CONCEPT OF AN ON-LINE COMPUTERIZED LIBRARY CATALOG

Frederick G. KILGOUR, Director, The Ohio College Library Center, Columbus, Ohio

A concept for mechanized descriptive cataloging is presented, together with four areas of research programs to be undertaken.

This paper will describe a concept of a catalog that is hospitable to mechanized descriptive cataloging, and will delineate major areas of research for production of knowledge necessary to implement such a catalog. To avoid an unnecessarily complex presentation, the discussion will treat only of printed books. Nevertheless, the catalog described will function equally well as a store for serials, journal articles, reports, and any other materials that carry bibliographic-like descriptions of themselves.

As used in this paper, a concept is an idea that combines experience with and observations of catalogs, and suggests further experimentation and observation. The merit of a concept is measured by its fruitfulness in production of new ideas and new experimentation and observation. The purpose of the concept proposed in this paper is to suggest avenues of investigation that will yield findings useful in development of mechanized, descriptive cataloging.

The paper opens with a brief discussion of the objective and functions of a library catalog. Next there is an analysis of the principal contribution of information retrieval during the past quarter century and a proposal for applying this advance to cataloging of books. The third section describes a plan for a new style of catalog, and the fourth shows how it will be possible to prepare entries mechanically from title pages for inclusion in such a catalog. There follows an outline of major research investiga-

tions to be undertaken to produce knowledge necessary for activation of the new style catalog and mechanical cataloging. The paper concludes with a brief estimate of the success the system may be expected to attain in achieving the objectives that appear at the start of the paper.

OBJECTIVES

The principal objective of a future library will be active participation in the program of the community or institution of which it is a part by furnishing members of the community with bibliographic, textual, and other recorded information when and where they need such information. The passive service functions that librarianship has developed during the past century are no longer adequate to maintain a library as a viable organization within its environment. Effective special libraries have rid themselves of the passive service concept and aggressively participate in the programs of their companies.

An extreme example of active participation in institutional programs is the library-like sections of intelligence agencies. Here collectors and processors of new information do not place that information on shelves or in files with the expectant hope that someone will use it. Rather the collection and processing staff immediately brings new information to the attention of those making decisions that the new data may affect.

A fruitful concept of a library is as an external human memory. Since Aristotle, it has been recognized that memory is necessary for creative thinking. The process of creative thinking requires raw materials from memory; but for centuries it has been impossible for man to retain within his memory all data that could fuel his creative thinking. Indeed, one great triumph that sets human beings apart from all other animals is the ability to store data in an external memory such as a library. However, to support creative thinking, the external memory must transfer data to the human mind with as great speed as possible to prevent hindrance of thought that admits distraction. It is this lack of speed that generates frustration among library users. If a library is to participate effectively in programs of its institution or community, it must simulate human memory in furnishing a human mind with data when and where that mind needs data. Current development in computers, and particularly in their memories, holds out the hope of highly effective external memory operation at some point in time beyond the foreseeable future, but in the meantime it is entirely feasible to strive for simulation of human memory with speedy recall of bibliographic information.

It has been pointed out elsewhere (1) that productivity of library workers is not continuously increasing as is that of workers in the general economy, so that library unit costs are rising at a much more rapid rate than are those in the economy as a whole. If libraries are to attain their objectives in the future, they must invoke a technology that will enable them to lower their excessively high rate of unit-cost rise. It appears that

mechanization, or more specifically, computerization, is the only avenue that extends toward the goal of economic viability.

INFORMATION RETRIEVAL

During the last quarter-century, there have been important developments in information retrieval that have yet to be applied to book collections. Such pioneers as W. E. Batten, G. Cordonnier, Calvin Mooers, and Mortimer Taube made a major innovation when they developed coordinate indexing. This technique coordinates index terms at time of searching, employing Boolean logic. Coordinate indexing has greatly increased flexibility of searching and number of accesses to documents in contrast to the precoordinated headings in traditional catalogs that are inflexible for searching and for up-to-date maintenance.

