PETER SIMMONS

Choosing Data Conversion Equipment

Since the automation of libraries requires files of bibliographic data in machine-readable form, librarians responsible for automation activities need to compare the equipment available for data conversion. Keypunch and typewriter keyboards must be considered, as well as devices which encode punched cards, paper tape, and magnetic tape, and on-line terminals. Once suitable machines have been identified, two other major criteria must be considered—price and reliability—though the latter can rarely be predicted accurately.

THE GOAL OF the Library of Congress' MARC Distribution Service is to encompass the entire spectrum of current LC cataloging. Though that goal may not be reached for many years, the promise of centrally captured and distributed bibliographic information in digital form for all current materials is bright. But since one of the academic library's unique assets is its ability to transcend the present —to control and make available materials without regard for their age—the creation of a machine-readable store of bibliographic information must of necessity accommodate retrospective as well as current information. Despite the likelihood that in the future retrospective cataloging information will also be centrally distributed,[1] many libraries that intend to take advantage of computer technology will become involved in converting their own bibliographic data. Since machine-readable files of bibliographic data are a prerequisite for automated library functions, it is difficult to imagine such a library that will not want to add local elements of information to centrally produced records and local records to centrally produced files.

Given the desire (or need), personnel, and finances to accomplish data conversion, the remaining requirement is for suitable equipment. Though few librarians are likely to become enraptured with the analysis of character sets, transmission rates, and parity checks, it is nevertheless important that librarians responsible for implementing automation realize that their work requires machines that can input a larger number and variety of characters than are used in most computer applications. Therefore they will need at least a passing acquaintance with the various devices that are currently available for the task of converting bibliographic information to digital, or machine-readable, form.

The great majority of conversion devices presently on the market combine some sort of keyboard with an encoding mechanism. This brief survey will first consider types of keyboards, then the various methods of encoding, and finally some of the criteria for choosing among the various types of machines.

The easiest way to classify the several kinds of keyboards that are used on encoding devices today is simply to count their keys, since their number will usually (though not always) provide an in-

[1] The Council on Library Resources has funded a study, being undertaken at the Library of Congress, to consider conversion of the retrospective records of the Library of Congress to digital form. The stated purpose is to provide data for the entire library community.

Mr. Simmons is Assistant Professor of Librarianship, University of British Columbia.

dication of the number of characters the keyboard can encode. Although smaller devices are in use, the smallest keyboard of interest to librarians is probably the keypunch, which generally has thirty-four keys and can record sixty-four unique codes. It will be recognized, of course, that this number is far too small to encode all of the characters required for bibliographic data. In fact, no fewer than 175 characters are required to represent adequately roman-alphabet languages and romanized forms of non-roman alphabets, even if most scientific and technical signs are spelled out (*e.g.*, "square root" rather than $\sqrt{}$) and certain liberties are taken with some diacritical marks.[2]

There are several possible ways of dealing with the severe size limitation of the keypunch keyboard. The first method is the most obvious: simply to limit the character set to the ten numbers, twenty-six letters, and twenty-eight symbols of punctuation that appear on the keyboard. The obvious inadequacy of this solution has not prevented its widespread use by librarians as a simple expedient, apparently in the belief that the sacrifice of typographical niceties in exchange for the efficiency of computer-printing is necessary. The production of printed cards, books, and lists using only uppercase letters is the direct result of this technique. Another escape from the 34-key limitation is multi-punching, which consists of pressing a combination of keys to produce a code that no single key can create. This raises the maximum number of characters to the limit of the medium on which the information is encoded. On punched cards, for example, this number is 2^{12}, or 4,096 unique codes, each of which can represent a character or diacritical mark. The major difficulty with multi-punching is that no useful printed record can be made of the punched character at the time of keying; for example, the characters 4 and Q might be punched together to represent an umlaut, but the characters printed by the conversion device would be merely an unreadable jumble indistinguishable from almost any other combination of characters. Furthermore, multi-punching is generally extremely slow and, because of the lack of visual verification, of unreliable accuracy.

