Retrospective Conversion

Richard W. Boss

Editor's Note: *This is an edited transcript of a presentation made at the NCLA/RTSS Automation Symposium, April 9, 1981.*

Building the Data Base

Before you undertake a retrospective conversion project you should decide on the nature of the data base that is to be built. That is saying the obvious, yet the fact that many non-standard and incomplete data bases are being created in an era when most libraries are talking about moving toward on-line catalogs justifies its repetition. We tend to get too preoccupied with the obviously high price tag of hardware and software and overlook the basic fact that the data base, if properly built, will outlast several generations of hardware and software. So your data base will outlive the first computer you buy—and several others after that. After all, the 100 percent rag catalog cards have outlasted several generations of catalog cabinets in our libraries.

You should also think in terms of software being outlasted by the data base. Now software, in my opinion, should last considerably longer than hardware. The higher level programming languages being used in software packages, including most of those now being used in turnkey packages, are are more oriented to the programmer than to the hardware. The ease of writing new code or of revising existing code is emphasized, rather than efficient utilization of the machine. We can afford to have a little inefficiency in the use of the machine in order to have efficiency of people.

It is unfortunate that some libraries that went into automation very early—before 1974—built data bases with records averaging fewer than 120 characters, but many of them had little choice because computer storage was limited by the small capacity and high cost of the available technology. By 1975 libraries were creating data bases with an average record length of 300 characters. At that time, some of the libraries that had automated earlier upgraded their data bases to more complete bibliographic records. By 1978 the average length of a record had increased to 400 characters, and in a few cases libraries actually expanded their records once again in order to have a better data base. Recently, the trend has been to full-MARC records averaging 700 characters, and at least one library is expanding the records in its data base for the third time. Thank heaven that they haven't got any place to go beyond 700. Once you are at full-MARC you have reached the last step—at least we hope so.

Weeding

A number of people have commented today that one might reduce the magnitude of the effort by weeding before doing a retrospective conversion.

This is true, but you can get bogged down in a weeding program, and weeding is not an inexpensive proposition in itself. There are two ways in which you might proceed with the weeding. One is a quick cut approach in which you weed the obvious things, multiple copies, textbooks, etc. The other approach is a comprehensive weeding program in which the entire collection is reviewed. In my opinion, the last thing you want to get into is a laborious title by title examination of the collections by professional staff.

If you are going to undertake a weeding program, think in terms of the kind of weeding you can do that keeps the cost per title weeded low and the time commitment minimal. As a rule of thumb, plan about a twenty-five to thirty-cent investment per title. If you have a cost-effective enough retrospective conversion technique, in balance you are better off avoiding a major weeding.

Eight Major Options

I want to spend most of my time talking about various ways that you might proceed with a retrospective conversion. I will not advise a single approach for all of you, but will seek to give you an idea of the range of options available.

MARC Data Bases

I suppose there is hardly anyone who participates in a bibliographic utility or who has had a COM catalog supplied by a COM vendor who doesn't think of utilizing his/her own machine-readable data base. If you have a MARC data base already as a result of participation in a bibliographic utility, that is certainly the logical first element in creating the data base for an in-house or shared system for circulation and/or a patron access catalog. If you have an OCLC or other bibliographic utility archival tape, the cost of the reformatting into the appropriate operating format should not be more than one cent per record. A higher quotation should be carefully checked.

MARC is a standard format and virtually every vendor has software to accommodate it. The less your existing records conform to MARC, the more expensive it is going to be to convert the records into the operating format of the system you have selected. A reliable way to determine costs and other problems you may encounter is to take a random sampling from the total file and send it to a vendor for a price quotation.

If you are installing a turnkey system, one for which a single vendor supplies hardware, software, installation, training, ongoing maintenance of hardware and software, and software enhancement, have the vendor load your existing data base, if any, at his site before shipping the system, so that you will already have some of the bibliographic file in the system before you start building your files. The tape loading facilities available to the turnkey system vendor are much faster than those which can be cost justified for an in-house library system.

Keep in mind that what you are building is a bibliographic file. The contemporary approach is to have a file separate from the bibliographic file for information about the volumes and copies. This file may be called an item or copy file. It includes such things as call numbers, copy numbers, location codes, identification numbers, and other item specific information not appropriately part of a bibliographic record, that might be transferred from one file to another.

Previous Circulation and Acquisition Systems

Another source of bibliographic records may be any machine-readable records that were created for a previous circulation or acquisitions system, even for a batch system that used the old Hollerith IBM cards with only 80 punched columns. The idea is to try to match these brief records against a full-MARC data base. It is normally done by going to one of the COM vendors such as Autographics, Science Press, etc., and having them match tapes of your brief records against their data base(s). Usually the matching is done by comparing the records on several different points or characteristics.

The vendor tries to match on the LC card number or the ISBN/ISSN, but those are not always available. A second choice is to match on an author/title key by taking a certain combination of letters from the title and matching them against records in a data base. You may be faced with reams of printouts for editing because this approach does not always result in an exact match. The cost involved in sitting down and editing all that material can be very substantial. It is a good idea to undertake a small pilot project in which you experience firsthand what is involved in this process.