Early information retrieval systems dealt with relatively small files of documents, such as patents and internal reports, that were not subject to traditional bibliographic control. Moreover, indexes to these files were housed in various manual devices. With the advent of the computer it became feasible to apply coordinate indexing techniques to large files of documents, including materials under classical bibliographic control. However, up to the present time, the techniques of information retrieval have been applied primarily to huge files of journal articles; an outstanding example of the application of coordinate indexing to journal articles is the pioneering Medlars project, in which the primary approach to article is via coordination of subject indexing terms.

Retrieval of books from a library collection is an information retrieval process irrespective of whether the borrower uses subject headings or author-title entries in the catalog. The user who obtains a book from a library by employing the book's author-title label is logically engaged in the same information retrieval process as is the user who searches out a book under a subject heading.

At first reading of the previous paragraph it might appear that a reader who obtains a novel by use of an author-title entry in a catalog is not engaged in information retrieval in its customary narrow sense. However, it is clear that the reader of a novel or poem is acquiring information in the same sense as is the reader of a book on computers, although knowledge he gleans from a novel is not for immediate practical application, but rather to enable him to understand what it is like to be a human being, and more specifically what it is like to be a human being in some of the precise circumstances of life.

For the library user the words in an author's name, a title, and subject headings, are index labels that he uses to find a book that contains information he needs. The traditional use of these index labels, at least since the Middle Ages, has been some variety of an author and title citation form. The user knows externally to the citation that the book so labeled has information he wants. Apparently three-quarters of the information

retrieved from a library by use of a library catalog is via an author-title entry (2, 3) or a known document search (4).

A librarian's use of a catalog (except for reference librarians who represent users) is to discover whether or not the library has a given book. The librarian does not use the author-title entry as a label, but rather as information *per se*. Librarians include sufficient data in catalog entries to enable them to decide from the description whether or not a book at hand or a book described by another citation is the same book as that in the catalog records.

In short, users employ a library catalog to direct them to information they require; librarians use the catalog for the actual information it contains.

COMPUTERIZED CATALOG CONCEPTS

Several libraries employ nonconventional design for computerized catalogs or lists of other than bibliographic items. The Stanford University Library's on-line system uses a sequential file of entries to which there is an index of words in the author and title elements of the entry as well as other words in other elements (5). Index files for various Stanford data collections are "Author, Title Word, ID Number, Corporate Author, Conference Author, Keyword, Citation," and for certain files, topical subject indexes. In general, this system is widely employed in the organization of computerized files, but the Stanford application has a unique feature in that it uses a derived key consisting of only the first three letters of index words. For example, the derived keys for author and title words in Figure 1 are VIC, ONb, RET, SYS, and THE. The computer calculates the position of the word in the index from these trigrams so that it is possible to locate the index word with great speed. This technique of employing a derived key to compute location takes full advantage of a computer's major characteristic, namely, ability to compute rapidly.

The Washington State University Library has developed a similar system for its on-line acquisitions file. Access to an entry in this linear file is by purchase order number. A random number generator uses the purchase order number to compute position of the entry in the file. From early trials, this technique appears to make possible exceptionally efficient use of random access file space.

Yale's Machine Aided Technical Processing System uses derived keys to locate entries for book funds in the system's commitment register and entries in a name and address file employed for addressing notices and claims. The Yale Technical Processing System also uses a derived key technique to detect duplication of purchase orders entering the master file. This system operates using the first four letters of the author's name, first three letters of the first non-article word of the title, and first letter of the second title word if there is a second word. A routine run on 23 June, 1969, compared 1,237 new entries against 63,641 entries already in

the file. The comparison produced 199 couplets containing possible duplicates of which 115 couplets actually were duplicates; of the 115, only forty would have been obtained if an equal compare had been made throughout the author and title fields.