These drawbacks apply similarly to a third possibility—the designation of certain keys as escape codes. In this case one character is interpreted as being a signal that the code which follows represents a character that does not appear on the keyboard. An obvious example would be to precede each capital letter with an asterisk to give the capability of encoding both upper and lower cases, which the standard keypunch keyboard cannot otherwise accommodate. Thus the sacrifice of a single character produces the ability to represent an additional sixty-three characters for a total of 126, and the use of two escape codes permits the encoding of 186 characters.

Although the use of escape codes is a legitimate and commonly used solution to the character set limitation, it can be used even more efficiently in conjunction with a typewriter keyboard. Since the standard typewriter keyboard includes forty-four character keys and two case shifts, it can encode eighty-eight unique characters before requiring an escape. Moreover, most conversion devices with typewriter keyboards also have at least one key that is normally used to create a code not associated with any character. Although these keys are designed to control some aspect of the encoding machine itself, there is usually no reason that they cannot be interpreted as escape codes by the programs that handle the data after it is input to the computer. Thus, with forty-

[2] For a thorough analysis of the characters required for romanized bibliographic data see Lucia J. Rather, "Special Characters and Diacritical Marks Used in Roman Alphabets," *Library Resources & Technical Services, XII* (Summer 1968), 285-95.

four character keys, two cases, and just one escape code, a typewriter keyboard can be used to code 176 unique characters. The additional major advantage of a typewriter keyboard derives from the ease and speed with which experienced typists can be trained to be keyboard operators. When the keying procedures are designed to imitate typing as closely as possible, the training period can be reduced to a few hours; in places where employee turnover is high or there is insufficient work for a full-time operator, this may represent an important consideration.

The two standard types of keyboards —the keypunch and the typewriter—account for the vast majority of keyboard-controlled conversion devices. Other kinds of keyboards, including the ten-key decimal keyboard that can handle only numbers, the twenty-four-key stenotype keyboard that can record over sixteen million unique codes through multipunching, and large keyboards devised for special applications, are certainly available. But most libraries will doubtless find that a standard keyboard, or some variant (such as a keypunch keyboard with upper and lower case capability) will prove suitable when used in combination with one of the several kinds of encoding mechanisms.

Since input devices are modular by nature—the keyboard and the encoding device controlled by the keyboard may be considered separately even though they are usually sold as a single piece of equipment—it is possible to combine any kind of keyboard with any kind of encoding device. Encoders fall into two categories: those which record codes in a medium suitable for temporary storage and subsequent input into a computer, and those which send coded impulses directly to a computer. It is assumed here that computer input is the primary reason for digital conversion and that storage of the encoded medium after input is both unnecessary and undesirable.

Of the various currently available keyboard-controlled conversion devices, the most familiar and still most popular is the punched card, which has been all but synonymous with mechanized information handling since Herman Hollerith patented his eighty-column card in 1889. The major strengths of the card punch result directly from its popularity; new and used machines are readily and inexpensively available, repair service is usually fast (and required infrequently), and card readers form an integral part of the vast majority of computer systems. For the librarian, card punching is usually the most convenient means of converting data, especially where it is possible to take advantage of existing equipment, personnel, and procedures. Even when this is not the case, supplementing the existing resources or locating a service bureau to punch cards at a flat rate is rarely complicated.

However, the limitations of the punched card for bibliographic information, though few, seriously challenge its impressive advantages. Not only are cards inconvenient to handle and store in large numbers, but their fixed length of eighty characters bears no relation to the variable length format of bibliographic data. The result is an awkward group or decklet of cards representing a single bibliographic record. Part of every card must be reserved for a control number indicating the relationship of that card to the other cards in the decklet, to permit sorting into correct sequence should a group of cards be dropped or otherwise scrambled. In order to insert the necessary control codes, the keyboard operator is required constantly to interrupt the flow of data, a process which slows down the keying and promotes inaccuracy. Moreover, even a single typographical error necessitates the correction of the entire eighty-column card on which it appears. Since the punched card is in such common use for a variety of kinds of information, it must be obvious that these inefficiencies are easily ignored or accommodated, but the fact remains that the fixed length format of the card is basically unsuited to the variable length nature of bibliographic information.

