in Wonderland,"* Chap. 6.

A WORD ABOUT the following Glossary. First, it is not, and does not purport to be, exhaustive. Secondly, it is intended to be informative rather than definitive, and therefore stresses understanding of general principles rather than familiarity with technical details. Thirdly, it follows standard terminology as far as can be ascertained; there is no intent to create a new vocabulary, or to give exotic new meanings to old words. Fourthly, the reader is reminded that all three of the fields from which the terms are drawn (computer-electronics, documentation, and optical character reading) are relatively new, and their vocabularies are not fully stabilized, so that in some instances there are several terms for the same process (e.g., "mask-matching," "template-correlation," and "image comparison" describe essentially the same operation). Finally, a conscious attempt has been made to avoid jargon; the subject is quite complex enough without complicating it by obscure terminology, of which the following, taken from the recent literature, is a typical example:

"Observable data samples are represented by n-tuples of real numbers, and classes by (possibly disconnected) regions of Euclidean n-

space. Research has prescribed methods of application of the Minimax and Neyman-Pearson Criteria and combinations of the Minimax, Neyman-Pearson, and Bayes Criteria for all multicategory cases. The use of other criteria is being considered along with various methods of separation of categories in machine implementation. Feasibility of analog implementation of the analysis has been shown."

In this connection the author of this book has observed a strange phenomenon in the literature on the subject: those who are actually making optical character readers seem able to describe them in understandable, even though technical, language, while those who are merely writing about them from a theoretical aspect often seem to delight in linguistic obscurity for its own sake.

(*Note:* The capitalized words in the body of the definitions are themselves separately defined in the Glossary. Many of the words and concepts are explained more fully in the text, and their locations therein may be found be referring to the Index.)

ABSTRACT. A short summary of the essential thoughts contained in a DOCUMENT (especially a technical article or report), See also AUTO-ABSTRACTING.

ALPHANUMERIC OR ALPHAMERIC. (Adj.). Said of a reading machine VOCABULARY which includes both numerals and letters of the alphabet.

ANALOG, ANALOG CONCEPT. An electronic status or activity characterized by a smooth, continuous, uninterrupted *flow or continuum,* typified by the VOLTAGE of an electronic current or the rotation of a shaft. It is best understood in contrast with the DIGITAL concept, in which the action is pulsating, or broken into separate steps. One would use *analog* action in pushing open a door against a heavy obstruction on the other side, whereas he would use *digital* action in striking the door with repeated sharp blows to break the lock.

ANALOG COMPUTER. A type of electronic computer "which solves problems by translating physical conditions such as flow, temperature, pressure, angular position, or voltage, into related mechanical or electrical quantities, and uses mechanical or electrical

equivalent circuits as an analog for the physical phenomenon being investigated. Thus an analog computer *measures* continuously, whereas a digital computer *counts* discretely." *

So far as is known, analog computers have not been used for natural language information processing, but their principles of voltage measurement have wide application in the design and functioning of optical page readers.

ANODE. An element within an electron tube which *collects* and channels the free electrons given off by the CATHODE. It is usually connected with the tube's output terminal, which conducts the SIGNAL. The anode is also sometimes referred to as the "plate".

AUTO-ABSTRACTING. The preparation of an ABSTRACT of a DOCUMENT by machine, and more especially by computer. One of the most difficult of all literary tasks to automate because of the large amount of empirical judgment required in deciding what points in the document are the most important, and because of the extensive paraphrasing needed for adequate condensation. There is a fundamental difference between auto-abstracting and AUTO-INDEXING, with which it is often linked: the former deals with a single document and aims at preserving its essential import, while the latter usually deals with many documents and can record only a few words indicative of their subject matter. In brief, abstracting must preserve what the author *says* about a topic, while indexing can list only the topic itself.

AUTO-INDEXING. The INDEXING of SOURCE DOCUMENTS by machine, especially by computer. Note the distinction between auto-indexing and AUTO-ABSTRACTING, with which it is sometimes confused.

BINARY, BINARY SYSTEM, BINARY COUNTING. An ingenious numbering and counting system developed for use in digital computers. It is necessitated by the fact that computers have only two numerals or digits in their vocabulary, one and zero, representing the states of "on and off," "positive and negative, or clockwise and counterclockwise polarity," "pulse and no pulse," as the case may be. To

* Quoted from the *ADP Glossary of Terms,* compiled by the U.S. Bureau of the Budget, and reprinted by *Datamation* magazine, F. D. Thompson Publications, Inc., New York City.

represent large numbers the ones and zeros are used as markers in a *binary scale,* represented graphically as a series of columns or positions, each of which is headed by a number twice as great as that of the column on its right. Since the rightmost column is headed by 1, the column on its left is headed by 2, the next by 4, the next by 8, and so on to the left as far as desired. A 1 marker has the numerical value of the column in which it appears, and the actual or decimal value of a binary number is obtained simply by totalling the values represented by the 1 markers, as will be apparent from the following illustration. (The zeros are employed only to fill in the unused columns).

