DEVELOPMENT OF COMPUTERIZATION OF CARD CATALOGS IN MEDICAL AND SCIENTIFIC LIBRARIES*

Frederick G. Kilgour

Various scientific libraries are computerizing their card catalogs; some produce catalog cards and others have gone to book catalogs. At least one library associated with the military establishment is developing an information retrieval system for catalog card information employing a large, high-speed computer. Still, relatively little work is being done on computerizing the retrieval of catalog and index information—well-known projects being the Medical Literature Analysis and Retrieval System (MEDLARS) at the National Library of Medicine and the information retrieval system of the American Society of Metals. Both of these systems employ sequential searching of magnetic tape. However, this paper will not attempt to survey these burgeoning activities completely but will report only on the Columbia-Harvard-Yale Medical Libraries Computerization Project.

It is in searching the file that the Columbia-Harvard-Yale Project differs from most others. It is intended that the file will be in a random access memory device and will be on-line for each of the libraries. One of the specifications for the design of the system is that the answer to an inquiry should begin to come out of the system within a minute or two after the question has been inserted. Furthermore, it is hoped that articles indexed by MEDLARS in some 260 journals supplying upwards of 75 per cent of recorded use in the three libraries will be included in the computer file. However, this paper will be confined largely to the discussion of the computerization of book cataloging.

Frederick G. Kilgour is Librarian, Yale Medical Library, Yale University, New Haven, Connecticut. The author is grateful and indebted to his colleagues, Thomas P. Fleming, Librarian, Columbia Medical Library, and Ralph T. Esterquest, Librarian, Harvard Medical Library, for making possible and directing with imagination and wisdom the Project described in this report.

*A national Science Foundation grant supports the work reported in this paper.

The Project is based in large part on the fact that there exists in scientific and medical libraries a relatively small core of the book collection that supplies upwards of three-quarters of the use of the library. The Yale Medical Library possesses some 350,000 items of which more than 110,000 are European theses. Still, a study of recorded book usage in the Yale Medical Library done three years ago showed that books published in the previous twelve years, or about 10,000 volumes, furnished 79 per cent of the recorded use.[1] More recently, a study at Columbia and Yale has shown that of the 2,000 journals then being received at Columbia and of the 1,500 at Yale, 262 furnished 80 per cent of the recorded use of recently published journals.[2] It is these heavily used cores that make the on-line computerization of catalogs and indexes economically feasible.

The Columbia-Harvard-Yale Project began to get under way in the autumn of 1961. Important for its initiation and present prosecution was a suggestion made by a group at ITEK Corp. whose most prominent members were Lawrence Buckland, Ben-Ami Lipetz, and David Sparks. This group pointed out that it had become possible to produce cataloging information in machineable form that could be used to continue the production of card catalogs and could be accumulated over several years to be put in a computer file when sufficient cataloging information was available to justify using a computer. Columbia, Harvard, and Yale drew up a request for a grant from the National Science Foundation (NSF) late in the summer of 1962 and revised this request at the end of 1962. The purpose of the grant was to finance the development and initiation of a computerized library catalog system. NSF made the award in the summer of 1963. However, the new procedures had been initiated in the late winter of 1962/63 so that at Yale all books possessing an imprint of 1963 and later have been processed in the new procedures.

The goal of the system is to increase the speed and completeness with which a user obtains catalog and index information in a library. The Project is attempting to play a significant role in the development of computerized catalogs which will undoubtedly be the next major step towards increased speed and completeness of library services following the nineteenth-century introduction of the card catalog, and the abstract and index journals. The purpose of the Project is to design a computerized catalog system and to demonstrate the feasibility of such a system. In addition, it is anticipated that the system, including computer programs, can be taken over and used by a majority of conventional libraries.

As far as library-like activities are concerned, it is convenient to think of information retrieval from a large corpus of information as being in three categories. In one category is the library supplying heavily and moderately used information rapidly—often within a few minutes. Next there is the bibliographical variety furnishing

relatively little-used information slowly—perhaps in a day or two. An example would be the MEDLARS project at the National Library of Medicine. Finally, there is the documentation category of information retrieval wherein specific, detailed data are furnished, usually after a period of time which may amount to a week; the documentation project of the American Society of Metals is an example of this category. The Columbia-Harvard-Yale Project is in the first group.

