

Ben-Ami Lipetz
Head, Research Department
Yale University Library
New Haven, Connecticut

A QUANTITATIVE STUDY OF CATALOG USE

Among people who are concerned with the management of libraries, it is now almost universally accepted that the traditional manual card catalog must sooner or later be replaced by an on-line computerized catalog of some sort. This is accepted almost as an article of faith; there is almost never any questioning or disputing of its inevitability. I have no intention of questioning or disputing its inevitability in this paper; but there *are* questions regarding the computerizing of library catalogs which ought to, and indeed do, trouble conscientious library managers. These are the crucial questions of *how* to computerize and *when* to computerize. The work I will report on was prompted mainly by concern with these questions.

The notion of computerized library catalogs has been with us for many years. Computerized library catalogs were, in fact, set up at libraries here and there as far back as a dozen or more years—which means during the era of the first generation of large computers. They operated in batch mode, of course, and on rather restricted document collections; but they operated. And, as the years have passed, the catalogs or indexes of more and more document collections have been committed to computers.

The appeal of computers is obvious. There is, first of all, the speed and accuracy with which they can perform basic functions, such as filing in of new data, compiling statistics, transcribing data for human reading, and transmitting data for use by other machines. There is the ability of computers to perform complex logical searches, at least on pre-designated elements of the stored data. And, very important, there is now the ability of computers to serve numerous users simultaneously at diverse locations, by means of time-shared terminals, thus obviating the need for the users to be in physical attendance at the catalog storage location.

Nevertheless, the use of computerized catalogs today is still highly restricted. It tends to be confined to applications where the document

collection is relatively small, where the catalog information is very simple and limited, where there is an unusually high value attached to rapid or remote catalog service, and where large computing capacity is already available for purposes unrelated to the library. This is because of the negative aspects of computers: the high cost of converting existing catalogs to machine-readable form, the high cost of computers, the unavailability of really large-scale rapid-access memory, and the limited reasoning capacity of existing computer programs.

Because the negative aspects of catalog computerization have been particularly serious for the very large general purpose library, of which the Yale University Library is a prominent example, there has long been a tendency for management in these libraries to regard catalog computerization as probably inevitable but clearly remote. Therefore, it could be dismissed from serious attention. That attitude can no longer be justified. Recent events have indicated that the time when conversion will be practical for large libraries may not be so remote after all—indeed may be only a few years away. Events contributing to this change have included: the steady growth of rapid memory capacity of computers, the falling cost of computing capacity, the improvement of equipment and of programs for remote-terminal time-sharing, the establishment of the MARC system to make new catalog data available in machine-readable form at low cost, the development of regional library groups which have the potential to make existing catalog data available in machine-readable form at low cost through cooperative effort, and the development of standard machine formats which will make data interchange possible and economical.

So the decision on *when* to computerize the catalog of the very large libraries may soon become a matter of tactics, rather than strategy. At this point, the question of *how* to computerize the very large catalog is in need of urgent attention. The natural tendency, of course, would be to create a computerized catalog in the image of the existing manual card catalog, preserving all features of present-day catalog content and file organization. Tradition tends to be very strong among catalogers in large libraries. Yet tradition must be resisted, or at least questioned. Existing card catalogs are not necessarily the ultimate in human wisdom and ingenuity. Certainly some of the features in their design are attributable to the inherent limitations of cards and card drawers. There is no need to perpetuate the weaknesses of present catalogs in future catalogs. Before computerizing our catalogs, it would be very desirable for people in large libraries to take a hard look at what they would want from an ideal catalog, and then to see what sort of design in a computerized catalog would most closely approach that ideal. The key question is "What do we want from a library catalog?" One of our research projects at the Yale University Library is endeavoring to provide an answer to this question.

The approach we have taken is very direct. We are trying to learn what a future catalog should be by studying, quantitatively, what our library patrons are trying, successfully or otherwise, to get out of our present catalog. This study is supported, in part, by the Office of Education.¹ The basic idea of a catalog use study is not at all new. There are quite a few such studies

already reported in the literature; most are master's thesis projects. Unfortunately, almost none of them inspire any confidence in the results because of gross deficiencies in experimental design, sample size, or both. Our own study was carefully designed to anticipate and obviate any foreseeable criticism. It is a two-year study which began in late 1967 and will be completed late in 1969.

