MICHAEL GORMAN
Head
Bibliographical Standards Office
The British Library
London, England

The Economics of Catalog Conversion

It will come as no surprise that, as an employee of a national library and one associated with the revision of Anglo-American Cataloging Rules (AACR), I did not agree with everything said by Mr. Kilgour in his opening paper. On one thing I am, however, in full agreement with him. I agree completely with his statement that the card catalog is dead. Many among us believe that it is dying, a few even believe that it is still alive and kicking. It is, however, as dead as a doornail. What I shall address in this paper is, in fact, the decent and economical disposal of the remains.

The title of this paper is as daunting to the person who must deliver it as it probably is to those who must listen to it. The topic of catalog conversion is a scattered one; it is something that has been carried out in recent years in a variety of different institutions and in a number of different ways. Major studies, such as the RECON and CONSER studies, have been made but no single method has emerged as the best and most economical. As a result, although the literature of the subject is extensive, the hard statistical and economic data contained within that literature are conflicting and, of course, are constantly being falsified by technological advancements on the one hand, and the ever-present inflation in the Western world on the other.

What I wish to do in this paper, therefore, is to sketch the processes involved in catalog conversion, and secondly to try to indicate the relative economic factors which apply to the various processes and strategies involved. I shall focus on one particular conversion project—carried out within the British Library—for which I had some responsibility, and I trust that this project will yield relevant information to librarians wishing to convert catalogs in the United States.

One of the most important aspects of the application of automation to library processing is the conversion of one's existing bibliographic files. In fact, a major constraint in many people's minds on the application of electronic data processing has been the query that it raises about the necessity for, and the problems of, conversion of existing files. I would like, first of all, to define what we mean by conversion. Conversion is the transfer of the bibliographic records of a library or group of libraries from manual to machine-readable form. This process can be enormously complex, expensive, and, in fact, daunting in its implications for the librarian and for the persons responsible for the establishment of an automated library system. The chief cause of the anxiety which many people feel about library catalog conversion is the expense of the process. For example, in 1970, John Jolliffe estimated that the conversion of the British Museum catalog to machine-readable form would cost £750,000.[1] In the intervening years, that figure has almost certainly doubled. This means that the cost of converting what is admittedly one of the world's largest and finest catalogs is now something in excess of $3,500,000. Similar figures would undoubtedly apply to the conversion of any similarly large bibliographic files.

The first task that arises, therefore, is to examine the benefits that one might hope to achieve from the expenditure of such a considerable sum of money. There are, broadly speaking, two benefits. The first is the fact that a converted file will provide the original data base for one's automated library system, to which can be added records derived from current cataloging and processing, or from centrally provided machine records, with the aim of providing an integrated file. The second benefit is the indisputable fact that machine-readable systems provide a better service to the users of the library than do manual systems. Even in these economically difficult times, it is thus possible to argue that one should endeavor to have one's existing files converted to machine-readable form.

Two fundamental points about catalog conversion shoud be noted before examining the details of the process. The first is that the most economical way of carrying out the process of catalog conversion will be to base it on the prior conversion of extensive regional, national and international files, and the widespread availability of records from these files. In other words, it is economic lunacy for every library to embark on its own individual, independent, and self-financed conversion program. Why is this? The answer lies in the simple fact that if one were to convert, let us say, the files of the two or three largest libraries in the English-speaking countries, one would have a reservoir of converted records which would account for a very high percentage of the items held by individual libraries within those countries. The conversion process would therefore not be replicated in a potentially nonstandard way, but would be done once in a standard way which would provide high quality machine-readable records, at a much lower cost than a library could achieve by developing its own conversion process. For example, the back files of the British National Bibliography (BNB) (some 400,000 records relating to British monograph publications since 1950) have been converted to machine-readable form and are to be made available to libraries wishing to establish machine-readable catalogs. In Britain, the authorities or bodies responsible for local public libraries, have been amalgamated and reorganized to form fewer and larger library systems. Some estimates of the percentage of the holdings of libraries covered by the BNB-converted file and another linked project are as high as 90 percent.[2] I will return to this project later, but wish now to establish the fact that this relatively small conversion operation (costing approximately £75,000 or $150,000) has made possible the distribution of MARC-compatible records at a much lower unit cost than would have been the case had these libraries started their own conversion projects. In fact, independent conversion may be simply impossible for many library systems which lack the economic and human resources required for such a project.