Because it permits the uninterrupted recording of long strings of information without the insertion of special sequence codes, paper tape is theoretically better suited but frequently less satisfactory. The dissatisfaction generally stems from the method for correcting typographical errors made and then discovered by the keyboard operator. Any paper tape typist who has spent an agonized hour (or more) handling yards of punched paper tape searching for a single elusive character, and has then attempted to correct it, has no doubt cast envious glances toward the card puncher, who needs only to find, remove, and repunch a single card. The fact that cards usually contain printed as well as punched characters, while paper tape as a rule contains no guide for human eyes of what the punched code represents, further complicates the handling of paper tape by people. But if the user of paper tape sometimes feels like some machine-age Laocoön, it is most likely because the potential of the computer for manipulating data has not been exploited. There is no reason why error correction cannot be accomplished by computer program after the information has been input rather than by the encoding device. When this is done, procedures can be written to ensure that paper tape need never be handled or searched manually, and from a human point of view the resulting system is generally far more efficient and effective than a punched card input system.[3] In such cases an entire day's keyboarding can be stored on a single roll of punched paper tape, a comparatively inexpensive and convenient form for temporarily storing data prior to computer input. Now that a variety of paper tape readers, for use both on-line and off, are available and widely used, the punching of paper tape is becoming more widespread among libraries involved in converting bibliographical data.

Yet even as the use of paper tape grows, its logical replacement follows hard on its heels. For almost every kind of device that records information on punched cards or paper tape there is a similar machine, often made by the same manufacturer, that encodes magnetically on recording tape. Some of these tapes can be placed directly onto a computer tape drive—others must first be read through a reader in the manner of punched cards and paper tape; but in any case, magnetic tape encoders possess certain characteristics that make them desirable data conversion devices. Since their only moving parts are the keys and the tape transport medium, they are quieter and more reliable than machines that must employ mechanical punches. Being able to record on magnetic tape, they can also erase and rerecord, thus permitting the operator to reverse the tape (commonly by use of the backspace key) to reach a typographical error and retype the erroneous section. This moves error correction from the computer, where errors are most conveniently deleted in paper tape input systems, to the conversion device, thereby simplifying computer programming and reducing computer processing time. Furthermore, since the magnetic tapes encoded by these machines can carry more information (anywhere from twenty to 800 characters per inch) then paper tape or punched cards (which hold ten and nine characters per inch respectively) and are reusable, they are ultimately more convenient and, in terms of raw materials used, less expensive than devices which punch holes in paper or cards. In several currently avail-

[3] To be fair it must be said that a computer program can similarly handle correction codes punched on cards, but this method of error correction is rarely, if ever, used.

able magnetic tape encoders, for example, the information found on over 30,000 printed catalog cards can be recorded on a standard 2,400-foot reel of computer tape, which is then ready for immediate computer processing without requiring the use of an auxiliary converter.

In order to find a more convenient medium than magnetic tape for recording information, one must look at what is, in any case, surely the most logical kind of input device—one which enables the data to be transmitted directly to a computer-controlled storage module: the on-line terminal. This machine is connected directly to a computer either by cable or common carrier (such as a commercial telephone line), enabling the information to travel from the keyboard to the computer without intermediate storage in any tangible medium. Many (though not all) of these devices permit an auxiliary display of the information at the time of input, usually by printing on a typewriter (which most on-line terminals resemble) or on a cathode-ray tube. With the recent introduction of cathode-ray tubes capable of displaying ninety-six different characters and the ability to edit the displayed information with an electronic "light pen," the use of cathode-ray tube keyboard terminals will no doubt increase in the future as the computer costs associated with on-line equipment drop.

A development which has contributed to the popularity of on-line input has been the marketing of small, special-purpose computers used exclusively for receiving data from terminals and for temporary storage. Such systems, which offer few of the manipulative or computational capabilities normally associated with digital computers, usually consist solely of keyboard terminals, magnetic disks, and a small control unit. As the disks are filled their information is transferred to a larger computer or to another medium more suitable for long-term storage. This independent system thus leaves the larger general-purpose computer free for computation and provides a device dedicated purely to the task of converting large amounts of information to digital form.