Our study attempts to find out what our users want from a catalog, but it does not stop there. It also attempts to find out the extent to which our present card catalog satisfies the needs of the users. And, furthermore, it attempts to find out whether there are practical methods, manual or mechanized, to satisfy needs that are not now being met. Thus, even if we do not computerize our catalog for many years, the study should be useful in perfecting our traditional card catalog in the interim.

Because the study is still in progress, I am unable to give any final results. The collection of data is more or less complete, but many of the projected analyses of the data have not yet been accomplished. Therefore, I will confine myself mainly to describing how the study has been carried out and stating what we should be able to learn from it. I will state some of our preliminary findings, but I must emphasize that all figures to be quoted here are based on incomplete data and are subject to possible revision in our final report.

The public catalog of the Yale University Library is located in the main entry hall of the Sterling Memorial Library. It contains some seven million cards, housed in some 7,000 file drawers. It is a single-alphabet catalog. It contains full catalog card sets for the more than three million volumes housed in Sterling Memorial Library and only main-entry cards for the two million volumes housed in other libraries at Yale. Since the numerous school and departmental libraries have more complete catalogs for their respective collections, users of the main catalog are generally in search of books that are housed in the collection at Sterling Memorial Library. The stacks of Sterling Memorial Library are open to all Yale faculty and students, and to a rather large number of authorized outside users of the Library. The catalog, as you can imagine, takes up a rather large area, and is the scene of constant activity throughout the hundred hours a week that the Library is normally open.

A catalog search is basically a word-matching procedure. The searcher seeks to match some known clue, which is commonly a word or a phrase or a name, against the headings in the file; if he succeeds in finding a file item which matches his clue, he can expect to find some associated information in the file (e.g., a call number) which is the object of his search. In a nutshell, the aims of our study are to find out: 1) what clues the catalog users possess when they begin a catalog search, 2) how well our present catalog responds to (i.e., matches) the clues that the user brings, and 3) whether the responsiveness of the catalog might be improved through some change(s) in catalog design.

We are finding out what clues the users bring to their catalog searches through interviews with a representative sample of catalog users. The interviewees are approached at the instant that they reach for a catalog drawer to begin a search; they are asked a number of carefully worked out questions

designed to elicit very precisely what the searcher is trying to accomplish through the catalog and what information he has brought to the search. We also collect background information about the searchers (but we do not ask for their names). The interviewers are all trained to follow a standard interview outline. At the beginning, the questions are very general and nondirective, to avoid leading of the subject. ("Could you please tell me what you were about to do here at the catalog when I interrupted you?") Only after the subject has had ample opportunity to say whatever he wants to, in his own way, do the questions become more direct and specific. Clues available to the searcher are recorded in full detail. If he carries them in the form of a printed bibliography or as handwritten notes, they are photocopied by the interviewer. If he carries them in his mind, they are transcribed by the interviewer, taking pains to determine and preserve the searcher's personal version of the spelling of author names and unusual words.

An average interview takes about ten minutes; but it may take as little as two minutes or more than fifteen minutes, depending on the nature of the searcher's problem and the amount of information which he brings to the search. When the interview is concluded, the subject is left alone to carry out his search, but is observed discreetly from a distance. The catalog drawer which he uses is noted. When he appears to have finished, he is approached again and asked if he was successful. If so, the interviewer notes the call number(s) of the item(s) which satisfied the search. Later on, we can examine the catalog cards for these call numbers, and we can examine the books themselves, to see how well the existing catalog matched, and how well it *might* have matched, the clues which the user had when he began his search. This follow-up activity to examine the catalog cards and the books they represent is considerably less glamorous and exciting than face-to-face interviewing, but it is every bit as important to our study and it actually takes more time and effort than the interviews.