The second fundamental point is that before conversion of one's file is begun, using the maximum number of externally available previously converted records, one should establish a current automated cataloging and processing system. I believe it is very poor strategy to try to establish a current processing system and to carry out the conversion at the same time—or to attempt to begin the conversion without a complete definition of the current system—its structure, methods and aims. To begin a conversion project before the current processing system is operating will have at least two bad consequences: (1) converted records will run the risk of being unsuitable for the finally developed current system, and (2) the conversion project will absorb some of the financial and human resources which should be concentrated on the current system.

The various aspects of the conversion process involve: (1) the selection of data to be converted, (2) the coding of that data to make it acceptable to the machine system, and (3) the transfer of the coded data from human-readable to machine-readable form. Consider first the selection of data. A primary objective should be to establish for which items held by your library records are already available in machine-readable form. For libraries in the United States there are two main sources for these records: (1) the existing MARC data base built up by the Library of Congress, and (2) the data bases which have evolved from the establishment of cooperative systems and networks. For example, the OCLC system is capable of providing records for a large number of more recent publications, and the CONSER project will provide records for more than 250,000 serial publications. The unit cost of a MARC record supplied by the British Library is twenty cents. Comparative figures for the independent production of records are notoriously difficult to obtain and interpret, but in most circumstances the unit cost to the library will be much higher. Furthermore, such a simple cash comparison does not take into account either the standard of the centrally obtained record (as compared to the locally produced one), or the staff hours involved in complete local production. The use of these externally provided records will, of course, dictate one of the standards necessary to establish for a conversion project, that is, the format in which the records will be received will be the MARC format or at least will be in a MARC-compatible format.

It is, of course possible to convert a MARC record into a nonstandard format, and there may in some instances seem to be advantages for an individual library in constructing a format which is of particular use to them. Nevertheless, I would strongly urge that all conversion projects and, by implication, the ongoing cataloging and processing automated systems, be based on the use of standard formats. I would urge this for one very simple reason: any automatic transfer from one format to another can at best only *maintain* the level of analysis and definition contained within the format from which one is converting, and more often involves a loss of definition and analysis. One can always create less out of more but can rarely create more out of less. Therefore, the evolving bibliographic order will demand, as seems inevitable, the use of a standard format for the exchange of records and, one would hope, for the use of records within local, regional, national, and international systems. There must be a mechanism for determining which items in one's library have externally produced records available. Ideally, this would be by a standard numbering system, such as identifying the Library of Congress card number or the International Standard Book Number for an item; if such numbers are not available, a search strategy will have to be evolved, depending almost certainly on the author and title of the work. These access points are difficult to determine and use in a fully standard way, and it will therefore be necessary to address the problem of identifying a record and establishing its identity with the item in one's library. Various strategies have been advanced in connection with the MARC/RECON project,[3] and it seems likely that the strategies established by the OCLC network users will be of great significance for libraries wishing to assess records which are outside the commonly used numbering systems.

We now come to the consideration of the conversion of entries relating to those items for which no externally produced records are available. As stated above, the question of the format in which these records will be created has been settled. That is, it should be an agreed basis for the conversion program that records be created in the MARC format. Other standards, however, are not as immediately apparent. Does one attempt to transfer all information held in one's current catalog? Does one attempt to change that information to ensure that it conforms to currently accepted national and international standards? To answer the first question, it is not necessary for the library to convert all the information which is held. It might be desirable, in an abstract way, that our future machine systems hold very full bibliographic information, and this is recommended in the report of the RECON pilot project.[4] However, practicality and economics seem to dictate that it will be necessary in many cases to establish a *minimum* set of data for the items held in one's library, and that the criteria for the establishment of this minimum set of data should be: (1) which elements provide the most important access to the whole record, and (2) which elements are necessary for the identification and adequate description of an item. The traditional catalog entry has been a very full one. To some extent, it harks back to a previous era of bibliographic description and contains much information which many would regard as not germane to the purpose of a modern local library catalog, even those catalogs which represent very large collections. I would like to suggest a minimum set of data to be included in all converted records:

1. the class number and/or call number,
2. the various author headings,
3. the subject heading(s),
4. the uniform (filing) title (if present),
5. the title proper (as defined by the ISBD(M)),
6. the edition statement,
7. the publisher and date,
8. a truncated form of the physical description of the item, and
9. a short series statement.

These elements will, I believe, provide enough information to access the record and will adequately describe it to the user once the record has been found.