32	16	8	4	2	1			
					1	=	1	
				1	0	=	2	
				1	1	=	3	
			1	0	0	=	4	
			1	0	1	=	5	
			1	1	0	=	6	
			1	1	1	=	7	
		1	0	0	0	=	8	
		1	0	0	1	=	9	
		1	0	1	0	=	10	
	1	0	1	0	0	=	20	
	1	1	1	1	0	=	30	[16 + 8 + 4 + 2 + 0 = 30]
1	0	1	0	0	0	=	40	

Note that since the numbers to the left of the scale double each time, it is possible to represent extremely large numbers in straight binary notation. It is also possible to devise other coding systems, using only the first four positions (8-4-2-1) to represent single decimal digits, thus: 0110 0100 0111 1000 to represent 6478.

A more complete explanation of the binary system will be found in the chapter dealing with the recording of ALPHANUMERIC CHARACTERS on magnetic tape.

BIT. The smallest unit in computer language. It is a contraction of the words "BINARY digit." A bit may be expressed in several ways, depending on the MEDIUM on which it is recorded. Thus on a punchcard or paper tape a bit consists of a single hole punched in the paper; on a magnetic tape, disk or drum, it is a microscopic magnetized spot; and in CORE STORAGE, it is a single positively

or negatively charged magnetic CORE. A unique pattern or configuration of bits is assigned to each numeral, letter, or punctuation mark of the alphabet. Other patterns of bits are employed as special codes for use only within the electronic machine.

CARRIER. That component of an optical page reader which holds and transports the printed page.

CATHODE. The element in an electron tube which gives off free electrons when stimulated by electric current, heat, or light. See also PHOTOCATHODE.

CHARACTER. A single numeral, letter of the alphabet, punctuation mark, symbol, or other unit of printed matter. A character is usually represented in machine language by two or more BITS, with a separate arrangement or pattern of bits for each character. The automatic conversion of characters into their respective bit patterns is the chief function of the optical reading machine.

COMPUTERIZE. (v.) A handy vulgarism meaning to convert into a form that can be utilized by an electronic computer. Related words are *automate* and *mechanize.*

CONCORDANCE. An alphabetical listing of every word (with the possible exception of common words such as *the, and, of* and *in*) in a collection of SOURCE DOCUMENTS, with their respective text locations. The concordance is usually designed to be stored on magnetic tapes or disks, and to be MANIPULATED by computer for DOCUMENT RETRIEVAL.

CONCORDANCE INDEXING. A type of indexing in which KEYWORDS and key phrases are listed alphabetically, together with their REFERENCE NUMBERS. The index may be printed and employed in manual search, or stored in a computer and MANIPULATED by machine. In either case, a group of words describing the desired reference is assembled, and the search is for a reference number common to all the words. It may be noted that the concordance index is the antithesis of the HIERARCHICAL INDEX.

CORE, CORE MEMORY, CORE STORAGE. A single core is a tiny, doughnut-shaped ring of ferro-magnetic material (usually selenium) connected by hair-like wires to a power source and to neighboring cores in such a way that the polarity, or direction of flow of current circulating within it can be changed at will by con-

trolling the amount of current supplied to each core. Each core represents one BINARY BIT. When the current is flowing clockwise the core represents a BINARY 1; when in the opposite direction it represents a BINARY 0.

Cores are mounted in planes within rectangular frames, which are then stacked one upon the other to the desired size, providing a three-dimensional block. When so mounted they constitute the "core memory" or "core storage" of a computer.

A more complete description of cores and core memories will be found in the chapter dealing with the recording of characters on magnetic tape.

CORRELATION. The process, carried on in the RECOGNITION circuits, of comparing the visual or electronic image of an UNKNOWN CHARACTER with the images of prototype or "ideal" characters stored in the recognition circuits. Synonymous with MASK MATCHING and TEMPLATE MATCHING.

DATA RETRIEVAL. Recovery of statistical information, usually in numerical form, in response to such questions as "How many employees have had two or more years of college?", "How many of such-and-such items are on hand?" or "What is the total vote for Candidate X at the moment?" The answers may require the computer to assemble data from various storage sources and to make certain mathematical computations, all of which can be done under control of appropriate programs. Most "business data processing" applications are of this kind. It is a comparatively easy type of retrieval to handle by machine, where numerical data have been stored on some form of machine-readable MEDIUM in predetermined categories, corresponding roughly to the column-headings on an accountant's worksheet.

