RICHARD DE GENNARO

A Strategy for the Conversion Of Research Library Catalog To Machine Readable Form

This paper describes in very general terms a strategy for converting the retrospective catalogs of the nation's research libraries into machine readable form. The method envisages a class-by-class conversion and printing out in main entry order of the shelflist of the Library of Congress. The larger libraries would compare their shelflists against these lists adding their location symbols and unique titles to the master machine record and pulling from the master record machine readable catalog copy for their own holdings in each class. The resulting augmented LC master record would become a kind of national union catalog in machine readable form.

UNTIL A FEW YEARS AGO librarians were rather skeptical about the technical and economic feasibility of converting the massive catalogs of multi-million volume research libraries into machine readable form. The view was generally held that while current input into these catalogs could be computerized the problem of converting the retrospective file into machine readable form was so enormous that future technological advances would have to be awaited before it could be undertaken. The science and medical librarians, citing the rapid obsolescence of their literature, concentrated their efforts where the most immediate payoff was available—in computerizing the record for current acquisitions. While librarians of humanistic collections could not completely turn their backs on the bibliographical heritage of the past, many of them were prepared to settle either for maintaining the retrospective catalog in its traditional format or for reproducing it in book form by offset photography. Thus, the computer would give us a powerful handle on current acquisitions but could not relate them to the total record. These views are beginning to change.

Many librarians are now becoming less pessimistic about the technical feasibility of converting mass catalogs. Practical experience in conversion has been acquired, photocomposition devices and print chains with upper- and lower-case and diacritical marks are available, keyboarding equipment has been improved, and new online keyboarding devices and techniques are being introduced. The extremely high cost of converting mass catalogs still remains a chief obstacle, but even here the picture is beginning to change and there is reason for optimism. With the federal government's growing interest in research libraries it seems reasonable to hope that funds may eventually be made available to convert to machineable form certain library catalogs or bibliographical records of national importance. Since the National Union

Mr. De Gennaro is Associate University Librarian for Systems Development, Harvard University.

Catalog is the largest and most comprehensive and therefore potentially the most useful record available, attention has been focused on it as the most likely candidate for conversion. One study has already been made of the feasibility of such a project and the techniques by which it might be accomplished, and a committee of the Association of Research Libraries is presently exploring the problem.

While there are many advantages to starting with the NUC there are also some serious disadvantages. It is an alphabetical file of fifteen million cards, all of which would have to be converted before much real use could be made of it, since a portion of an alphabet is of limited utility. The conversion of fifteen million entries complete with notes and added entries is a formidable undertaking and would require several years and a considerable investment of editorial effort, which might spell the death of the project if allowed to get out of control. The end product, in spite of its tremendous usefulness, would still be incomplete and inaccurate by the standards that are used to judge the catalogs of large research libraries. Advances in computer and communication technology will tend to make these standards even less acceptable in the future than they are now.

The purpose of this brief paper is to suggest as a possible alternative a method of converting the retrospective catalogs of the nation's research libraries and eventually creating a national union catalog in machine readable form as a byproduct of that effort. The strategy would be to avoid a frontal assault on a multi-million card dictionary catalog and a straight A-to-Z conversion, and to divide this massive single conversion project into a series of smaller and more manageable projects, each of which would utilize and build on the experience gained in the previous ones, generating useful outputs as the effort progresses. A similar approach is being used with considerable success in the Widener library shelflist conversion project at Harvard.[1]

The starting point for this conversion effort would be the shelflist of the Library of Congress, a bibliographical record that is relatively accurate and up to date. Since it is a unit-card shelflist, each entry is complete with notes, subject, and added entries, and once converted to machine form would serve as the basic record from which all other secondary records could be generated by computer. What is being suggested here is that the LC shelflist might be converted class-by-class to form the basis for constructing a master machineable bibliographical record in LC classification order and alphabetically in main entry order within each class. Other libraries could compare their shelflists against these basic LC lists, adding their own location symbols and unique titles to the master file and pulling from it machineable catalog copy for their own holdings in each class. The resulting augmented LC master record would eventually become an accurate and serviceable national union catalog in machine readable form. The problem is to develop strategies and techniques to facilitate not only the conversion of the basic LC file, but also for comparing and adding the new titles and locations for the titles held by each succeeding library as it enters the system and for enabling a library to extract catalog entries for its own holdings from the record.

If we can assume that a MARC-type standardized format for inputting bibliographical data into a system will have been developed and adopted within the next few years, then one could envisage a project being refunded to re-create LC's catalog in machine readable form using a class-by-class shelflist approach. Initially, a subdivision of a science class

[1] Richard De Gennaro, "A Computer Produced Shelflist," *CRL*, XXVI (July 1965), 311-15, 353.

such as physics or geology, and a part of a history or literature class might be selected as pilot projects to test assumptions and develop techniques. For the sake of discussion, however, let us suppose that LC started its conversion with the E-F or American history class. Upon completion of the conversion of the entire class or a logical segment of it such as U.S. history, a printout would be produced listing the entries alphabetically by main entry. The American or U.S. history holdings of another research library, that of a university for example, could then be compared with this list. One possible way of doing this would be to search the entries of the university library's American history shelflist against this alphabetical main entry printout. Each time a match was encountered, the local call number would be noted on the main entry printout. At the end of this comparison, the local library would have an annotated printout accounting for a large proportion of the titles in its collection. It could then pull those entries held in common with LC from the master tape by simply keyboarding the LC card number (or a special machine-assigned identification number) together with its own call number and other local information, and having the computer create a new local tape combining the LC entries with the local ones.