Several investigators are working on techniques for derivation of keys (6, 7, 8). Similar work on telephone directories (9) has yielded preliminary results indicating that an efficient formula for derived keys for personal listings is the first three letters of the surname and first three letters of the street name; and for business listings the first three letters of the first word and first three letters of the second word.

The Ohio College Library Center is working on development of a computerized catalog for traditional catalog entries wherein a computer will compute position of an entry in a file organized in a two-dimensional array from a precoordination of truncated strings of letters from words in the author's name and in the title. OCLC plans to use this technique because, as already noted, three-quarters of the use of a library catalog seems to be use of author-title entries. Precoordination of derived keys from these two elements will speed average retrieval time. The present design calls for a microcatalog containing, on the average, perhaps fewer than five entries to be located at each computed position. Having computed a location, a computer will search the microcatalog for entries possessing derived keys matching the original request, and entries satisfying this requirement will be displayed as a minilist on a cathode ray tube terminal. It is hoped that algorithms can be constructed that will yield minilists containing fewer than twenty entries 95% of the time.

Indexes to the proposed main entry file will be the equivalent of classical added entries. However, it is fruitful to view subject headings, title added entries, and author added entries as being continuous text, from which uniterms can be mechanically extracted. Under each uniterm will be a list of addresses of the microcatalogs containing the corresponding entries, and each entry could be looked upon as analogous to the concept of a microtheme proposed by T. P. Loosjes (10). Coordination of indexing at search time by the user need not be limited within subject words, or title words, or author words. Rather, coordination among these elements will greatly increase accesses to entries and will make possible retrieval of entries with a relatively slight amount of bibliographic information, particularly if each word is truncated as described above.

Although much research and new knowledge is necessary to achieve successful design of the type of catalog described above, there is no technical obstacle to its successful activation for experimentation. When practical implementation and routine operation are also successful, the user will be employing a minilist containing twenty or fewer entries most of the time. In other words, the reader will be using a catalog of twenty or fewer entries, and such a catalog makes it unnecessary to include bibliographical embellishments required for entries in huge card or book-

form catalogs. It would appear that for catalogs of twenty or fewer entries only information on title pages would be required; a scholar rarely, if ever, needs more. Hence, it seems feasible that mechanization of descriptive cataloging could be achieved in the foreseeable future.

MECHANIZED DESCRIPTIVE CATALOGING

The organization for a computerized catalog containing entries prepared mechanically from title pages would be somewhat different from that described in the previous section. If it proves impossible, as seems likely, to devise an algorithm that would mechanically identify author, title, and other elements on a title page, it would be necessary to arrange entries in sequential order. A computer could then prepare a mechanical coordinate index of substantive words on the title page that would make possible at search time coordination of words in author, title, and other elements without having to identify those elements. Of course, catalogers would still do subject classifying and indexing, as well as assigning of call numbers, but a computer would mechanically convert these additions to the uniterm type of coordinate indexing described in the previous section.

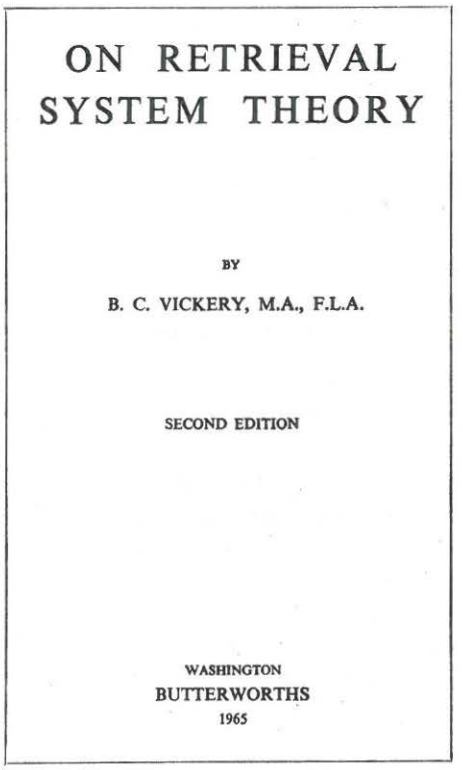
This proposal to construct a bibliographic record in the form of a transcription of a title page is not new. An early code for production of catalog entries, which the French Government issued in 1791 (11), prescribed transcription of the title page, and underlining of the author's name as the filing term. If the book did not have an author, the key word in the title was to be underlined. The code also provided for the title-page transcription to be supplemented by a physical description of the book.