DIGITAL, DIGITAL CONCEPT. An electronic state or activity characterized by being broken into a succession of separate steps, states, or pulses. The digital concept also embraces the idea of compartmentalizing or classifying things into separate categories. It is best understood in contrast with the ANALOG concept, in which the action is smooth, continuous and uninterrupted. One climbs a set of stairs in digital fashion but rides an escalator analog fashion. See also the related term DIGITIZE.

DIGITAL COMPUTER. "A computer which processes information represented by combinations of discrete or discontinuous data, as compared with an analog computer for continuous data." * In brief, a digital computer *counts;* an analog computer *measures.*

DIGITIZE, QUANTIZE. To convert to DIGITAL form from some other (usually ANALOG) form. More specifically, in optical reading machines, to determine by applying certain prescribed criteria, or THRESHHOLDS, whether a color sensed by the optical system is black or white and hence should be represented electronically by a BINARY 1 or 0.

DISSEMINATION. That part of the retrieval process which consists of delivering the retrieved information to the requester, or distributing it to others who presumably have a need for it. Dissemination presupposes a retrieval system operated at some distance from its potential users. See also SELECTIVE DISSEMINATION OF INFORMATION (SDI).

DOCUMENT. A single, self-contained, complete unit of source material. It is usually identified by a *document number* for machine storing and retrieval purposes. A document may be a title or citation reference, an abstract or headnote, or the full text of the original source, although it would be very rare for it to consist of an entire book. What shall constitute the "document" is one of the decisions confronting the designer of a storage and retrieval system. In some applications a lengthy document may be divided into SUBDOCUMENTS, which may be retrieved separately on the basis of the specific subtopic within a more general main topic, e.g., a subdocument on airedales in a document about dog-breeding, or a paragraph on George Boole in an article on modern logic.

DOCUMENT RETRIEVAL. Recovery of the actual language of original documents containing a discussion of, or the answer to, the topic inquired about.

DUAL INVERTED FILE INDEXING. A form of computer MANIPULATED index in which all the words in the SOURCE DOCUMENTS are stored (in their natural-language form) on one magnetic tape or

* Quoted from the *ADP Glossary of Terms,* compiled by the U.S. Bureau of the Budget, and reprinted by *Datamation magazine,* F. D. Thompson Publications, Inc., New York City.

disk, while only the significant KEYWORDS (in coded form) together with their text locations or REFERENCE NUMBERS, are stored on another tape or disk. The advantages are that only one grammatical form of a keyword need be used in preparing search requests, misspellings and variant forms are taken care of, and the length of the index FILE is reduced considerably. See KEYWORDS IN LOGICAL COMBINATION METHOD.

ERROR. The false identification of one CHARACTER for another, as a Q for an O. Many so-called "errors" will be found on examination to originate from misspellings, typographical errors or misprints in the original DOCUMENTS, in which case they will seldom be repeated. If the same error is repeated, the machine is probably at fault and should be inspected to determine the cause. Note the difference between an error and a REJECT, in which the machine reports, by the use of an appropriate special symbol, its inability to identify a character. However, if the machine *should* have correctly identified the character but failed to do so, the entry of the reject symbol is an error.

ERROR RATE. The number of ERRORS (false identifications) committed by the machine per a specified number (10,000, 100,000 or 1,000,000, as the case may be) of characters scanned. The error rate may be expected to rise as the quality of the source material deteriorates.

FILE. (1) In BUSINESS DATA PROCESSING, a file has been variously defined as "an organized collection of information directed toward some purpose," and as "a collection of records containing information about a group of related accounts, people, stock items, etc., such as an accounts receivable file or a payroll file." In this usage the concept is that the data contained in the conventional categories of business accounting information (payroll, accounts receivable, etc.) has been converted into some type of machine-readable storage such as magnetic tape or punchcards, which then become the "file." See also MASTER FILE, TRANSACTION FILE, REPORT FILE.

(2) In INFORMATION RETRIEVAL, the idea of having the information stored in separate categories on some form of computer MEDIUM is the same, but the categories are different in name, content, and

purpose. The word FILE is usually accompanied by a modifier, as in the case of INDEX FILE, SOURCE DOCUMENT FILE, VOCABULARY (or KEYWORD, DESCRIPTOR, or OPERATOR) FILE, etc.

HEADNOTE. A brief summary or digest at the beginning of a published decision or opinion of an appellate court, usually prepared by the law publisher's editorial staff.