Information Retrieval System

The main goal of the Columbia-Harvard-Yale Project is information retrieval—the rapid and complete retrieval of cataloging and indexing information. As already mentioned, only that cataloging and indexing which relates to the relatively small core collection in medical libraries will be computerized. Studies done at Yale indicate that books supply approximately 40 per cent of recorded usage, while journals furnish nearly 60 per cent.[3] It will be the cataloging of that part of the book collection supplying upwards of 75 per cent of use which will be computerized. The activation of the information retrieval system cannot occur before 1966, but it is hoped that it will begin operation in that calendar year.

At least a half-dozen significant achievements are expected of the information retrieval system. As already mentioned, it will increase and make more complete the supplying of cataloging and indexing information as compared to the present tedious card-by-card search. The primary approach to the catalog in the computer file will be by subject, but any given subject can be coordinated with perhaps up to four more subjects, with the date of publication, the language, and the place of publication. An example might be a search for books discussing the use of computers for information retrieval in science and published in English after 1962. In a card catalog there probably would not be an excessive accumulation of entries under any of the headings equivalent to these subjects. However, a search on the relationship between cancer and enzymes in the card catalog of the Yale Medical Library would involve going through 600 cards under cancer and 100 under enzymes. Another important advantage of the information retrieval system involving the catalogs of three libraries is that those three catalogs will be searched as one for users in each library. Since 55 per cent of book holdings are in but one or two of the three libraries, users in each library will enjoy increased access to literature, albeit that some would not be available for a day or two. In addition to the coordinated subject searches, there will also be an increased depth of subject cataloging—a third benefit. As is well-known, libraries now keep to a minimum the number of subject

headings for each title in order to slow the engulfing growth of the card catalog. At the Columbia and Yale medical libraries, subject headings per title average 1.7;[4] at Harvard, 1.8;[4] and at the Library of Congress, 1.6.[5] Since this need for repression will disappear with the advent of the catalog in a computer file, it will be possible to do more adequate subject analysis in depth. During the early months of the application of the new procedures in the Yale Medical Library, the number of subject headings assigned to each book rose from 1.6 to 3.2. At the present time the figure is higher, with the goal being an average of five subject headings per title.

A study of the use of the subject cards of the catalog of the Yale Medical Library was carried out last autumn and yielded results[6] which seem to indicate that present-day subject cataloging is somewhat less than adequate to meet the increasing demands for information. The subject cards were used but 12 per cent of the time when catalog use by the technical staff of the library was included. When such utilization was excluded, use of the subject catalog was 18 per cent.[7] These low percentages indicate that the present subject card catalog is not the favored tool of users. However, it is hoped that with speeded access to the subject catalog and with a greater depth of subject cataloging this tool will increase in usefulness.

Greater completeness in catalog searching can be assured because it will be possible to teach the computer always to search "see also references." It is impossible to teach users to search "see also references," and they thereby miss useful titles at times.

Another achievement of the information retrieval system will be a relative ease in printing out the catalog in book form. Certainly author and title catalogs will be produced in book form, and it would be equally possible to print subject catalogs. Copies of these book catalogs could be placed outside the library in various laboratories, thereby increasing the availability of materials in the library. Moreover, book catalogs are far easier to use than card catalogs and stimulate catalog "browsing." Finally, a computerized information retrieval system will also provide for selective periodic dissemination of new cataloging information. "Current awareness listings" of certain subjects could be furnished on a periodic basis. Indeed, the first product—actually a by-product—of the Project was the mechanized production of the monthly Bulletin of the Yale Medical Library which lists accessions for the previous month.

There are also several administrative benefits which will accrue from a computerized catalog system housing the catalogs of two or more libraries. Cataloging expense will be reduced, or amount of cataloging increased, because such a computerized catalog is in effect a union catalog. For instance, an investigation of duplication amongst the Columbia, Harvard, and Yale medical libraries revealed that 66, 84, and 83 per cent respectively of each collection

is in one or both of the other two collections. If acquisitions of books were completely at random in the three libraries, each library would need to catalog only half the percentage of its collection duplicated in one or both of the other two libraries. It therefore appears reasonable that perhaps one-third or one-quarter of present cataloging costs will be eliminated. Of course, the more libraries that participate in one system the greater will be the decrease in cataloging costs providing all libraries acquire works at approximately the same rate and in the same subjects. Another administrative benefit is that such mechanized procedures are faster and more accurate. Once a decklet of punched cards containing cataloging information about a title has been produced and verified, it can be employed for each subsequent activity rapidly and without need for repeated proofreading. Another administrative advantage, albeit perhaps not a major one, is that the ballooning card catalog could be housed elsewhere than in a library's busiest and most desirable location. A computerized catalog will be in the computer file, presumably in a computer room outside the library and enormously reduced in size. Catalogs in book form could be widely available, but as is well-known, book catalogs occupy far less space than card catalogs.

