A National Bibliographic Data Base in Machine-Readable Form: Progress and Prospects

RICHARD DE GENNARO

Librarians are gaining experience with localized computer systems, they are struggling with the problem of how to integrate the use of MARC data into their technical processing operations, and they are contemplating the intriguing possibilities of a national library network. As they do so, they are becoming increasingly aware of the necessity for converting their retrospective catalog records to machine-readable form which will be the foundation of the complex automated systems that the future requires. This article will address the question of why retrospective conversion is necessary, and it will attempt to show that it is a feasible objective by citing significant research and recent, continuing large-scale conversion projects. It will explore the means by which retrospective conversion might be accomplished as well as cost and time projections. Emphasis will be placed on the Library of Congress' current and forthcoming activities in this area because they are of particular significance in the creation of any national bibliographic data base in machine-readable form.

Why is retrospective conversion necessary? Most librarians have accepted the idea that conversion of current and future catalog records to machine-readable form is both a desirable and a necessary step in the automation process. Conversion of retrospective records, however, has always appeared to be such a formidable undertaking that few have been willing to face it. The case against retrospective conversion has been made by science and medical libraries on the valid grounds that most of the use of their collections is based on...

Richard De Gennaro is Senior Associate University Librarian, Harvard University.

APRIL, 1970
recent or current materials and that time will solve the problem. The same rationale does not hold true for general research libraries because the older materials in their collections are used more heavily and they continually acquire substantial quantities of retrospective materials. If the entire retrospective bibliographical record is not converted, these libraries will always be obliged to maintain their old manual systems along with their machine systems, and they will never get the maximum benefit from automation.

Libraries cannot seriously begin to design and implement "total" or "integrated" systems until they come to grips with and solve the problem of converting their retrospective catalogs into a machine-readable data base. This data base would become the foundation for subsystems for various operations, such as circulation control, searching, cataloging and catalog maintenance, and interlibrary loan service. It would also provide the means of generating the management statistics and information that are needed to improve library operations. Moreover, this comprehensive data base is obviously the foundation upon which networks must be built if the network concept is to become a reality. Actually, few librarians would question the desirability of having their entire catalogs in machine-readable form; they merely cannot believe that conversion can be accomplished, or accomplished at a reasonable cost. Let us discuss feasibility first and costs later.

Six years ago, no major research library had even begun in any serious manner to convert its retrospective catalog records to machine-readable form. There was no standard bibliographical format; the coding and printing of upper and lower case and diacritical marks was still poorly understood and difficult to accomplish with the available equipment. Almost no one had gained any significant experience in converting large files of complex bibliographical data and few librarians would have known how to use the products of such conversion if they had been available. Today the picture is drastically different. A considerable body of experience has been accumulated and a great deal of solid research and development has been done in the conversion of mass catalog files. Harvard's Widener Library shelflist conversion project was one of the early entries in the field. Routine conversion of its limited-entry shelflist in a local format was begun in 1965, and well over a half million entries of the estimated total of 1.6 million have been converted to date. The Meyer undergraduate library at Stanford and the Ontario New Universities Library Project at Toronto, although more limited in scope, yielded
A National Bibliographic Data Base

much valuable experience in conversion techniques, The Universities of Toronto, State University of New York-Buffalo, and Syracuse, among others, have converted portions of their shelflists or catalogs. The Library of Congress has developed the MARC II format and complex input systems, and has converted over 100,000 entries into machine-readable form in the two MARC formats. In England, the University of Newcastle-upon-Tyne has converted its entire catalog and Oxford University has embarked upon a major project to convert the pre-1920 Bodleian catalog using OCR (optical character reader) typewriters as the input medium and a format recognition program to lessen the manual editorial burden.

The Institute of Library Research at Berkeley has done an outstanding job of research, development, and publication on the problems and techniques of mass bibliographic file conversion, principally in connection with its project to develop the design and specifications for a Technical Processing Center for the California State Library. Its publications set a standard rarely equalled in the library automation field. In 1970 it embarked on one of the most ambitious, well-planned, and technically complex conversion projects that has been attempted to date, i.e., the conversion into MARC II structure and subsequent publication in book form of the estimated 900,000 records that form the 1963-67 supplement to the University of California's printed catalog. The project will use OCR-font typewriters and a highly-developed automatic field recognition system to facilitate input and minimize manual tagging and editing. The completion of this project in 1971 will mark the beginning of a new era in file conversion and the experience gained should be of considerable value to LC's Project RECON and other conversion projects.