By this stage in the conversion process, one has identified the items for which externally produced records can be obtained, and has decided on a set of data to be recorded for the items to be converted locally. The next stage is to edit and code the entries in the manually produced catalog for their transference to machine-readable form. A full set of information must be used for this purpose. In most North American libraries the most suitable entry is the shelflist entry because this gives not only the descriptive details, heading and call number, but also the other headings for the item being described. This is necessary because a machine-readable record is more extensive than any one catalog entry. It is worthwhile to point out also that the use of the shelflist is frequently less disruptive of ordinary catalog use than is the use of other catalog entries.

Once the shelflist entries for the items to be converted have been identified, the next process is to code the information. This can be done in a variety of ways, and I will now describe and comment on the way in which this process was carried out for the BNB-conversion mentioned earlier. The basic entries were given to a group of persons consisting of professional librarians and library school students, to be edited on a "cottage industry" basis and according to a strictly prescribed set of rules. These rules were: (1) to select those entries which fell within the scope of the project (for example, repetitive entries for continuations were excluded), (2) to cross out from the entry any element not considered to be relevant in the conversion process, and (3) to add distinctive punctuation and numerical and alphabetic coding to indicate the class of information to be converted. In this application, "cottage industry" means that the people did this work at home at a fixed price for a certain number of entries. At that time it was £2.00 (about $5.00 in those days) for 100 entries—though it would now be more. The advantage of the system is a rather simple—and to a certain extent brutal—financial advantage: work done at home is not subject to "overheads." In his study of the costs of conversion of bibliographic records to machine-readable form, Duchesne established that in 1971 the overheads were equivalent to 100 percent of salary paid, and it is unlikely that in the intervening years this percentage has decreased.[5] In other words, to organize work in a factory, library, etc., will cost more than $2.00 in real costs for every $1.00 of salary paid. The advantages, therefore, for the library and for the persons thus temporarily employed are, I think, obvious.

So far as the coding is concerned it is sufficient to say that enough information was added to the records to produce a MARC-compatible record. For example, a simple numerical code was added by the side of the heading to indicate the class of personal or corporate heading to which it belonged. It was not necessary to code the heading further, because in a sophisticated and homogeneous file like BNB the same typographic conventions had been used over the years so that, for instance, once one has coded a heading as being a personal name, an element in roman type preceded by a comma must be a forename, and an element in italic type preceded by a comma must be an epithet. Thus, an element of automatic format recognition (AFR) can be combined with the precoded information, even in situations where complete use of that technique is not possible. This combination of partial pretagging and AFR was also used in the RECON pilot project.[6] In the rest of the entry, each element was separated from the next by a standard punctuation mark. This particular coding exercise was done directly onto photocopies of the printed BNB entries. It would be possible, of course, to photocopy cards directly from a shelflist card catalog. Another method would be to devise a standard worksheet with precoded pigeonholes or boxes; professional transcribers would then fill in this sheet from the evidence provided by the entries. This latter course has several disadvantages in that it involves paying professional staff to write or type data, the possibility of transcription error is increased, and it takes longer. For these reasons I would suggest that the coding of existing data is more economical. At this point one shoud question whether professional (or trainee professional) work is required for coding. In our experience it is necessary, partly because the coding or the transference of entries to a worksheet is not an automatic or a clerical process, and partly because it may be necessary (and in the case of a catalog which has been built up over a period of time and which has a considerable history, it *will* be necessary) to amend the bibliographic information before its transference. For example, the BNB entries reflected for most of their history the 1908 cataloging rules and the converted entries were intended to form part of a data base which was based on the 1967 cataloging rules. It was necessary to decide for example, the relevancy of certain elements of personal headings, and the form of many corporate headings. These decisions cannot be made by clerical staff without a considerable amount of training. Experiments we did using nonprofessional staff for coding and amendment indicated that the amount of error created by lack of knowledge of cataloging rules was unacceptably high; these errors had to be corrected at the proofreading stage, which is even more expensive than the stage of the creation of the coded data. This amendment of bibliographic data is necessarily limited. The form of the bibliographic entry can be changed, but such matters as the different choice of main entry by different cataloging rules, and the changes in class numbers necessitated by different editions of (say) the Dewey Decimal Classification, cannot be done exactly without reference to the actual item itself. It is an axiom of catalog conversion that the only economical way to carry it out is by using the data already present in the system. Any attempt to recatalog or reclassify the actual items will prove to be ruinously expensive.

