Determining the Optimal Number of Volumes for a Library's Core Collection

by RICHARD W. TRUESWELL

The concept of the core collection in a large library is not new. The thought behind such a concept is to provide a separation of the more frequently used volumes from those that are infrequently used. There have been, however, difficulties in determining what volumes are to be included within the core collection. It is usually necessary to establish a committee or group of professional people who determine, usually by enumeration, those volumes that should be included in the core collection. Often the results of such a determination merely reflect the opinions of the individuals, and although valid in many cases as desirable reading, the core collection that results may not be a reflection of user requirements.

It was with some of these difficulties in mind that the author undertook the study described in this article. It was felt that a more effective and easier way of determining the bulk of the volumes to be included in the core collection was needed. The core collection also offers the additional advantage that those books not included in the core collection may be stored in some form of compact storage, either physically or in some machine readable form.

The technique described in this article pertains to books of the monograph variety and not to serials, reference materials, or periodicals. It is possible that the approach could be applied to these other types although this has not been done here. It is felt that the core collection concept has possible applications in libraries presently having closed or open stacks. Regardless of the current operating policy of the libraries, the technique described may lead to a way of separating the frequently and the less frequently circulated materials. The latter might be considered as research material and would of necessity still have to be made available in some form. The technique, therefore, should have application in both research-oriented libraries and in circulation-oriented public libraries.

The criterion used in this study is circulation use. For the moment, we are ignoring in-stack or browsing use. There is some evidence of correlation between these factors and circulation (Fussler, 1961). One method of meas-

Revision of a paper presented at the joint meeting of: Operations Research Society of America and the The Institute of Management Sciences, Minneapolis, Minnesota, October 7–9, 1964.
The address of the author is University of Massachusetts, Amherst, Mass.

Libri 1966: vol. 16: no. 1: pp. 49–60

uring circulation use would be to examine the total number of times each book in the library (or in a sample) has circulated (if this information were available) and to rank the books according to the number of circulations, including a factor about the distribution or range of use dates. This would require a book card or record of all transactions for each volume which in general would be rather difficult to obtain. This difficulty occurs primarily because the circulation data for the highly used books fill up a book card rather rapidly and a given book, although used considerably, may have only a few circulation entries on the new book card. Some approaches to measuring circulation and book use have been made by the application of this method but difficulties were encountered in collecting data. Morse of the Massachusetts Institute of Technology (Morse, 1964) has developed a Markov Chain expression for book use in certain subject areas at the MIT Science Library.

As a more readily obtainable statistic, the last circulation date was used in the study described in this article. The last circulation date is defined as the last date stamped on the book card, representing the previous transaction or loan of the book. Actually, the last date appearing on the book card may stand for either the date on which the book was borrowed or the date on which the book was due back in the library, depending upon local library practice. If the due date is stamped on the card, the date the book was borrowed can be determined by subtracting the charge period from the due date.

The method of data collection was as follows: Book cards for each book circulated during the data collection period were examined and the last circulation date was recorded. Thus the data reflected the previous transaction of the book. This information was then plotted in a cumulative frequency distribution as shown in Figure one. The coordinates are "percent of circulation having a last circulation date within the cumulative time period" versus "time period." The data for Figure one were taken at the Technological Institute Library of Northwestern University.

For example, in Figure one we note that approximately ninety percent of the volumes with current circulation sample have last circulation dates within a twelve-month period. This means that ninety percent of the current circulation sample has circulated at least once during the previous twelve-month period.

The same data can be considered in another way and Figure two is a plot of the "percentage of circulation having not previously circulated within the cumulative time period" versus "time period." Thus ten percent of the current circulation sample has not circulated within the last twelve months. This information is of value in determining the cost and work load of in-

stalling a new circulation control system in a library (Trueswell, 1964) such as a punched card system. In such systems, it is frequently proposed that a punched card be used as a book card and that this be used in an appropriate charging device which in turn would record the transactions remotely. Installations have been made of such devices; The University of Southern Illinois, for example, has an IBM 357 Document Originating machine which records the information about the transaction from a punched book card and a patron badge. One difficulty in establishing the new system is the apparent need to prepare punched book cards for every book in the library's holdings. This can be and usually is a very costly process, and in addition, there is the task of matching the newly punched book cards with the books in the stacks or in circulation.