The body of experience and knowledge gained in all these projects, together with the many improvements and developments in hardware and software that have taken place in the last few years, clearly indicate that the state of the art is now sufficiently advanced to support the large-scale conversion of complete bibliographical entries in the MARC II format.

While experience has shown mass conversion to be technically feasible at this time, it has also demonstrated that the cost is extremely high—in a range of one to two dollars per entry. Input keyboarding is only one of the costs and by no means the major one. Tagging the elements and editing the copy require the greatest effort and are the most difficult to accomplish since they demand personnel with train-
ing and experience in bibliographical work, and such persons are in extremely short supply. Computer and other machine costs are also significant, as well as project direction, administration, space, and other overhead costs. Another major category of expense which is frequently overlooked or underestimated is the very high cost of software development—systems design, programming, and program maintenance. Expense is not the only problem; it is difficult to find and hold the highly-skilled persons who are needed to do this complex technical work.

The issue is no longer whether the retrospective record can or should be converted, but rather how it should be converted and at what cost. Here we come to a critical point. Unless some over-all national plan for centralized conversion of a standard record in a standard format is developed and implemented in the near future, many libraries will begin (as many have already begun) to convert their own catalogs on an individual basis. The result will be the repetitive creation of expensive local conversion systems producing non-standard or sub-standard machine-readable entries. The combined cost of these separate efforts will exceed substantially the cost of a single centralized conversion effort which would provide a common bibliographical data base in the standard MARC II format from which libraries could draw a significant percentage of their catalog entries.

The RECON Working Task Force under the chairmanship of Henriette D. Avram has recently completed and published a comprehensive study entitled, Conversion of Retrospective Catalog Records to Machine-Readable Form; A Study of the Feasibility of a National Bibliographic Service 5 (hereafter referred to as the “RECON Study”). This excellent study, which was underwritten by the Council on Library Resources, Inc., has in one stroke raised the prospect of a national centralized retrospective conversion effort from the discussion and speculation stage to a level of systematic analysis and concrete planning. Since the RECON Study is now the basic document on LC-based conversion, this paper will necessarily draw upon and summarize many of its ideas and conclusions. For the serious student of retrospective conversion, library automation, or bibliographic networks, no summary can take the place of the full text of that report.

Most of the arguments that can be made in support of LC’s centralized cataloging and card distribution service apply equally well to centralized record conversion. Indeed, if the MARC distribution service is a logical extension of LC’s current card distribution service,
then creation and distribution of retrospective catalog data in machine-readable form is an equally logical extension of that service as well as of the MARC service itself. Conversion of its retrospective record is essential for LC's own future internal automation as well as for the card distribution service. Most of the experience, development work, and computer software that has been created for the MARC service is directly applicable to an LC-based retrospective conversion project. In short, conversion of LC records in the MARC format by LC is clearly the most reasonable and economical course to pursue; this is the major conclusion of the authors of the RECON Study.

So far, we have argued that retrospective conversion is necessary, that it can be done, and that it should be done centrally at LC with LC records as the starting point. Three other major considerations remain to be discussed: 1) the catalog or data base at LC which would be the most appropriate, 2) the principal technical and cost considerations, and 3) the over-all method of implementing the project.

With regard to the selection of the catalog or master data base to be converted, the RECON Study cites three important factors to be considered. There should be a high rate of duplication between materials covered by the data base selected and the collections of other libraries. The entries should have a high degree of accuracy and completeness, and certain types of entries should be excluded, such as serials and non-book materials. With these and other factors in mind there are only four catalogs which can be seriously considered for conversion: 1) the National Union Catalog, (NUC), 2) the LC shelflist, 3) the LC official catalog, and 4) the LC card division record set (a catalog of printed cards in LC card number and, therefore, roughly chronological sequence).

The NUC seems at first glance to be a likely candidate because of its size and comprehensiveness. In addition to its four million LC records, it contains seven million records that represent analytics, dissertations, local publications, foreign-language titles, etc., which are not on LC cards. These, however, are titles which are not held by many libraries and these entries do not come up to the standards of accuracy and completeness that are desirable in a master data base. Therefore the NUC was eliminated from further consideration.

Drawing on experience gained from converting the Harvard shelf-list, the LC shelflist was this author's candidate for conversion in an article in College & Research Libraries published in 1967. There would be many advantages to approaching conversion through the
LC shelflist if the shelflist were a reasonably accurate, up-to-date, and legible record. When the RECON Working Task Force considered the LC shelflist, it found that this file “contains a mixture of temporary, incomplete, and printed records with essentially no corrective changes beyond revision or updating LC class and book number. Nor are the cards legible enough to be microfilmed to provide a readable guide to locating the master records in the Official Catalog.” Because of these deficiencies the shelflist was eliminated from further consideration.