An on-line computer catalog is made possible by locating in each library an information station possessing a telecommunication connection with the computer. It is likely that the computer will be located in New Haven, but actually the Yale Medical Library will be no closer to it electronically, than the Columbia or Harvard medical libraries. Present specifications for the information retrieval system include a time lapse of no greater than one minute at any information station—no matter how remote—between the end of the inquiry going into the computer and the start of the reply from the computer. These information stations will probably consist of an electric typewriter and a card reader. The typewriter will be used to put questions to the computer and will type out the replies in the form of catalog references and will include call numbers. The card reader will process cards to add to the computerized catalog.

A second basic technical feature of the system will be a random access memory unit that will hold the catalog files. Information sought in such files can be found in a fraction of a second. There has not yet been agreement upon the specifications for the arrangement of information in such a file, but one possible configuration would be to have the equivalent of catalog cards in one section of the file. A second section would have the addresses of subjects, and under each subject would be the number of the catalog card on which that heading occurred. The analogy to a card catalog would be an author catalog not filed by subject but with each card having, perhaps, a sequential identification number. Each identification number would then be listed on a subject card which would be filed by a coded, numbered

address in a different drawer. In a computer when a work on three different subjects is sought, the numbers under those three subjects would be brought from the random access files into the core of the computer where the numbers would be compared. Each number that occurred under all three subjects would then be brought from the equivalent of the card file and dispatched electronically to the remote information station.

Mechanized Catalog Card Production

Although the information retrieval system is in the design period, the mechanized production of catalog cards is in the early stages of operations. The procedure starts with the cataloging of a book on a 8-1/2" x 11" worksheet (see Fig. 1). The cataloger makes a format of the card on the worksheet and the keypunch operator

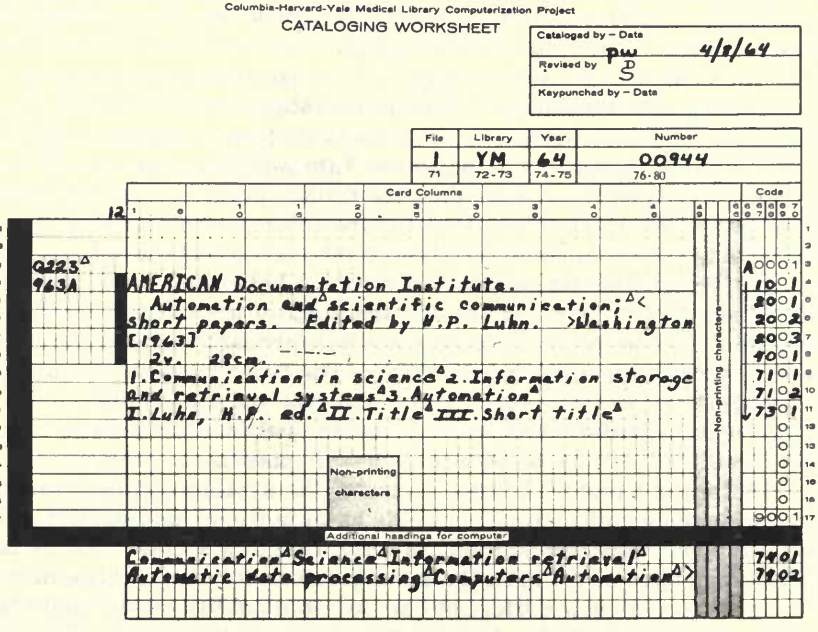

Figure 1

punches one card for each line on the sheet. The resulting decklet of cards, with many other decklets, is then fed into a computer behind a program which expands the decklet into the number of cards required and puts these cards on magnetic tape. After a computer

has sorted these records by filing entry, they are punched out on punched cards. These punched cards drive an electric typewriter which produces the cards in their final form and in alphabetical order ready for filing in the catalog.