As an alternative to the above method, one might punch a book card during the time the book is on loan if the book does not already have a book card from a previous transaction. Under this system, a punched book card would be prepared during the time the book is on loan and the punched book card would be placed in the book when it is returned. Thus punched book cards would be made only for those books borrowed from the library. We can now relabel the coordinates of Figure two as is done in Figure three: "percentage of day's circulation not already on a punched card system" versus "months after installation of a punched book card system." During the first month under the new system, about one hundred percent of the books presented for loan will require new punched book cards. The figure is actually slightly less than 100% because a few books may be borrowed, returned, and charged again during the same month. During the second month, some books will have been returned and borrowed again, and it will be necessary to prepare book cards for approximately ninety percent of the circulation. This ninety percent represents the portion of circulation for the second month that did not circulate during the first month. As each month goes by, a smaller percentage of the current circulation will have to be converted to the punched card system. Figure three shows that after thirty-six months, we find approximately four or five percent of the circulation requiring a new book card. Using this approach, it is possible to predict the conversion workload. Newly punched book cards are prepared while the book is in circulation, that is only for those books that are in demand. In effect, the books that now have punched book cards could be considered a core collection of books frequently circulated by the users of the library.

Thus referring to Figure one, after thirty-six months of converting to a punched card system, we would have books in the library that would satisfy all but about three percent of the user circulation requirements. Eventually after a given period of time, the curve would go lower, and it would be

possible to determine the one percent level, that is the time required to have the core collection satisfy ninety-nine percent of user requirements.

The approach described requires, as noted, the use of a punched card system. However, with certain assumptions, it is not necessary to use such a punched card system to predict those volumes in the core collection that would satisfy ninety-nine percent of the user circulation requirements. It is possible to generate a theoretical model of a core collection using the data in Figure two. Note that in the previously mentioned punched card system, the books in the core collection have, in addition to punched book cards, a last circulation date within the time period since starting the system. Thus books could be identified as being in the core collection if they had a last circulation date within a period of time that corresponded to the satisfaction of ninety-nine percent of the user requirements.

We might, therefore, predict the size of the core collection (ninety-nine percent user satisfaction) as follows. Assume that we go back in time several years, generate a theoretical system and define the core collection as consisting of books that have circulated at least once in the last x years. We, therefore, go back x years and starting at that point in time, we adopt in our model the procedure of placing a red × on the cover of each book borrowed. As time progresses more and more of the books borrowed will have a red × on the cover. After several years, we will reach a point where ninety-nine percent of the books brought to the desk for circulation will already have red ×'s.

Thus we can generate a model of the core collection by using the data of Figure two. We do this as follows: For the first month of our generating period, we say that one hundred percent of the books will receive red ×'s, i. e. monthly circulation times 1.0. During the next month ninety percent of the books will receive red ×'s and the other ten percent will already have red ×'s from the previous month's transaction. During the third month, approximately fifty percent of the books will require red ×'s. Knowing the monthly circulation for the library we can sum the products of these percentages by the corresponding monthly circulation, and generate the number of volumes having red ×'s without actually marking them. We would do this over the period of time required to reach the ninety-nine percent level of user circulation requirements and in effect determine the number of volumes in the core collection. Applied to the Deering Library at Northwestern University, a core collection was predicted of approximately forty percent of the library's holdings. This result is based on a limited amount of research and is subject to further study. This means we are predicting that approximately forty percent of the volumes in the library represents a core collection that will satisfy the requirements of ninety-nine percent of circulation

Note: This plot is essentially the percentage of current circulation sample having previously circulated within the cumulative time period.