From the point of view of up-to-dateness, completeness, and accuracy, the LC official catalog would appear to be the most desirable candidate for a master data base. However, there are serious difficulties in using it directly for this purpose. The name portion of the catalog contains some twelve million cards of all kinds, and the task of searching out the four million discrete records produced since 1898 would be formidable. These records frequently contain numerous additions and corrections and would be difficult to use as a source document for first conversion. For these reasons the RECON Working Task Force recommends using the LC card division record set for first conversion and then bringing up to date the resulting record after comparing it with the master entry in the official catalog.

The card division record set consists of a master copy of the latest revised reprint of every LC printed card, arranged by card series and, within each series, by card number. The chronological nature of this catalog, its subdivision by series, and its legibility are potent arguments in favor of making it the starting point for conversion. The chief disadvantage is that not all changes in a catalog entry are cause for reprinting and therefore this record will have to be searched, compared with the official catalog entry and corrected to insure the level of accuracy and quality that a machine-readable data base of this nature requires.

Even with this disadvantage, the record set is still the prime candidate for first conversion because, like the shelflist, it has one of the characteristics most essential in a data base for a mass catalog conversion project: it is a large file that can be divided into a series of significant subsets that can be tackled and completed singly and used effectively as they are completed. The importance of this feature should not be underestimated. The Working Task Force recommends that the record set be divided by language categories (a tedious manual process) and that these categories be divided by time spans
of card series according to the table below. This rearranged file would be microfilmed and a copy produced for the project; the original file would then be reconstructed for the card division.

<table>
<thead>
<tr>
<th>Category</th>
<th>Time span of card series</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. English language</td>
<td>1960-March 1969</td>
<td>386,000</td>
</tr>
<tr>
<td>2. Romance and German languages</td>
<td>1960-June 1970</td>
<td>381,000</td>
</tr>
<tr>
<td>3. English language</td>
<td>1898-1959</td>
<td>1,728,000</td>
</tr>
<tr>
<td>4. Other Roman alphabet languages</td>
<td>1960-June 1971</td>
<td>137,000</td>
</tr>
<tr>
<td>Nonbook materials</td>
<td>1960-June 1971</td>
<td>157,000</td>
</tr>
<tr>
<td>5. Slavic languages</td>
<td>1960-June 1972</td>
<td>225,000</td>
</tr>
<tr>
<td>6. Other non-Roman alphabet</td>
<td>1960-June 1973</td>
<td>256,000</td>
</tr>
<tr>
<td>languages</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Romance and German languages</td>
<td>1898-1959</td>
<td>698,000</td>
</tr>
<tr>
<td>8. All remaining catalog records</td>
<td>1898-1959</td>
<td>682,000</td>
</tr>
</tbody>
</table>

This table clearly demonstrates the advantage of being able to divide a large file into a series of significant segments to which priorities based on various considerations can be assigned. Such a strategy is a reverse chronological conversion sequence with priority assigned to the categories in greatest demand (and with the fewest problems). It will tie in nicely with the recommended rapid phasing-in of additional categories of current catalog data to be produced by the MARC distribution service. In no language category would retrospective conversion begin until the current records in that category were being produced by the MARC distribution service.

The RECON Study recommends that an initial conversion effort be made with English-language monograph records issued from 1960 to the beginning date of the current MARC service. This would be followed by conversion of Romance- and German-language monographs issued from 1960 to their beginning date in the MARC service (projected for June 1970). Both should be completed within four years. The third category would be English-language monograph records issued from 1898-1959. The conversion of other categories might follow the sequence of the table above or might be modified in the light of experience gained with the first three.
One of the very difficult and important problems that the study recognizes, but rightly reserves for further investigation, is how best to obtain standardized bibliographical records for items that are not now in the LC record set. In other words, how can the master database be expanded to become a truly national union catalog or database in machine-readable form? Just as the problem of retrospective conversion had to await certain hardware developments, the establishment of a standardized format, and the accumulation of some practical experience with conversion, so the larger problem of how to create and maintain a true national union catalog in machine-readable form must await additional hardware developments. It must also await the experience and knowledge that will be gained by the conversion, organization, and manipulation of a substantial body of retrospective and current records. The Working Task Force wisely recognizes this and recommends moving rapidly toward the design and implementation of a conversion system capable of handling a non-trivial pilot project of English-language records from 1960-68.