For titles for which a library has no machine-readable records, some libraries prefer to key partial records themselves and then have the vendor do the matching. The LCCN or ISBN/ISSN is used if available; otherwise, a brief author/title key is entered. The tape is then sent to the vendor for the matching. The library may not be able to do the keying less expensively than the vendor, but it can charge the cost to its regular salary budget, rather to a supplies or operating account. Whichever matching approach is used, be sure that the records the vendor is going to provide are MARC records, not MARC-like.

Keying from the Shelf List or the Books

Another approach is to key the records from the shelf list or from the books themselves. Most libraries that take this approach create only brief bibliographic records because keying is expensive. Most libraries which have taken this approach do not know their exact costs and claims run as low as fifty or seventy five cents per record for up to 450 characters. It is unlikely that anyone is, in fact, realizing so low a cost. By the time you calculate all of the hidden expenses such as the staff time spent on editing and the fringe benefits and overtime, the real cost is probably well over $1.00. A major commercial service bureau that does a great deal of keying for libraries charges $1.95 per thousand characters; that is approximately $1.40 for a full-MARC record.

Renting a Data Base

Another thing that you can do is temporarily load a data base. This involves the renting of a data base from a vendor such as Blackwell North America, which has over 3 million records in its data base, and loading it on the in-house system. The books or shelf list cards are then brought to the terminals, and matches are sought by LCCN or ISBN/ISSN; or by brief author/title key. The advantage of this approach over the vendor matching method is that editing can be done on-line rather than by reviewing printouts.

The great danger in this approach is that the in-house system, which was sized to support circulation and/or a patron access catalog for that particular library may not accommodate the temporary loading of a massive data base. This approach makes the most sense when the size of the computer system installed is well in excess of the immediate needs of the library because several other libraries are to be added later.

The nature of the data base used is again a very important consideration. Many of those available for lease are not MARC data bases and the majority of them reflect the vendor's history of working with public libraries. An academic library might, therefore, realize a lower hit rate than a public library. Random sampling of the data base to be used is extremely important.

MINIMARC and REMARC

Yet another approach is the use of MINIMARC. This system is a micro-computer-based stand-alone cataloging support system. The data base consists of the LC MARC data base on diskettes. The use of MINIMARC can be quite cost effective if yours is a library that has a collection that would be well represented in LC MARC tapes. We have seen libraries with hit rates against MINIMARC as low as 30 percent and as high as 95 percent. Again, you have to determine your hit rate in order to determine whether this approach is the right one for you. The libraries which have had the best experience with the data base are public libraries, although a large number of four-year and two-year colleges have also had extremely high hit rates. Productivity on the MINIMARC system is very high. We have seen anywhere from 48,000 to 90,000 plus retrospective conversions on a single system, and I am sure that, if you had the proper organization with multiple shifts, weekend shifts and the like, you could exceed 100,000. So, if you are paying $15,000 to have the system for a year and you can get 100,000 records converted, the approach might be quite attractive. Keep in mind that OCLC retrospective conversion would cost less than this only during the non-prime hours.

Now you might say that it is not going to do you much good because your library has a lot of older materials, you have a lot of foreign language materials, you are somewhat more research oriented, or you are somewhat special in your requirements. There are some other ways to use this approach. One of them is currently being tried in West Virginia, a combination of MINIMARC and REMARC. As you know, Carrollton Press is building the REMARC database by

keying the Library of Congress' non-MARC cataloging. They are working alphabetically by main entry and are offering the records for retrospective conversion use as they are completed. In West Virginia, the State Library uses a MINIMARC system to search for records. Failing to find a record in MINIMARC, the operator enters the LCCN, ISBN/ISSN, or author/title key on a blank diskette. When the diskette is full it is sent to Informatics, the vendors of MINIMARC, for reformatting onto tape and delivery to Carrollton. Carrollton matches the tape against the REMARC data base and extracts any hits.

OCLC

The most popular retrospective conversion approach is to use the OCLC system. The vast majority of your libraries are associated with SOLINET/OCLC. As you know, the present retrospective conversion price is sixty cents during prime-time hours and five cents during off-hours. The off-hours rate will go up to ten in July, 1981 and to fifteen cents on January 1, 1982. When OCLC first began to offer retrospective conversion, there was no charge. The reason why the service was free was that the retrospective conversion enriched the data base. Well over a million records were added to the data base in a very short period of time as the result of the no-charge policy. The State University System of Florida alone undertook nearly a million retrospective records conversions, of which over half were new to the OCLC data base. But the rate of enrichment of the OCLC data base began to drop off dramatically in late 1979. The amount of retrospective conversion being done had also begun to have a significant impact on computer resources. The dual pricing structure now in effect was therefore developed. The idea was to establish the concept of charging for retrospective conversion and to discourage libraries from doing the work during the hours that the computer system had its greatest load. Nevertheless, the rate of retrospective conversion has continued to increase. It has grown to such a point that it is going to be necessary to increase the capabilities of the system to support retrospective conversion. The money to do that has to come from somewhere. It does not take an exceptional crystal ball to guess that the price for retrospective conversion is going to rise until it pays its share of the operating costs.