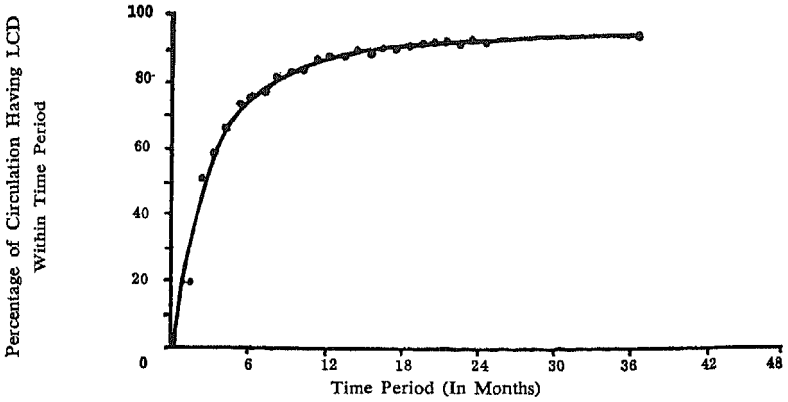

Figure 1. Tech. Inst. Library Northwestern University.

• LCD means last circulation date as defined in text.

Note: This plot is essentially the percentage of the current circulation sample having *not* previously circulated within the cumulative time period.

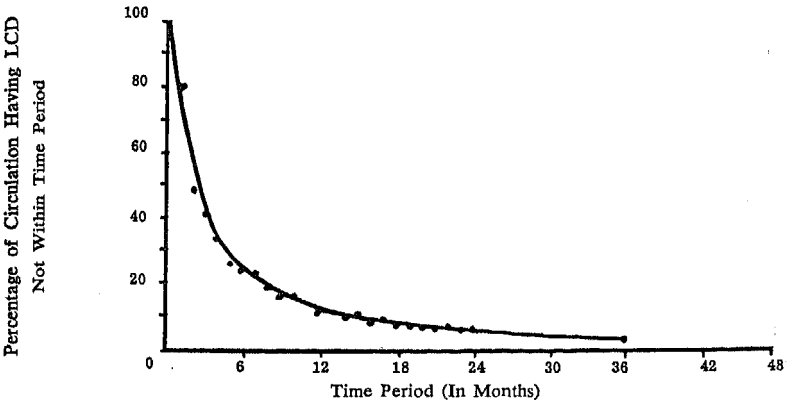

Figure 2. Tech. Inst. Library Northwestern University.

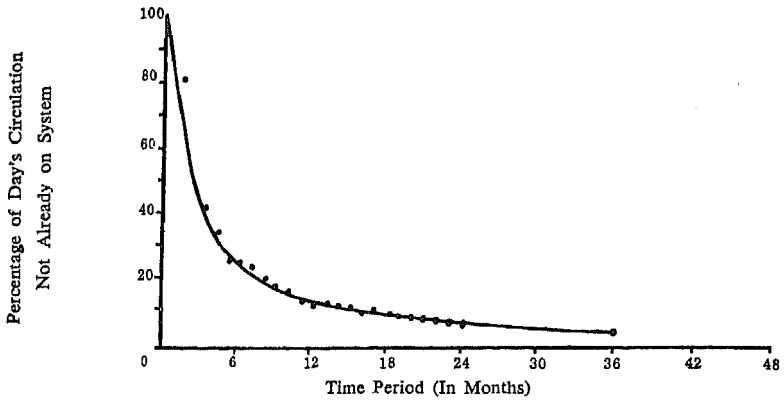

Figure 3. Tech. Inst. Library Northwestern University.

Note: This plot is essentially the percentage of the current circulation sample having *not* previously circulated within the cumulative time period.