One other problem remains at this stage of dealing with any file—the identification of duplicate records. These duplicate records either refer to different copies of the same item or to items which are too similar to require separate description in the converted file. I will mention here two ways of identifying duplicates. Call numbers, if they have been consistently applied, can be matched and will provide a list of suspect duplicates should the same call number occur on more than one record. Another possible strategy is the matching of certain data fields. A coincidence of title proper and date will provide a list of almost certain duplicates. A more interesting and advanced technique is that which was developed by the Project LOC[7] which was devoted to records relating to early printed books. This is the technique known as "fingerprinting." The theory behind fingerprinting is that it provides a unique identifier for an issue of a book by taking certain arbitrary data. To take an imaginary example, these might be the first four letters of the title proper, the first three letters of the publisher's or printer's name, and the first and last three letters on page twelve. It has been demonstrated that such assemblages of data provide a very high degree of accuracy in identifying duplicates even in old established catalogs where items have been acquired over a great many years and where the descriptive cataloging practices have varied a great deal over those years. The fingerprint also provides a control "number" for the converted record. In the absence of an ISBN, LC card number, or other unique identifier, it is necessary to devise a control number system. The only criterion upon which this should be based is that the number should be compatible (of the same length and type) with the control numbers used for records in the established or projected current cataloging system.

The next stage in the conversion process is the transference of the coded information into a form which can be fed into the computer system. In the BNB system to which I have been referring, the coded and edited data were keyboarded by persons who not only were not professional librarians, but also were not familiar with the content and structure of bibliographic records. This was possible because the BNB entries were printed, and therefore the basic information on them was of a high level of legibility, and also because the simple coding and punctuation conventions meant that data could easily be transcribed by persons who were unfamiliar with library techniques. Once the records were keyboarded in the coded form they were very simply converted by program into MARC format. This process did, of course, not create complete MARC records, but did create records that were within the parameters of the holding MARC format standard (ISO 2709) and hence were compatible with existing UK MARC records.

In this system the keyboarders were linked via a minicomputer to a CRT display, on which they could see the records which they were keyboarding. There were very simple amendment and error-correction techniques. Among these were certain automatic validations; for example, certain fields and elements (e.g., the title and date) were mandatory, and if they were omitted a signal was shown on the screen. The program also supplied an automatic series of tags in the normal sequence. This reduced keyboarding costs by eliminating the necessity for keyboarding the tags, and also helped to eliminate error by presenting the keyboarder with the logical next element. Rejection of an element thus became a conscious decision (because an element of data was not present in a record) rather than an unconscious error. The costs of keyboarding data are analyzed by Duchesne,[8] and inflation and increased labor costs have combined to increase these figures considerably. A reasonable current estimate is sixty pence (approximately $1.20) for one thousand key depressions. Any strategy which reduces the number of key depressions is, therefore, of great value. To take a simple example, the decision to supply the period after the third digit of a DC number by program, rather than keyboarding it, saved more than one-half million key depressions in the BNB conversion. Other punctuation, subfield codes and capital letters can also be supplied by program.

The method used for keyboarding data in the BNB conversion is, of course, only one of a great many ways of submitting the data to the computer system. Other significant methods include the punching of paper tape or cards, the preparation of magnetic tape for batch processing, and the presentation of data directly on-line to the data base. Another method other than straightforward keyboarding which has been widely discussed is that of optical character recognition (OCR). This technique depends on the creation of cataloging data in a special typeface in which the letters, numbers, and symbols are sufficiently individualized to be read by electronic scanners and automatically transferred into machine-readable form. Investigations into using OCR techniques on conventional typefaces or conventional typewritten characters have been carried out. It seems unlikely, however, that in dealing with the catalogs which we have at present, with their inconsistencies in type and presentation, that such techniques can be used directly. In other words, there will have to be in the use of OCR techniques, for the foreseeable future at least, an intermediary stage of transference from conventional letterpress or typewritten characters into electronically readable characters. The most economical method for the future is likely to be the direct interaction, using on-line techniques, of the keyboarder and the data. That is, records will be immediately processed and presented on a CRT to the keyboarder or keyboarder/editor so that the information can be seen to be correct as it is added to the file. Recent advances in technology have made this method both feasible and likely to be economical.