The interview program, concluded only this month, was conducted over a full calendar year. We gathered data from some 2,000 interviews. The catalog users were cooperative beyond our wildest dreams. Fewer than 1 percent of the people approached refused to be interviewed—generally it was because they had to rush off to a class. Most interview subjects were delighted to be asked about their activities and eager to respond to all questions. Because of the accidents of random sampling, some people were interviewed two or three times during the year, and they still remained fully cooperative. To put it simply, the library users were very happy to learn that somebody actually cared about them.

At this point, I should explain how the interviewees were selected in order to provide a representative sample. Long before we began any interviewing, we had already begun collecting gross statistics on observed traffic in the catalog area and on various activities which occur in the catalog area. There happen to be five different entrances to our catalog area. By counting the number of people entering through each doorway at various times on different days, we constructed a preliminary projection of expected traffic by day of week and time of day. We then decided how large an interview sample we wanted (at least 1 percent). To get this, we worked out a precise interview

schedule for each doorway in which the interview times and dates are in proportion to the expected traffic. Thus, each of our interviewers (two full time, with a third available to help in emergencies) was assigned to be at a specific doorway at a specific hour and minute; the first catalog user who entered through that doorway before a fixed interval elapsed was the person to be interviewed. Then the interviewer would go on to his or her next assignment, which would generally be a different doorway. Assignments were spaced to allow reasonable time for completion of one interview before starting the watch for the next one. Sometimes no one would come through the doorway during the scheduled interval and so there was no interview; however, this is a random event which does not affect the value of the sampling technique.

What *can* affect the value of the sampling technique we used is seasonal variation in traffic pattern. Therefore, we continued the gross traffic counting program for more than a year in order to detect such variations. Differences between the observed pattern and the preliminary projection on which the interview scheduling was based will be compensated for by applying appropriate weighting factors to the results of interviews conducted at different times and times of the year, so as to make the statistical results entirely representative of observed traffic.

Having provided a background on the study, we can now discuss the ultimate question: What do we expect to get out of the study that can do anyone some good?

Let us start with our gross observations of traffic and other activities in the catalog area. We can plot traffic by time of day, day of the week, and time of the academic year, and can thus produce a clear picture of expected volume and variation of catalog use. This can be of immediate value to the library administration—particularly in planning for the provision of reference assistance, and in scheduling of catalog maintenance—and it can be important in helping to determine the peak simultaneous access capacity which must be provided in any future computerized catalog facility. Of course, librarians already know quite a lot about traffic patterns from long years of experience, so we do not expect any earth-shaking revelations from this particular result of the study.

Other aspects of our observation of catalog traffic are more novel. We have collected much information on the amount of time which users spend at the catalog. What proportion of users spends one minute per use, two minutes, five minutes, fifteen minutes, etc.? From this we can tell what kind of queuing to expect in the catalog area, not only with the present level of activity but with increased future activity as our user population grows. This should give us a sort of yardstick against which to measure the performance of contemplated computerized systems, to see whether they are worthy of serious consideration. We have collected extensive data on the number of catalog cards which users actually look at in the course of a catalog search, and on the number of references which they tend to copy from the catalog cards during a search. These data may or may not prove useful in furthering our understanding of the catalog user.

We have collected data on precisely which catalog drawers were consulted by searchers at times when traffic was being observed. This should tell

us whether all catalog drawers tend to be consulted equally or whether there are high-activity areas and low-activity areas in the catalog. This will have an important bearing on the level of queuing to be expected in a computerized catalog for any given memory access arrangement. All of these results will be based on very simple objective observation of the catalog area—merely counting people, and timing people, counting their hand motions in writing down references or flipping cards, and noting and recording catalog drawer numbers. These measurements require no interviewing at all.

The interview data will yield a wealth of potentially useful results. For one thing, they will add some useful details to our picture of catalog traffic. Since we record the academic status of persons interviewed, we will be able to describe separate traffic patterns for students, faculty, staff, and outsiders—and see whether they differ significantly. We will be able to do the same for newcomers to the University (students or faculty), as opposed to old-timers. We will be able to do the same for different departmental affiliations or areas of study.