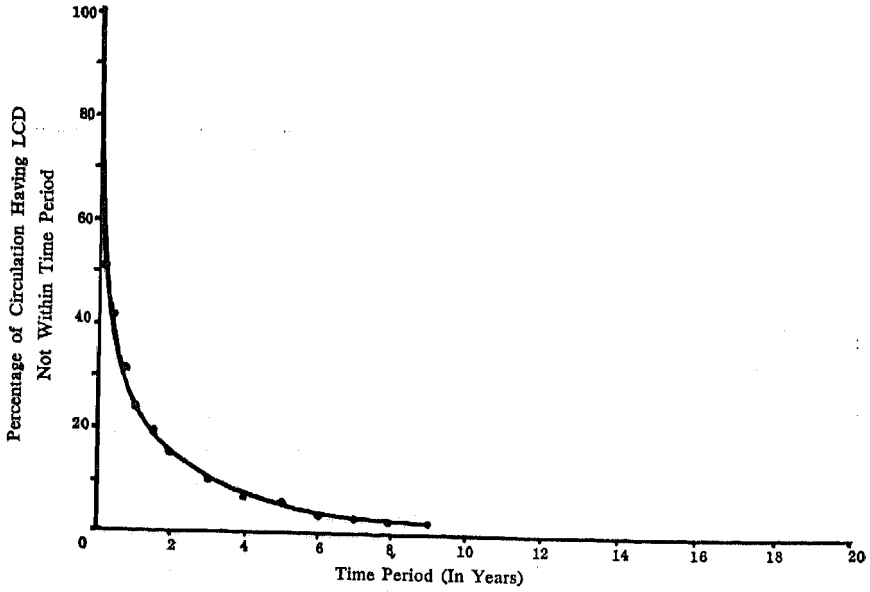

Figure 4. Deering Library Northwestern University 820's.

Note: This plot is essentially the percentage of the current circulation sample having *not* previously circulated within the cumulative time period.

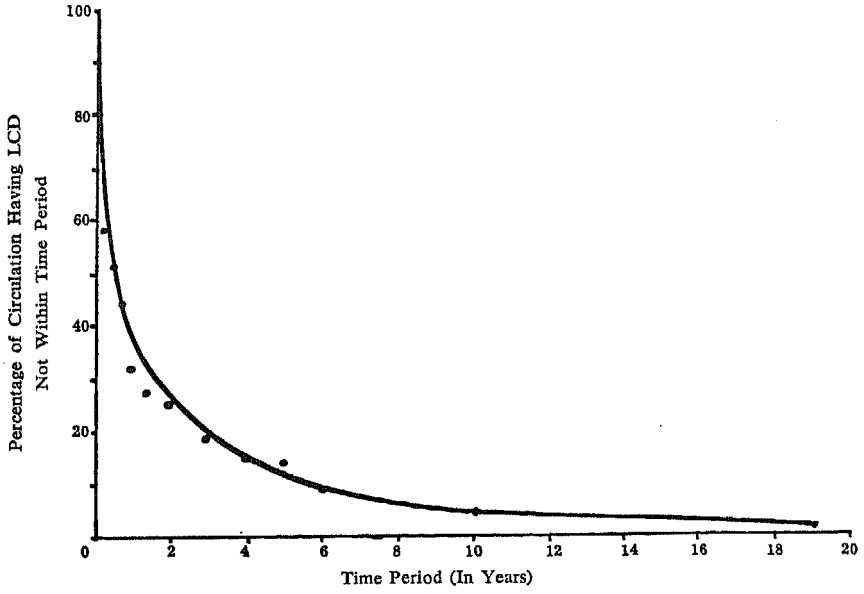

Figure 5. Deering Library Northwestern University 830's.