The computer programs, except for the sorting program, are designed for an IBM 1401, 4K core, 2-tape drive computer. The census of computers in the March 1964 Computers in Automation showed that 37 per cent of all computer installations were 1401's. The nearest competitor was the IBM 1620, but the 1401 had nearly five times as many installations as the 1620, a small scientific computer not particularly suitable for non-numerical data processing. Clearly, the 1401 is the most widely available computer and because of this fact the Columbia-Harvard-Yale Project programs have been written for this computer so that the programs would be of the widest possible use to other libraries.

In the cataloging procedure the only change from the operations in many libraries is the use of a worksheet (see Fig. 1). The worksheet shown is one designed for a cataloger who prefers to write out the catalog card. At Columbia a worksheet which can be used in the typewriter is employed with success. At Yale the catalogers find that the worksheet enables them often to establish the main entry in its final form at the catalog, thereby eliminating some recopying. The format of the catalog card on the worksheet mimics almost exactly Library of Congress format, the only difference being that topical subject headings, name subject headings, and other tracings each begin on a separate line.

Certain positions on the worksheet must be precise and flags or signals to the computer programs must be placed in a half-dozen locations on each sheet. The "Directions for Use of the Cataloging Worksheet" occupy about 5-1/2 typewritten pages; in other words, they are not tremendously long and complicated. Nevertheless, the location of the call number and of the beginning of the main entry, the title, the collation, and other groups of information on the card are precisely defined, as are the locations of the flags. For instance, if a short title tracing is employed, a delta must be inserted in the space preceding the first letter of the first word of the short title, and similarly, in the space following the last letter or punctuation mark. A "less-than" sign is always placed at the end of the title or the title added entry and a "greater-than" sign before the imprint.

The worksheet in Figure 1 makes it possible to use one set of subject headings on the printed catalog cards and another set for the information retrieval computer. As an example, the Yale Medical Library is continuing to use Library of Congress subject headings in its card catalog, but of course employs Medical Subject Headings (MeSH) for the information retrieval computer. The MeSH are written on the lines at the bottom of the worksheet. The

programs which produce the catalog cards disregard these headings.

The completed worksheet goes to a keypunch operator who prepares a punched card for each line on the worksheet with the exception of the call number. The call number is punched on the first card of the decklet with the delta separating each line of the number. At the Yale Medical Library the same person keypunches who formerly stenciled cards, and the cards are punched more rapidly than they could be stenciled.

Next, a group of punched card decklets are fed into a 1401 computer behind a program having a card punched to set up the number of packs of catalog cards which will be needed. For instance, the Yale Medical Library needs to have one pack of main entry cards to go to the National Union Catalog, two packs of main entries for the Yale University Library, one pack including main entry and all subject and added entries, and finally two packs of shelf list cards—one for the library shelf list and the other for insurance purposes. This first program expands each decklet into the number of catalog cards required to make up the group of packs and writes the data for each catalog card on magnetic tape. A second program prepares the data on tape for sorting and the sort is carried out on an IBM 709. The 709 produces a magnetic tape with catalog card data alphabetized by the filing entry within each pack.

The magnetic tape produced by the 709 is then put on a tape drive of the 1401 and manipulated by a program which, like the first, has a control card. This card can be punched to determine the format of each of four types of headings: (1) topical subject headings, (2) name subject headings, (3) title, short title, and series added entries, and (4) other added entries. The control card can be punched so that any one or all of the headings will appear at the top or bottom of the card, at the left-most position, first indention or second indention, in upper and lower case or all upper case, and in black or red. Since the required heading is printed in the proper location on each card, tracings appear only on the shelf list card. The data are read from the magnetic tape, each line of the card formated in the computer, and the characters recorded on punched cards.

The punched cards produced by the third 1401 program are then placed in an IBM 870 Document Writer. This contrivance is basically a card punch electronically coupled with an electric typewriter. The typewriter which the system uses has eighty-eight characters including sufficient diacritical marks to enable the system to handle twelve different languages in entirety. The cards punched by the 1401 are placed in the card feed of the 870's keypunch whence they travel through the reading head on the keypunch. Information read off is communicated to the typewriter which types out the cards on a continuous card form which was originally designed by Phillip Bagley of the Mitre Corporation working with the Dennison

Manufacturing Co. of Framingham, Massachusetts. As the cards come out of the Document Writer, they are in final form and in alphabetical order for filing.