I have no way of knowing how fast the rate will go up or to what level, but I suspect that the cost to OCLC of supporting a retrospective conversion is a lot closer to sixty cents than it is to the five cents figure. So it would be wise to launch your retrospective conversion program now, ideally with a written agreement fixing the price. There is no assurance that you will get it. You can certainly try and, depending on the length of the retrospective conversion program, it may be possible. OCLC has had some protected price agreements in the past, but they have been rare. I am not aware of any that have been made recently except where OCLC is doing the actual work. That is, OCLC has bid some projects such as for the Philadelphia Free Library, where it has said that it would actually provide the labor for the retrospective conversion work for a

fixed quoted price. Most of the agreements of this type provide for OCLC to use a copy of the shelf list to find matching records in the data base, add local information, and create a new record for a price of seventy-five to ninety cents per record.

This price compares quite favorably with the in-house cost of doing a retrospective conversion on OCLC. The State University System of Florida estimated that its cost for converting nearly a million records was sixty-seven cents each at a time when OCLC was not charging for the service. At today's rates that would be seventy-two cents if undertaken during off-hours.

Now keep in mind that you may have to get additional terminals in order to undertake a significant amount of retrospective conversion, and the waiting time for the additional terminals can be significant. You may be eating into those remaining days of OCLC's nickel rate while waiting for the additional terminals to arrive, so it is not at all a sure thing. I would encourage you whenever you think of adopting a retrospective conversion technique to have a backup approach in mind should your first choice cease to be cost effective.

Optical Character Recognition

Another retrospective conversion approach that I was asked about here in the hall today is that of optical character recognition. Why can't we scan catalog cards and translate the images into machine-readable form? We have heard about the Kurzweil machines that scan printed books and synthesize speech for the blind. If it is possible to convert printed information into machine-readable form and synthesize speech from it, why can't you scan catalog cards? We have been to Kurzweil in Cambridge three times with stacks of catalog cards in the hope that as they continue refining that system, it will be possible to do just that. To date, the results have been absolutely miserable. Unfortunately, the scanner is geared to deal with a full size page and when you put 3x5 cards on it the machine will not register properly. The variety of type fonts and quality of catalog cards pose additional problems. It may be five or more years before this attractive new technology will be practical for libraries.

Evaluating the Technology

How should one evaluate a retrospective conversion technology? Obviously, the first thing you are going to look at is cost. Tally up all of the costs. How much are you going to have to pay somebody outside the library? That is only part of it. How much are you going to have to spend in terms of the value of the time of the staff in your library to do the necessary editing and all of the other things associated with the retrospective conversion effort? If you do it all in-house, be sure to count in more than just the salaries. The cost of fringe benefits for staff and the like all are part of your institution's real cost, even if they do not show up in your budget. If you invest in special equipment that will be used solely for the retrospective conversion, include those costs. If your figure isn't getting close to $1.00 per record by the time you include all of these factors, do it

again and do it more carefully. I submit that anyone who tells you that their figure is fifty cents or less has discovered a wonderful new method or has failed to calculate all of the costs.

A vital factor in evaluating the cost of any approach is the hit rate. Obviously, two quite different prices quoted on a cost per record basis may, in fact, be comparable if the hit rates of the two approaches are different. To assess the hit rate, check a random sample of your titles, using the various approaches you are considering, in order to determine the relative percentages you will be able to convert with each method without doing original data entry. If you pay less per hit, but you get very few hits, it means you have a disproportionately larger number of things that you are going to have to convert with an alternative and presumably more costly technique. The way to do that is one of two things. Pull a random sample and use a MINIMARC or an OCLC terminal or send them to a vendor and get a match against the data base to get a fix on what percentage you will be able to convert without doing original data entry.

As important as the cost of a retrospective conversion is the quality of the records you get. Non-MARC records will cost you more in the long run because it will be more difficult to share or exchange data bases with other libraries, and you will pay reformatting charges every time a vendor has to work with your records. Brief records may save a little bit of computer storage cost now, but you will probably pay by having to expand the records at some time in the future when patron access catalogs become common.

Yet another factor is the length of time the retrospective conversion will take. If the lowest cost option will take several years and you need the data base within one year, it makes sense to examine other options.

Unfortunately, I can't tell you from this rostrum which of the retrospective conversion techniques you should use in your library. There is no single retrospective conversion technique that is right for every size and type of library under every circumstance. You have to do that careful analysis in order to determine which is right for you. As long as the OCLC five cent rate prevails, however, OCLC retrospective conversion proves to be the most attractive approach more frequently than any other.

One final thought: Don't enter any retrospective conversion program without a written agreement of some type that sets forth the rights and obligations of both parties. You have an obligation to the institution for which you are working to protect it in the future against all types of circumstances, even the possibility that you will become the president of OCLC and have to raise several million dollars to pay for an expanded computer system.

Richard Boss is president, Information Systems Consultants, Inc., Bethesda, MD.