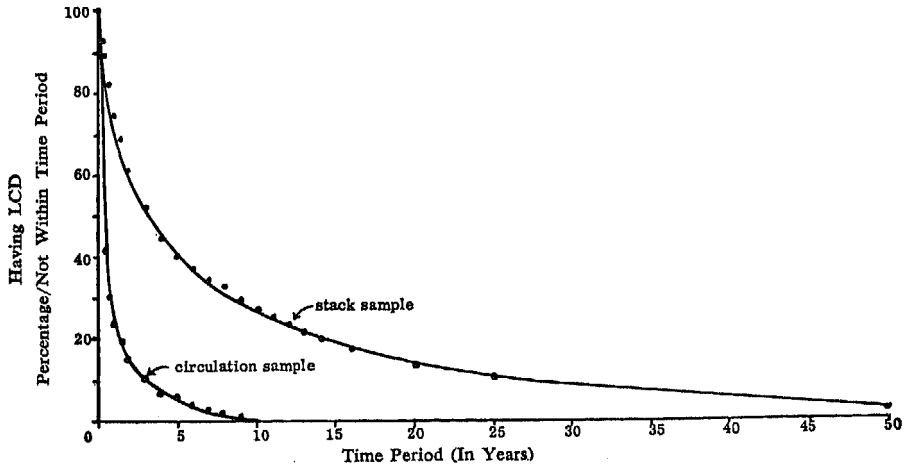

Figure 6. Deering Library Northwestern University 820's.

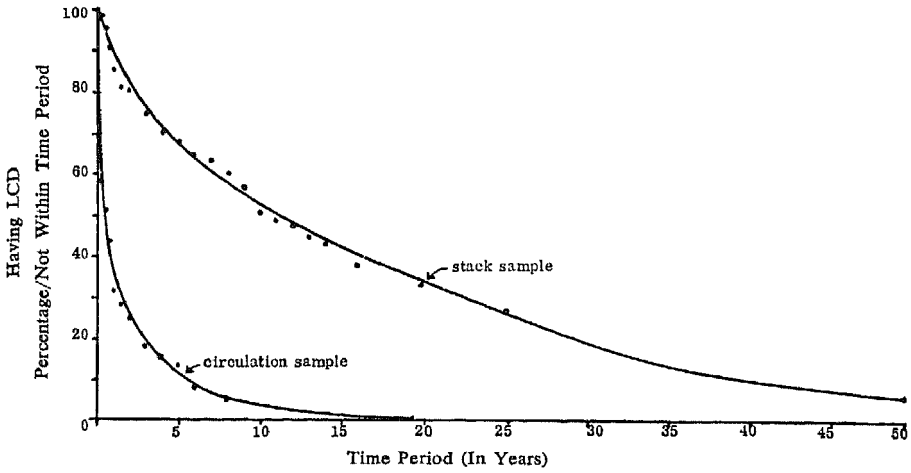

Figure 7. Deering Library Northwestern University 830's.

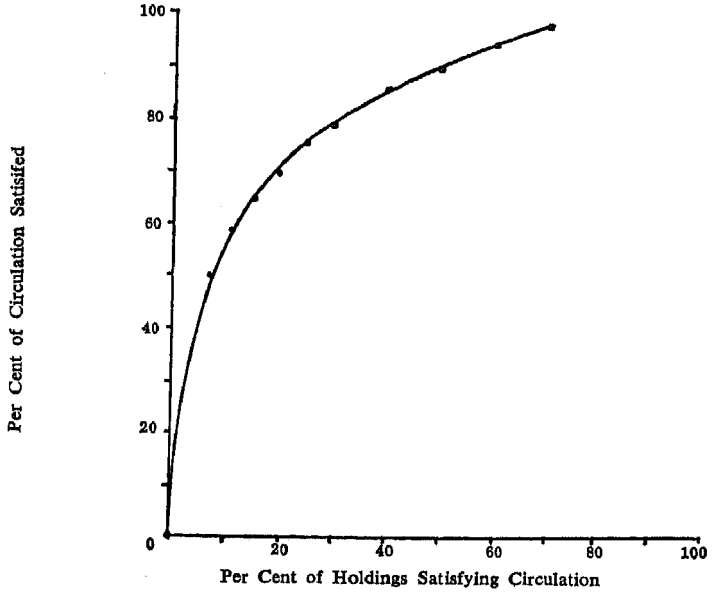

Figure 8. Deering Library Northwestern University 820's.

needs. We can identify these core collection books in the stacks because their last circulation date is within the time period defined by the ninety-nine percentile point of Figure one. To select or identify these books in the holdings it would simply be necessary to examine the book card, note the last date, and if it were within the time period corresponding to ninety-nine percent of circulation requirements, we could consider the book as being within the core collection. A rather general assumption made in this approach is that the circulation pattern as measured today is typical of the circulation pattern five, ten, or twelve years in the past. This may not be an unreasonable assumption based on the original and subsequent data.