This proposed new concept for a computerized library catalog closely relates to the Stanford design and the planned OCLC design. However, in contrast to the new concept, the Stanford file organization requires identification of record elements from within which words are extracted for inclusion in indexes, and the indexes are so tagged. Similarly, the present OCLC plan also requires identification of author and title elements for calculation of location, and hence for retrieval, as well as flagging of other retrieval elements, such as record number and call number; but the OCLC system will not make necessary identification among types of added-entry elements. The proposed new concept expands this last device to the entire record.

A computer simulation has been carried out of an on-line computerized catalog containing descriptive entries prepared mechanically. Access to the simulated catalog was by coordination of non-structure words in titles via single-level indexes. Simulation of user inquiries at a peak rate of five per second, processed on an economically feasible computer, revealed that utilization of the computer's central processing unit was only 19.87 percent. It is known from other simulation studies that library use of such

a computerized catalog would raise utilization by only two percent at the most. Hence it follows that there is at least one (and probably several) existing, economical computer system that can be employed for such a catalog.

Mechanical descriptive cataloging of the title page depicted in Figure 1 would be efficient and effective. The only character strings on the title page that would not be useful in coordinate indexing are "BY" and "M.A., F.L.A.". However, the title page in Figure 2 contains at least seven, or perhaps eleven, words and symbols that would not


ON RETRIEVAL SYSTEM THEORY

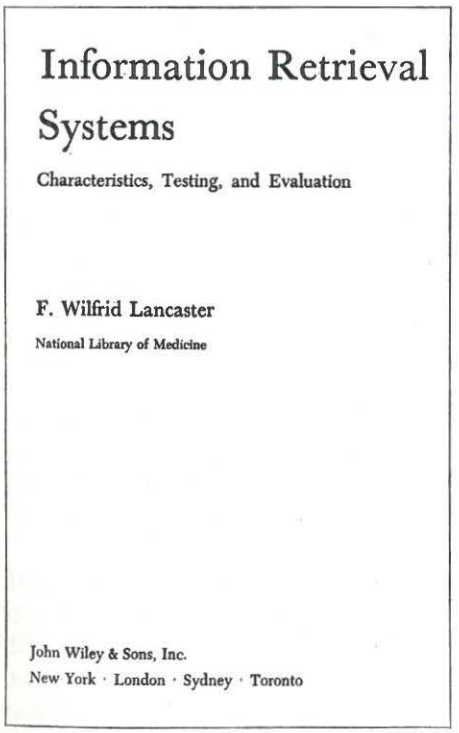
BY

B. C. VICKERY, M.A., F.L.A.

SECOND EDITION

WASHINGTON
BUTTERWORTHS
1965

Fig. 1. Title Page.


Information Retrieval Systems

Characteristics, Testing, and Evaluation

F. Wilfrid Lancaster

National Library of Medicine

John Wiley & Sons, Inc.
New York · London · Sydney · Toronto

Fig. 2. Title Page (undated).

be employed in coordinate indexing. If these eleven were to be included in the index and remain unused, they would approximately double the size of the index for this particular title page. Such inefficiency is too large to tolerate. Morever, the title page in Figure 2 does not contain date of publication. Effect of absence of publication date from entry and index is not known, although a recent study suggests that date of publication may be of relatively little use as a retrieval element (12).