Another method of input that is likely to gain favor as the sophistication of the machines grows is known generally as OCR, for "optical character recognition." There are several kinds of machines that can optically recognize or "scan" characters, and at present their limitations are large. Until recently, character recognition was limited to mark sensing (as used on answer sheets for standard examinations), magnetic ink (seen most often on bank checks), and the optical recognition of the ten arabic numerals—hence the post office ZIP codes. The recognition of alphabetic characters has been and often still is limited to certain type fonts designed specifically for OCR machines, and typewriter manufacturers have been quick to offer one or more of these fonts on their equipment in the expectation that the use of machines to read typed documents would rapidly rise.

These expectations have not been in vain, and the growing use of OCR has resulted in the development of machines that can recognize a variety of typewriter fonts and even hand printing. Though the day when machines can read handwriting, even library handwriting, has not arrived, it is now possible to find equipment capable of automatically converting the information on a typewritten form to machine-readable codes. This permits a single system to be used both for the conversion of existing files of typewritten cards and for inputting current information that is first typed onto a form, then scanned and converted. And without a doubt before many more years pass the machine that can read printed LC cards will be developed. Meanwhile, OCR systems continue to find new users as they develop flexibility and sophistication. Although they are still extremely expensive to purchase, in many areas they are available for lease at hourly rates that make them competitive with other means of input.

Once a number of suitable conversion devices have been identified, consideration must be given to costs, which have not been listed in this survey both because prices vary among manufacturers and among users, and because price, though important, should never become the prime criterion. The general rule, as one might suspect, is that the more recent and sophisticated devices cost more than the older, simpler ones. But there are significant exceptions. Certain magnetic encoders are less expensive than comparable paper tape punches, and when actual production is measured and analyzed, the cheapest machines sometimes prove unexpectedly expensive. On the other hand, a recent report[4] finds that on-line costs run almost twice those of paper tape, suggesting that while online terminals have their uses, input is one of them only for those who can afford the added expense. Computer costs are constantly falling, however, and we soon may be approaching the day when on-line input to a computer will prove as economically advantageous as it is desirable.

The initial cost (that is, the purchase or rental price) of a keyboard device is only part of the total expense. With all methods of conversion other than encoding on computer-compatible magnetic tape and on-line transmission, a device is needed to read information from the encoded medium (cards, tapes, etc.) into the computer. That these machines impose their own technical and financial problems hardly requires stating, though equipment salesmen have been known to neglect to mention the subject. Furthermore, additional costs are often imposed by the need to keep one or more machines available as backup, so that production can continue when one machine requires service.

A third criterion for the evaluation of a conversion device, beyond suitability to the task and cost, is reliability. This aspect is the most difficult to judge accurately, for there are no indicators of reliability other than the manufacturer's claims, the salesmen's opinion (these two may not coincide), and the experience of others. While the last of these may appear the most trustworthy, one needs to be sure that the information comes from the best-informed source (a keyboard operator may point out problems her supervisor is unaware of) and that the past experience bears some relation to the proposed use. Many punching devices that are satisfactory when used intermittently several times a day fail completely when subjected to forty hours of use every week. Moreover, a given model of machine will often differ as much from other models made by the same company as it will from the products of other companies.

But these warnings of pitfalls and potential hazards should not be a cause of pessimism—only of cautious skepticism. Librarians have never before had such a variety of suitable equipment from which to choose, and the choice grows every year. Cheaper, more sophisticated, and more reliable equipment is constantly being developed. Entirely new techniques of input—as different from today's methods as optical character recognition is from punching holes in cardboard—are without doubt being designed and developed as this is written. But even when input systems that can accept information as handwriting and as spoken words become as widespread as punched cards are today, librarians who cautiously investigate the various machines available before choosing one will find the expenditure of time amply rewarded. ■■

[4] Alan R. Benenfeld, *Generation and Encoding of the Project Intrex Augmented Catalog Data Base* (Cambridge, Mass.: Electronic Systems Laboratory, Department of Electrical Engineering, Massachusetts Institute of Technology, 1968).