Retrospective conversion is no longer a technical problem awaiting the development of better keyboard input equipment or even the long-heralded advent of direct-read optical character recognition (OCR) equipment. The last few years have seen the development of several input keying devices including the magnetic tape inscriber and the OCR-font typewriter which are well suited to the mass conversion of bibliographic data. One of the most surprising findings of the RECON Study is that the cost of conversion by direct-read OCR equipment, when it is perfected, would be slightly more than conversion of unedited records by magnetic tape inscriber when all systems costs are considered. The reason, according to the RECON Study, is that these devices will not be capable of reading non-Roman characters, diacritical marks, and other special characters. The machine would have to be programmed to reject records with an excessive number of unreadable characters and these records would have to be manually keyboarded. It was estimated that the number of records rejected might be as high as 10 percent. The cost of keyboarding corrections and unreadable records, added to the relatively high estimated cost of the OCR equipment itself, makes this alternative much less attractive than one might think. In any case, the importance of input keyboarding and the selection of input devices have been given more attention in the past than they deserved. These two factors together account for only 16 to 20 percent of the total unit conversion.
cost in the LC environment and whether one device is slightly more efficient than another is a relatively minor matter. The selection of an input device in a conversion project may well be made on the basis of criteria other than cost.

Thus the need to await further breakthroughs on input equipment before undertaking large-scale conversion of bibliographical records has disappeared. We have the hardware for conversion, and we either have or know how to create the necessary software. The cost per record will probably never be much less than it is now, since 85 to 95 percent of the costs can be categorized as manpower, and only 5 to 15 percent as machine costs, and it is a fact that manpower costs are rising and machine costs are falling. It should be stressed that these cost ratios apply to the LC environment where the data has to be edited and corrected to the highest standards possible in order that it be acceptable for a national data base. In local environments where less complex procedures could be adopted and lower standards of accuracy could be tolerated, the total unit conversion cost might be significantly less and the manpower-to-machine cost ratio might be more nearly even. Indeed, a greater utilization of automatic format recognition programs might yield a more favorable result even in the LC environment. In any event, it appears that the chief obstacles to conversion are no longer technical; they are financial, political, and managerial.

While detailed consideration of technical and cost factors is not appropriate for this general overview, a brief review of these factors is essential for a basic understanding of how the records can be converted and at what cost. The RECON Study considered six input devices: 1) keypunch, 2) paper tape typewriter, 3) magnetic type inscriber, 4) on-line typewriter, 5) OCR-font typewriter, and 6) direct-read OCR (still under development). The keypunch and paper tape typewriter were eliminated as being technically unsuitable. The on-line typewriter and OCR-font typewriter were eliminated after a cost analysis showed them to be more expensive than the magnetic tape inscriber and direct-read OCR. The magnetic tape inscriber was deemed to be the most appropriate and least expensive device now practicable, while the direct-read OCR, although it is not fully developed, was retained as a possibility on the assumption that it might be used for some portion of the file when it is perfected. The elimination of the OCR-font typewriter as being too expensive is probably justified in the LC environment, but there is considerable evidence to
suggest that it may well be the most appropriate input device in other environments.

The manpower and machine unit costs of twenty technical alternative methods were analyzed and the four best ones were selected for detailed consideration. They were: 1) direct-read OCR (assuming its perfection in a few years) using a format recognition program, 2) unedited copy using a tape inscriber and a format recognition program, 3) partially-edited copy using a tape inscriber and a format recognition program, and 4) fully-edited copy using a tape inscriber. The resulting copy would in all cases be manually compared against the LC official catalog and corrected. The total cost per entry of the four alternatives ranged from a high of $1.87 to a low of $1.51 in the third alternative with 94 percent of the cost ascribable to manpower and 6 percent to machine costs. Of the manpower costs in this alternative, $0.52 is for partial editing which in this context includes partial coding prior to input, post-editing to correct and augment the output of the format recognition programs, and editing of new data derived from comparing the interim records against the LC official catalog. This study confirms the conclusion that human editing in cataloging conversion projects is one of the most important cost factors and the trained personnel required are in short supply. That fact accounts for the intense interest that has developed in writing and utilizing automatic format recognition programs in such centers of conversion as Oxford, Berkeley, and the Library of Congress.

The format recognition program envisioned by the RECON Study analyzes the data in a partially pre-edited machine-readable record and automatically assigns tags or content designators and coded information which make explicit what is implicit in the textual information on the catalog card. Partial editing means that the records have been pre-processed by a human editor who has supplied some cues which increase the accuracy and reliability of the format recognition program. The utilization of these techniques reduces significantly the cost and difficulty of the conversion process by putting the burden of tagging and coding on the machine where it belongs. The Bodleian Pre-1920 Catalogue Project is successfully using a format recognition program that was initially written by John Jolliffe for the British Museum general catalogue. The Institute of Library Research has successfully tested the ILR Automatic Field Recognition System, on several significant samples. The goal of the System is to achieve a full MARC II record without any pre- or post-editing/tagging. The
computer recognition algorithms work with the existing format of the catalog card, and have no special input requirements." The system was developed for use in the University of California Union Catalog Supplement Project and promises to reduce substantially the amount of human editing necessary and thus reduce the time and the cost of the project. This approach, if it proves feasible in actual operations, could be used for inputting current as well as retrospective records and might even be adaptable to the LC environment.