The text of a title-page would be displayed as a string of characters and not rearranged as is done in traditional catalog entries. No doubt sophisticated algorithms will be devised to format displays, but even a simple algorithm produces a useful representation of title-page information. For example, by employing the simplest of algorithms that would insert two spaces at the end of each title-page line, the title page in Figure 1 would appear as follows on a terminal.

ON RETRIEVAL SYSTEM THEORY BY B. C. VICKERY,
M.A., F.L.A. SECOND EDITION WASHINGTON
BUTTERWORTHS 1965

Readers have used title pages successfully for centuries and will surely experience no difficulty in using them displayed in this manner on terminals.

It is hoped that ultimately it will be possible to use optical character recognition techniques for mechanical transcription of most title pages. Until effective OCR techniques are available, it will be necessary for clerical staff to transcribe title pages, and such employment for human beings is undesirable. However, libraries now employ clerical staff to transcribe bibliographic information for entries in essentially the same manner, so that continuance of an existing practice in this instance cannot be looked upon as invocation of a machine to convert human beings to machine-like activity. Nevertheless, machines should replace such activity at the earliest opportunity.

RESEARCH

There are at least four major areas of unknown on which research must be carried out to produce knowledge needed for development of a computerized library catalog hospitable to descriptive cataloging entries produced mechanically: 1) use of library catalogs; 2) specification for derived keys; 3) identification of title-page words useful and not useful for coordinate indexing; and 4) extent and type of coordination necessary to ensure successful retrieval.

Extraordinarily little is known about users' employment of library catalogs to obtain information from books. Yet successful design of a catalog must be based on firm knowledge of catalog use. Some areas of the broad pattern of catalog usage are known, but much more must be discovered before an effective catalog can be designed.

Descriptive cataloging rules have long been derived from rationalized principles of title-page and catalog formats. As yet there has been no major effort to derive these rules from the bibliographic practices of library users. For example, there has been no general effort to construct rules for descriptive catalog entries that match scholarly bibliographic references in such a way that a scholar could always expect to find in a library catalog essentially the same entry presented to him in a biblio-

graphic footnote. To design new catalogs based on the various scholarly traditions of citation will require a series of analyses of citation practices that will ultimately yield descriptions of minimum regularity.

The section of this paper on computerized catalog concepts has referred to research on specification for derived keys. Such specification is required to enable swift access to files and at the same time to diminish human error in search requests. Traditional designs of computer files are inadequate for management of huge files of millions of bibliographic entries.

At the present time it appears that the truncation algorithms already referred to may be able to cope successfully with a majority of catalog entries. However, it is clear that such truncation techniques will not provide uniqueness of all keys adequate for efficient on-line catalogs. Therefore, it will be necessary to carry out a series of investigations that will identify classes of entries for which a basic algorithm does not operate satisfactorily and to devise a supplementary algorithm to improve uniqueness of keys for those entries for which the basic algorithm essentially failed. Presumably this cycle will be repeated as long as inadequacy of key uniqueness persists. In other words, research in this area will continue as long as retrieval inefficiency exists for the user.

Uniqueness of key depends on uniqueness of the serial combination of words from which the key is derived. Hence analyses of frequency of word occurence on classical catalog entries, title pages, and in subject indexes, should be carried out with the aim of deriving a generalized description of such frequency distributions. Such findings will be necessary for sophisticated logical and physical file organization.

To organize an efficient, huge file of bibliographic entries it is necessary to develop a method for computing scatter storage addresses that provides a very high percentage of unique addresses, thereby avoiding a collision with an entry already in an address. Of course, it is necessary to furnish a hash-coding, or scatter-storage, algorithm with keys that possess high relative uniqueness; otherwise, the most efficient of scatter-storage algorithms would yield non-unique addresses in ratio to the degree of non-uniqueness of keys. P. C. Mitchell and T. K. Burgess (13) have introduced random-number generation for computing scatter-storage addresses and have shown their method to be more efficient than division hash coding. Other investigators are working on techniques for minimizing queues resulting from repeated collisions. There is need for continuing imaginative investigation that will yield results like that of Mitchell and Burgess before huge bibliographic files and their indexes will be accessed efficiently.