The Project has on order together with Florida Atlantic University and the University of Toronto Library an upper and lower case printing chain for the IBM 1403 printer, the printer in the 1401 configuration. The three institutions are sharing the cost of developing special characters for such a chain and each is acquiring its own chain. When the chain is available, the third 1401 program will be altered so that the catalog cards will be printed out directly and much more rapidly on the computer. However, it will not be possible to have red headings, but the catalog cards will be produced at the rate of perhaps thirty per minute instead of at the rate of one in somewhat less than a minute on the 870 Document Writer.

At Yale the original decklets of punched cards have been used for the past half year to produce the monthly Bulletin of the Yale Medical Library which lists the accessions of the previous month. The only extra work which involves the catalogers is indicating by a letter in column 66 of the worksheet whether or not that title is to go into the Bulletin. A 1401 computer prepares copy for the Bulletin. The program used in this operation necessarily takes out all special characters and flags since the printing now must be done all in upper case. Formerly it took one person one week of each month to prepare copy for the monthly Bulletin, but it is now prepared in less than thirty minutes of computer time and at a cost in the vicinity of $25. Moreover, the present Bulletin is 50 per cent larger than the earlier manually-prepared issues.

Perhaps the most difficult problem to solve in the computer processing of bibliographic information is the alphabetical sorting of entries. There is no possible way that a computer can be programmed so that it will know when to alphabetize "St." as "saint" and when as "street." The Columbia-Harvard-Yale solution to this problem is somewhat inelegant but appears to work. Other libraries employ essentially the same technique. Whenever the sorting characters of the filing element differ from the actual characters—which is not often —the cataloger writes out the sorting characters on a line at the bottom of the worksheet and assigns that line a special code. In this line numbers and abbreviations are spelled out. The sort program uses this line to alphabetize the entry.

The system is being designed in anticipation of the inclusion of acquisitions and circulation activities in mechanization procedures. Indeed, at Columbia work has already progressed to the trial stage in the use of punched cards in acquisitions with the same cards being subsequently employed in the cataloging process. Similarly, it should be possible to expand information retrieval systems based on the present procedures.

Conclusion

The Columbia-Harvard-Yale Project has developed several principles of small magnitude that are, nevertheless, effective guides past various pitfalls. First of all, flags or signals should not be characters used in routine print-out and should occupy only one column on a punched card. Also, signals which are special signals for the computer should be avoided except in computer output at the very end of the computer operation. Among the flags used by the Project are the "at sign," delta, less-than sign, greater-than sign, lozenge, and group mark, but the group mark appears only in the cards to go into the 870, these cards being punched out at the end of the computer processing. Another principle is that sub-fields within the overall record length should not be a fixed length. The third principle is that the data used for sorting should be as long as possible; in the Columbia-Harvard-Yale Project system the sort control includes the first fifty characters from the filing entry. Another principle is that it is most desirable to have at least one application from as near the beginning of a project as is practicable. The Yale Medical Library's Bulletin has served this function and served it well, for several difficulties were detected in writing the program for the Bulletin production, and in the processing of the cataloging data which goes into the Bulletin. Some of these stumbling blocks would have become major entanglements had they remained undetected until the information retrieval system was being activated.

The Columbia-Harvard-Yale Medical Libraries Project is attempting to design a fast, on-line catalog and journal index information retrieval system together with mechanized catalog card production that can be used as a base to expand to total computerization of only the catalog and indexing of the relatively small, heavily-used core collections of books and journals furnishing upwards of 75 per cent of recorded usage. Finally, the goal of the present system is to increase speed and completeness in supplying users with cataloging and indexing references.

REFERENCES

1. Kilgour, Frederick G. "Recorded Use of Books in the Yale Medical Library," American Documentation, 12:266-269, Oct. 1961.

2. Fleming, Thomas P., and Kilgour, Frederick G. "Moderately and Heavily Used Biomedical Journals," Bulletin of the Medical Library Association, 52:234, Jan. 1964.

3. Kilgour, op. cit., p. 266, and other unpublished findings.

4. Sprenkle, Peter M., and Kilgour, Frederick G. "A Quantitative Study of Characters on Biomedical Catalogue Cards - A Preliminary Investigation," American Documentation. 14:204, July 1963.

5. Dubester, Henry J. "Studies Related to Catalog Problems," Library Quarterly, 34:99 Jan. 1964.

6. Brooks, Benedict, and Kilgour, Frederick G. "Catalog Subject Searches in the Yale Medical Library," College and Research Libraries, 25:482-487, Nov. 1964.

7. Ibid., p. 486.