The next stage in the conversion process is the checking of the data by the professional staff—the "proofreading" process. It is clear that to put the maximum amount of professional effort into the coding and editing of the data before keyboarding will save on proofreading and amendment costs. These costs are likely to be rather high for two reasons. One is that the proofreading activity—the activity of editing the file—is something that must be carried out by high-level professional staff. It cannot be reduced to a clerical routine, nor can it be entrusted to relatively inexperienced professional staff thus making per hour costs high in the proofreading stage. Additionally, the amendment process generally requires substantially more in computer resources per record than does the creation process. For these reasons correction of errors at the proofreading stage is expensive and time consuming. Another reason for the expense is that amendment costs are also high; either the file must be accessed to put in replacement fields (or in extreme instances, whole replacement records), or the records must be called on-line so that they can be checked and amended. Either way this is an expensive and time-consuming process. A certain amount of proofreading and amendment will be necessary even when the input is of high quality, because catalogs and bibliographic files are complex structures. They are more than a heap of glittering baubles—a mere assemblage of records; they are systems which allow the user to relate one record to another, to establish the identity or nonidentity of two or more items, and to survey groups of records sharing a common characteristic. Therefore, an overview is needed. We took the view in the BNB conversion process that a sample proofreading technique would be adequate. This was because we had devoted a considerable amount of time to high-quality editing at input, and also because the entries from which the input was taken were similar in quality and level of bibliographic standardization. Even with such a situation, however, we found that a sample proofreading of one item in every ten (paid for at a rate double that of the editing rate) did not result in satisfactory quality control. It seems to be necessary to maximize the quality of input and to supervise that quality very closely, but also to make sure that there are resources available to ensure a complete overview of the results of the conversion process.

One important general question which arises in considering the conversion of bibliographic files is that of training staff for editing, keyboarding, and proofreading. Undoubtedly the best staff for a conversion project consists of those who are familiar with the creation of records for a current automated system. The reason for this is fairly obvious—one should have an idea of the purposes of a task in order to carry it out effectively. It is a minimum requirement that the supervision of the project be carried out by a trained librarian who is familiar with the application of computers to bibliographic processing. It is a grave error to plan such a project or to carry it out without at least this measure of professional knowledge and control. If the other staff engaged in the conversion project are not familiar with library automation and cataloging techniques, then a training system must be evolved. This can be a costly activity but one which will recoup the costs of a good training scheme over and over again during the conversion. This type of activity (a major one-time project) is one where on-the-job training is not at all suitable, and may have dire results. Certain aspects of the BNB conversion have proved this. The training system should embrace not only the techniques to be used in this particular project, but also an overview of the bibliographic standards to be used (or which the project is trying to achieve), and the use to which the converted records are to be put. The editor must be aware of the consequences to the future cataloging system of a particular editing decision. It is particularly important because the editing process, although it can be rendered mechanical to a very great extent, will always have a residue of decisions to be made. To take one very simple example, there is the question of uniform titles (or filing titles) found in some entries. One simply cannot specify the addition of uniform titles or the transference (or nontransference) of uniform titles without understanding the context of the subsequent use of the records.

This has been a brief and necessarily sketchy overview of a large and complex subject, and I can hope to have done no more than to indicate the main strategies and the relative economic consequences to those wishing to convert their back files into machine-readable form.

I would like to thank my colleagues in the British Library, Andrew Phillips and Bruce Royan, for their help during the preparation of this paper.

REFERENCES

1. Jolliffe, John W. "Retrospective Conversion of the British Museum Library Catalogue: Techniques, Strategies and Costs." *In* Department of Education and Science. *The Scope for Automatic Data Processing in the British Library*. London, H.M.S.O., 1972. (Available from National Lending Library for Science and Technology, London, England.)

2. Plaister, Jean. "The LASER/MARC Project," *Catalogue & Index* 34:3-4, Summer/Autumn 1974.

3. RECON Working Task Force. *National Aspects of Creating and Using MARC/RECON Records*. John C. Rather and Henriette D. Avram, eds. Washington, D.C., Library of Congress, 1973.

4. RECON Pilot Project. *RECON Pilot Project: Final Report*. Henriette D. Avram, project dir. Washington, D.C., Library of Congress, 1972.

5. Duchesne, R.M. "Unit Costs of Conversion of Bibliographic Records to Machine-readable Form," Appendix 5 to Annex A: "Cost Estimates for Implementing Recommendations." *In* Department of Education and Science, *op. cit.*, pp. 145-58.

6. RECON Pilot Project, *op. cit.*

7. Jolliffe, John W., dir. *Computers & Early Books: Report of LOC Project*. London, Mansell, 1974.

8. Duchesne, *op. cit.*