Secondly, the interview data will yield quantitative insights into what it is that catalog users are seeking, and will tell us whether different categories of users tend to bring different types of problems to the catalog. Fairly early in the study, it was observed that the objectives of catalog searches tend to fall into four rather distinct categories. One category, the “document search,” is where the user has a specific published work in mind and is using the catalog in order to locate a copy of that work. A second category, imperfectly called the “author search,” is where the user knows of a source of publication—usually but not necessarily an author or corporate author—and wants to find what works are available from that source (e.g., what are some books by Thomas Mann?). A third category is the “subject search,” where the user seeks to identify publications on a known abstract topic. The fourth category is the “bibliographic search,” where the user has no intention of borrowing any book, but is only interested in finding the catalog card for a known publication so that he may get some specific information from the catalog card itself (e.g., to complete the bibliographic citations in a paper he is writing).

The document search is by far the most common. Analysis of a portion of our data suggests that about 75 percent of the uses of our catalog are for the purpose of locating a specific known publication (which, to our surprise, is almost always available in our collection). The other three use categories are more or less equally divided among the remaining 25 percent.

These results are preliminary, of course. Even if they were final, they would be suspect, however. There is a strong possibility or presumption that the actions of a library user are shaped by the nature of the catalog facility that is available to him. Do library users tend to accommodate themselves to what our catalog can do very well, such as locate known works? We are getting an answer to this from a very innocuous sounding but highly revealing question that we ask in our interviews. It reveals that a significant number of the document searches performed at the catalog are really subject searches in disguise. Presumably there would be a smaller proportion of overt document searches if our library catalogs were better suited for subject searching. We

hope to get at the question of accommodation in yet another way, by looking for any difference in searching patterns between newcomers to the University and old-timers, or between newcomers at the beginning of the school year and later in the school year (when they have had a chance to adjust to reality).

A third, and also very important, type of result expected from our interview data will be the compilations and analysis of the search clues which catalog users possess at the start of their searches. By comparing the clues with the information available in the retrieved catalog cards and the documents they represent, we can assess the accuracy of the clues. For example, we can tell how often the catalog users start out with author names or titles that are inaccurate or misspelled, and we can analyze the frequency of different types of inaccuracies. This is fairly important for designing card catalogs, but it could be crucial for computerized catalogs. Computers make no concessions to misspelling unless designers take great pains to program around their punctilious and unyielding accuracy. The data collected from the interview program can be used to test the effectiveness of computer algorithms which are intended to produce matches despite inaccurate input from the searcher. We have already made quantitative evaluations of the effectiveness of two different data compression algorithms described in the literature by testing them on real data from our interview program.

Last, but by no means least, we will be able to use data from the interviews and from the retrieved catalog cards, and from the works corresponding to those catalog cards, to seek means to improve the quality and efficiency of cataloging rules and catalog structure. We will be able to say whether there are categories of data included on cards which are rarely wanted, or categories which are frequently wanted but rarely included. We will be able to throw some light on the wisdom of dividing a catalog into sections segregated by date of publication or by other unconventional distinctions. We should learn whether machine-like subject indexing which makes use of the key words occurring in book titles, or prefaces, or chapter headings, or indexes, etc., would match actual user clues as well as our conventional subject indexing (based on authority lists) does now, or whether it would be even better.

Of course, we are only studying one library at one university. Will our results be useful to people outside of Yale? We believe that they will be; but I would caution in advance against blind acceptance of any of our results as universally relevant. There are bound to be local differences among libraries and universities. To find out how significant these differences can be, it would be prudent to conduct studies similar to ours at a considerable number of large libraries of different kinds. I was very gratified to learn recently that a study of this type will soon be undertaken at the Library of Congress. But more studies are needed. I hope that they will not be long in coming since the computers are nearly upon us. With all the effort that has been going into research and development work on how to computerize catalogs, it would be nice to have more guidance on how to do it right.

Reference

1. Lipetz, Ben-Ami, and Stangl, Peter. "User Clues in Initiating Searches in a Large Library Catalog." In *American Society for Information Science, Proceedings* (Annual meeting, October 20-24, 1968, Columbus, Ohio). Vol. 5. New York, Greenwood Publishing Corporation, 1968, pp. 137-139.