The entries present in the university library's shelflist that were not present in the LC list could be duplicated by photography and converted, using the standard input format that had been used for the LC list. This could be done at the university library, but it might be preferable to send them to a central facility for further searching and conversion and for entry into both the master LC file and the university library file. These entries would also have to be assigned LC class numbers. The university library would then have in its tape file the bibliographical information it needs to re-create its shelflist and catalog and to produce other listings either in hard copy or machine form. The central master file would now be augmented to certain titles in the local library that were not held by LC along with locations for all the titles held by the local library. Several problems remain, such as reconstructing the syndetic apparatus or the complex of cross references in the catalog, and accounting for the titles in American history held by the university library but classified elsewhere for local reasons such as in reference or rare books collections, etc. The latter problem would be the responsibility of the local library while the former one would have to be dealt with by the central authority.

The same techniques could be applied to each successive segment of the LC shelflist as the conversion effort progressed. As classes were completed the computer could sort them into a single main entry list and eventually re-create a version of the dictionary catalog. After the contents of several major collections had been compared with and added to the augmented master LC file, the comparison and conversion effort of each additional library would be made increasingly easier because the number of titles not found in the master file would be decreasing. The comparison procedure would be easiest for those libraries which are classified according to the LC system because there would be a relatively close correlation of scope in the two shelflists. For this reason it might be better if the pilot comparison effort took place in such libraries rather than in those which do not use LC.

This problem of scope of shelflists could well be one of the most serious objections to the strategy being suggested. Many of the older libraries with rich collections, such as the New York public library, Harvard, Yale, etc., have classification systems which may be difficult to correlate effectively with LC's classes.

This difficulty might not be as serious as it may seem at first glance if one bears in mind that the comparison or searching is done in a printout of a class of the LC shelflist that has been sorted by computer into main entry alphabetical order rather than the list in classified order. Thus the American literature class of a library with its own scheme would be searched against the equivalent part of the LC schedule arranged by main entry. Nevertheless, the problem remains and should not be minimized. On the other hand, the catalogs of these libraries, because of their uniqueness, age, size, and complexity, are going to present serious problems of compatibility in any future national bibliographical system based on computers and sooner or later these problems will have to be tackled and solved.

The techniques outlined for comparing, searching, annotating, and adding to files are here described in terms of today's familiar technology for the sake of clarity. In an actual project the whole process would presumably be considerably streamlined by the use of advanced online computer technology with visual display consoles, mass random access storage, and sophisticated means of communication. Thus, instead of actually producing a computer printout of the segment of the LC shelflist to be used for comparison, it could be in random access storage and accessible through a cathode-ray tube or visual display console. The local card shelflist entries would be searched in sequence by calling for the appropriate part of the alphabet on the console display unit. Each time a match was encountered a symbol would be added to the machine record together with the local call number and any other necessary local information. This would greatly facilitate the entire process and reduce keyboarding to a minimum.

The ultimate goal of the effort is to create in machine readable form an inventory of the holdings of the nation's major research libraries. The method suggested looks toward building this record in a gradual, orderly, and economical manner. Each bibliographical record would be in a standardized format, and the master file would be the basic record which would be put into mass random-access storage for online long-distance consultation when these techniques become economically feasible in the future. The file would serve as a data bank from which extracts of various types and for various uses could be drawn. While it is theoretically possible to produce the entire contents of this file periodically in printed form, this would be extremely expensive and probable unnecessary. It might be far more useful to produce a large variety of shorter and more specialized lists based on class, subject, language, date of publication, etc.

Some of the principal advantages of this conversion strategy are summarized below.

1. The master record is based on a relatively accurate and solid foundation, *i.e.*, the current inventory records of LC and the participating libraries—their shelflists.
2. It is a gradual process which can be changed, developed, and improved with experience. It is flexible, unlike the single frontal assault required for an A-to-Z conversion of fifteen million entries.
3. It would not only give LC a tremendous impetus in its total systems effort but would also make possible a parallel development for the entire research library community by removing the chief bottleneck—conversion of the retrospective file.
4. The cost and effort of keyboarding a bibliographical entry would only occur once and in a favorable environment.
5. The funding of this single but seg-

mented effort might be facilitated because the subject approach would create interest and enthusiasm among the various segments of the research community including user groups as well as funding agencies. The E-F classes would interest historians, scientists would be eager to see the Q class done, etc.

6. The strategy and techniques could be inexpensively and meaningfully tested and costed in one or more pilot projects, such as the conversion of the Physics or Geology subdivision of the Science class, and a segment of a history or literature class. A decision to proceed with, modify, or abandon the strategy could be made on the basis of the experience and information generated in these pilot efforts.
7. There is no reason why, after suitable pilot projects, several classes could not be converted simultaneously. The work could be geographically decentralized by duplicating portions of the LC shelflist by photography and having the conversion work done outside of Washington, where space and personnel might be more readily available.
8. Useful lists of all kinds, such as shelflists, classed catalogs, subject bibliographies, chronological, alphabetical, and language listings, etc., could be created as each portion of the list is completed. There is no need to wait until the entire Library of Congress shelflist has been converted and augmented to obtain products of this kind.
9. Eventually the complex of cross references that tie a catalog together could be reproduced and all classes merged by computer into a single dictionary catalog in machine form.

The conversion of the present NUC or the re-creation of it in a new form is obviously an extremely complex and costly undertaking and one which has tremendous implications for the future development of libraries. This brief paper is not meant to give pat answers as to how it should be done nor does it pretend to be a detailed and carefully constructed master plan. The most that can be said for it is that it offers an idea for a strategy which may be worth considering along with others that are being discussed.

Whatever the strategy, the job of converting the massive retrospective record can and should be done, but it need be done only once in a standard format providing for full access. These millions of bibliographical entries were keyboarded several times before they came to rest as printed LC cards, and it is not unreasonable to suggest that they be keyboarded once more in machineable form to put the nation's research libraries firmly into the computer age. ▪▪