Therefore, we may be able to predict from circulation data the size of the core collection, that is, the percentage of the total number of holdings that will satisfy ninety-nine percent of the circulation requirements. We can do this with the last circulation date which in turn offers a very easily defined criterion for selecting or identifying these core collection books in the stacks. This has possibilities for the library problem of stack thinning or weeding of books. Thus one could say that books in the stacks having a last circulation date prior to a given cutoff point would be removed and placed in a limited access storage area or in an inter-library center. The remaining books should satisfy ninety-nine percent of the user requirements. It might be preferable to establish closer satisfaction limits such as ninety-nine and five-tenths percent. This technique also allows us to predict by the method described above the size of the collection required to satisfy a given percentage of the circulation. Thus we have an analytical way of predicting the number of books that would be necessary to satisfy any given percentage of user circulation requirements.

As a rather loosely defined check on validity, the method was applied to two subsets of the library's holdings at the Deering Library of Northwestern University. Last circulation dates for current circulation were recorded and plotted for the 820's and the 830's, English literature and German literature respectively. Figures four and five are plots of the "percentage of circulation having not circulated during the cumulative time period" versus "time period." Through the method outlined above the core collection was predicted at sixty-eight and five-tenths percent and sixty-five percent for the 820's and 830's respectively. These high percentages may be a result of the subject area. The ninety-nine percent cutoff dates were approximately 1955 for the 820's and 1945 for the 830's. Thus all books having a last circulation date in 1955 or later should satisfy over ninety-nine percent of circulation requirements for the 820's.

As a test of the validity of this approach a sample was taken of the books in the stacks and the last circulation date recorded for these books. In this

case, the last date on the book card is considered to be the last circulation date. The sample size was approximately one and seven-tenths of one percent in a holding of approximately 32,000 books in the 820's and one and six-tenths of one percent of 23,000 books in the 830's. The sampling procedure was to select the second book from each shelf within the 820's and with the 830's and to record the last circulation date. This information is plotted on Figures six and seven. From these plots it is possible at the 1955 and 1945 cutoff dates to determine what percentage of the stack sample has a last circulation date within that time period. It was seventy percent for the 820's and sixty-six percent for the 830's. These compare favorably with the predicted core collection sizes of sixty-eight and five-tenths percent and sixty-five percent of the holdings. As a validation technique this leaves a great deal to be desired because of the method of data collection, the limited sample size, and some of the assumptions involved. However, the closeness of the predicted and actual (as reflected by the stack holdings sample) core collections indicates a need for further exploration. This is currently being done at the University Library and the Mt. Holyoke College Library. One apparently difficult problem is the design of an improved stack sampling procedure.

With this approach we are in effect saying that there is a predictable optimal number of volumes for a library's core collection that will satisfy a given percent of user circulation requirements. In this example ninety-nine percent was used, but the satisfaction level must be determined by the administration of the library. Still to be resolved is in-stack use and it is obvious that some special purpose books regardless of last circulation date must be left for certain categories of in-stack use. Further studies are needed to apply some sort of marginal analysis to the ratio of unit increase of circulation satisfaction to increase of stack holdings core collection size.

Figure eight is a plot of the data for the 820's. Combining the data from the above stack holdings sample and the corresponding circulation sample, we obtain a plot (Figure eight) of "Percentage of Circulation Satisfied" versus "Percentage of Holdings Satisfying Circulation." It is interesting to note that approximately fifty percent of current circulation is drawn from about seven percent of the stack holdings. Thus if given the desired percentage of circulation requirements to satisfy (such as ninety-nine percent) we can predict the necessary or optimal size of the holdings. It appears from preliminary findings that it will be possible to predict this figure solely from circulation data and that the core collection size may be a function of the quantity of books circulated. Note Figure eight is for a small subset of the holdings and that the core collection defined at ninety-nine percent of circu-