Applying the unit cost of the least expensive RECON Study conversion method, i.e., $1.51, the 386,000 English-language records in the record set from 1960 to March 1969 could be converted for an estimated $581,000. The cost of converting the 1,728,000 English-language records from 1898-1959 would be $2,602,000. To convert the estimated total of 2,114,000 English-language records would cost nearly $3,200,000. Since this is approximately half of the entire LC record set, the cost of converting the whole set would be on the order of $7 million.

The cost of systems design and software for a conversion system is estimated at $569,000 and is constant regardless of the number of records to be converted. The cost of hardware is based on the total number of records to be converted over a period of years and is therefore an extremely complex factor. However, for purposes of this discussion, the conversion, storage, and manipulation of the four million entries in the record set would require a two-shift computer system costing an estimated $7 million over an eight-year period. This system would support more than mere conversion operations; it would provide equipment for a national bibliographical service.

No matter how it is viewed, the total cost of retrospective conversion, including the cost of the systems design and software to accomplish it and the cost of the hardware necessary to support a national bibliographical service, is formidable. However, as has been said earlier, these costs seem far more reasonable and acceptable when viewed against the alternative, which is for each library to attempt to do its own conversion, a course which would produce a rich profusion of non-standard and incompatible records and systems at an enormous aggregate cost.

Another element in the cost picture which deserves more emphasis than is given by the RECON Study is the fact that full-scale mechanization of the card distribution service, upon which LC has embarked, is dependent upon conversion of major portions of the retrospective record. There is no doubt that this dual use of the machine-readable
data base along with dual use of the hardware and software would make the costs a good deal more acceptable. It should also be emphasized that a machine-readable data base will be a valuable property. It can be used to produce many kinds of marketable services and products which could contribute significantly to the support of the initial conversion as well as to the maintenance of the bibliographical system that will be based on it.

The Working Task Force recommended that the MARC distribution service be expanded as rapidly as possible to include all current cataloging done by LC in order to arrest, or at least slow, the growth of the retrospective record. It suggested that the cost of this expansion, along with some of the cost of retrospective conversion, might be budgeted as part of LC's regular operations, supplemented by grants and transferred funds. The research and development costs might well come from grants from private and governmental agencies with an interest in libraries.

Fortunately, planning for retrospective conversion did not end with the completion and publication of the RECON Study. Continuing the momentum that had been generated, the Library of Congress applied for and received an Officer's Grant of $25,000 from the Council on Library Resources, Inc., to implement the first phase of a RECON Pilot Project. This grant was made to convert the 85,000 English-language monograph titles cataloged during 1968 and those English-language titles cataloged in 1969 but not included in the MARC distribution service.13 The conversion will provide a practical situation to test and study the various conversion techniques as well as the concepts and techniques of partial-editing and format recognition as outlined in the RECON Study so that the best methods for future conversion efforts can be determined. In addition, a representative sample of from five to ten thousand older titles in English and other Roman-alphabet languages will be drawn from the record set for further detailed analysis and testing.

Thus, the important task of retrospective conversion is moving from the study stage to active experimentation in the form of a pilot project. Experience with library automation has shown that this method of proceeding in stages is the one most likely to be successful in accomplishing a difficult and complex task. The development of the MARC II format and distribution service followed a similar pattern with excellent results.

The RECON Study leaves many technical, organizational, and pro-
cedural questions unanswered. How can such a massive file be organized and maintained efficiently and effectively? How can the data base be expanded into a true national bibliographical system with locations? How can other libraries draw entries from the data base for their own use and add their unique holdings to it? How will serial entries be handled? In what forms and on what financial basis will the data be distributed to libraries as well as to firms desiring to exploit its commercial possibilities?

Some of these questions will be answered by the first phase of the RECON Pilot Project; others will be answered only in later phases. In any case, despite the numerous problems and islands of ignorance that remain, a significant beginning has been made on the task of converting LC's retrospective bibliographical file to machine-readable form, and cautious optimism is in order.

References

8. Ibid., p. 31.
9. Ibid., p. 97.
10. Ibid., p. 99.