HIERARCHIC, HIERARCHICAL. A type of indexing or classification scheme in which the items within a class, and the classes themselves, are arranged in order of rank, with the more general terms at the top and the more specific terms at the bottom. At least three characteristics of a hierarchical index make it difficult, if not impossible, to mechanize: (1) the necessity for human intervention in determining that word A is more general or more specific than word B; (2) the fact that a word which is general in one context may be quite specific in another (for example, in a book on dogs, the word "dog" is very general, whereas in a book on the animal kingdom the word "dog" is quite specific); and (3) the fact that each word in a hierarchical index takes much of its meaning from the title of the class in which it is placed and from its neighboring terms, and thus may be non-informing or even misleading when standing alone.

INDEX. (noun) A reference tool consisting of words or phrases with their respective locations in a specific DOCUMENT, book, set of books, or LIBRARY, and so arranged as to indicate to a user the places therein where specific topics are treated. An index differs from a *table of contents* by being much more detailed and by being arranged without regard to the organization of the text into parts, chapters, sections or paragraphs. Every index contains at least two parts, the TERMS (topic words or phrases) and the REFERENCE NUMBERS (usually the volume and page or section number of the source material).

(verb) Depending on context: (1) To compile an index. (2) To position a document page within an optical reading machine in proper place for scanning.

INDEXIBILITY, INDEXIBLE. The characteristics of source material with regard to the ease or difficulty of indexing it.

INFORMATION RETRIEVAL. The term has two main connotations.

In the broader connotation, it is a convenient term for describing the processes of collecting, editing, storing, indexing, searching, finding, and disseminating information in usable form, and in this sense it includes DOCUMENT RETRIEVAL, DATA RETRIEVAL and REFERENCE RETRIEVAL.

In its more precise connotation it refers to a type of data recovery which would (if it could) provide "ultimate" answers to questions requiring the exercise of comparison, discretion, or advice-giving, such as "What is the difference between a capacitor and a resistor?" or "From the symptoms given, what disease does the patient have?" or "Is the Statute of Limitations in Texas the same as that in California?", or "Where is Uganda?" No information retrieval system yet devised can answer such questions entirely on its own.

KEYWORD. In the technical sense, an especially descriptive or significant word selected by a human indexer or an automatic indexing process to appear in an INDEX. It may or may not appear in the SOURCE DOCUMENT.

In the non-technical sense, a keyword may be any informative term in a source document.

KEYWORDS-IN-CONTEXT ("KWIC") or PERMUTED INDEXES. A form of computer-printed index developed within the last ten years, in which KEYWORDS selected by an indexer are arrayed in alphabetical order in the center of a page, and are preceded and followed by the words appearing on both sides of them in the original DOCUMENT, thus allowing the user to ascertain the context in which each KEYWORD was used. Special "packaged" computer PROGRAMS search for the prescribed keywords on the magnetic tape or disk file on which the source documents have been recorded and cause the computer line-printer to print the text line of which the KEYWORD is the approximate center. Other program routines sort the entries into alphabetical order and left-justify on the KEYWORD. The process is known technically as PERMUTED or PERMUTATION indexing.

The difficulty and expense of keypunching the entire text of source documents has usually limited the use of the KWIC technique to the compiling of "title indexes" for technical publica-

tions. KWIC indexing will come into its own for comprehensive general indexing when optical page readers convert the full text of important source documents into computer language quickly and cheaply.

"KEYWORDS-IN-LOGICAL-COMBINATION" METHOD OF DOCUMENT STORAGE AND RETRIEVAL. A basic approach to fully automated storage, indexing and retrieval of SOURCE DOCUMENTS. The documents are copied in full text on a keypunch medium and recorded on magnetic tape. The computer sifts the text and prepares a CONCORDANCE-like search index. Search inquiries are prepared in natural-English combinations of KEYWORDS selected by the user as descriptive of the contents of desired documents. The computer searches the stored index (not the full text) and reports all documents which have the specified KEYWORDS as a common denominator. Considerable precision is possible through the use of a request format providing the equivalent of the logical "connectors" *And* (requiring that *all* terms be present), *Or* (requiring that at least *one* of the terms be present) and *Not* (requiring that a certain term or terms *not* be present). The method has been applied successfully to the retrieval of legal literature, i.e., legislation, judicial precedents, and government agency rulings.

LETTERPRESS. A printing process in which the impression is made directly by inked metal type, plates or mats as distinguished from the offset process. For our purposes, the term also includes reproductions of original letterpress material, such as those done by duplicating machines.

LIBRARY. A collection of SOURCE DOCUMENTS.

MACHINE READABLE. Until the advent of optical reading machines, the term meant simply a form of language notation which could be sensed electrically or mechanically by a machine, e.g., holes in a punch card or paper tape, or patterns of magetized spots on magnetic tapes, drums, or disks. Since the coming of optical readers it has taken on an additional meaning, and now refers to a type of material which is capable of being read by an optical system. In this sense typewriting and letterpress printing may be said to be machine-readable, whereas unstructured cursive handwriting is not, as yet.