Identification of useful and non-useful words for coordinate indexing on title pages, including those in foreign languages, is related to catalog usage. At the present time there is no information that gives a clue as to size of a list of non-useful words. Much ingenuity and imagination will

be required to identify non-useful words and to construct efficient null lists.

Finally, investigation will be needed to determine amount and type of coordination necessary among author and title words. It will also be essential that a measure of retrieval success by author and title be developed. The need here is construction of a meaningful measure for retrieval of a single entry.

CONCLUSION

The proposed concept for an on-line computerized library catalog will make it possible for a user to obtain bibliographic information from a remote terminal rapidly. Use of derived keys would increase error tolerance well above that of present manual systems by diminishing effect of misspellings and by making it unnecessary for the user to have knowledge of catalog organization. Moreover, the concept is a step toward full mechanization and can indeed be viewed as a partial simulation of text processing.

The proposed catalog will also make it possible for libraries to take the first major step toward their economic goal of development of a continuously increasing productivity for both library staff and library user. It is anticipated that successive steps to come after mechanical descriptive cataloging will be automatic subject classification and indexing, to be followed ultimately by full text processing. When it is possible to achieve full text processing mechanically, and a decade or more may be required for that achievement, libraries will have succeeded in attaining their objective of participation as well as their economic goal of rate of cost rise equal to that in the general economy.

REFERENCES

1. Kilgour, Frederick G.: "The Economic Goal of Library Automation," *College & Research Libraries,* 30 (July 1969), 307-311.
2. Tagliacozzo, Renata; Kochen, Manfred; Rosenberg, Lawrence: "Orthographic Error Patterns of Author Names in Catalog Searches," *Journal of Library Automation,* In press.
3. Brooks, Benedict; Kilgour, Frederick G.: "Catalog Subject Searches in the Yale Medical Library," *College & Research Libraries,* 25 (November, 1964), 483-487.
4. Lipetz, Ben-Ami: "A Quantitative Study of Catalog Use" In University of Illinois Graduate School of Library Science: *Proceedings of the 1969 Clinic on Library Applications of Data Processing,* (Preprint).
5. Parker, Edwin B.: "Developing a Campus Information Retrieval System." In *Proceedings of a Conference Held at Stanford University Libraries, October 4-5, 1968* (Stanford, California: Stanford University Libraries, 1969), pp. 213-230.

6. Ruecking, Frederick H., Jr.: "Bibliographic Retrieval from Bibliographic Input; The Hypothesis and Construction of a Test," *Journal of Library Automation,* 1 (Dec. 1968), 227-238.
7. Nugent, William R.: "Compression Word Coding Techniques for Information Retrieval," *Journal of Library Automation,* 1 (Dec. 1968), 250-260.
8. Kilgour, Frederick G.: "Retrieval of Single Entries from a Computerized Library Catalog File," *Proceedings of the American Society for Information Science,* 5 (1968), 133-136.
9. Rothrock, Hamilton Irving, Jr.: *Computer-Assisted Directory Search; A Dissertation in Electrical Engineering* (University of Pennsylvania, 1968).
10. Loosjes, T. P.: "Document Analysis," *Proceedings of the Third International Congress on Medical Librarianship* (1969), Preprint.
11. Instruction pour Procéder à la Confection du Catalogue de Chacune des Bibliothèques (Paris: Imprimerie Nationale, 1791), p. 6.
12. Vaughan, Delores K.: "Memorability of Book Characteristics: An Experimental Study." In University of Chicago Graduate Library School: *Requirements Study for Future Catalogs* (Chicago: University of Chicago Graduate Library School, 1968), pp. 1-41.
13. Mitchell, Patrick D.; Burgess, Thomas K.: "Methods of Randomization of Large Files with High Volatility," *Journal of Library Automation,* 3 (March 1970), 79-86.