CREATION OF COMPUTER INPUT IN AN EXPANDED CHARACTER SET

Donald V. BLACK: System Development Corporation, Santa Monica, California (Formerly, University of California, Santa Cruz, Calif.)

Keypunching of an expanded character set for library catalog data is described. The set included 101 different characters. Source documents were shelf list cards, the master record at the University of California Library, Santa Cruz. At the end of February, 1967, some 50 million characters, representing more than 110,000 separate titles, had been punched. Some of the considerations leading to the adoption of this method for the creation of machine readable input are given, and details on costs and production rates.

For manipulation by a computer, data must be converted to machine readable form. There are still only a few reasonably flexible means of creating machine readable records, especially if the data include an expanded character set. Five possible methods utilize one of the following: standard keypunch, paper tape-producing typewriter, optical character reader, keyboard device that encodes directly onto magnetic tape, or a keyboard terminal that inputs directly into a computer. Descriptions of some of these methods are available in the literature. The Johns Hopkins University (1) used optical character recognition which can handle a full alphanumeric representation, whereas Southern Illinois (2) used mark-sense scanning to convert only a limited amount of information. Cartwright (3) and IBM (4) discuss direct computer input from a keyboard terminal. Buckland (5) discusses the use of the paper tape-producing typewriter. Hammer (6) and Kilgour (7) discuss keypunching. Patrick (8) discusses several methods of conversion, but only in the abstract.

Chapin (9) presents the results of a comparison test of the first three methods above.

This paper does not discuss the relative merits of these methods, but rather presents the details of a system that has converted approximately 500 million characters of library catalog data in more than 20 languages, with a set of 101 characters.

The University of California at Santa Cruz is one of three university campuses recently established by the State. It opened for business in the fall of 1965 with a core collection of some 55,000 titles in approximately 80,000 volumes. Early in the operation of the Library, it was decided to use machine methods as much as possible; therefore the existing catalog records had to be converted if the original collection were to be a part of the future machine system. The creation of the core collection for the three new campuses of the University of California has been described in the literature (10).

METHODS

Bids were sought to convert the catalog records during the summer of 1965. The shelf list record produced by the new campuses' project was the master record and was to be the source for conversion. Unfortunately, the shelf list consisted of both printed Library of Congress cards and cards produced at the new campuses' project from typewritten multilith masters. No editing was to be done on the shelf list cards. The only addition was the stamping of an arbitrary number using a five-digit automatic numbering machine, the purpose of the number being to keep individual punch cards together for each entry.

Weighing the responses to the request for bids was a disheartening experience. Only four responses were received from a total of 15 requests sent out. The bid request did not specify the method to be used to convert to machine readable form, but only the resulting machine readable record. Since the specifications had used punch cards as an example, perhaps this limited the thinking of some of the organizations involved, with the result that they did not choose to bid.

Three bids were based on keypunching. One was from Florida and the complexities of the task made the choice of such a distant company impossible. If problems had arisen during the course of the conversion, travel costs would have been excessive.

Another response estimated the cost to be about $1.50 per record. Clearly, this was too costly, and since bids of this nature are apt to be conservative in the matter of ultimate total costs, we felt the choice of such an organization to do the job would, indeed, result in a target figure that would be too high.

Only one bid used optical scanning as the method of conversion. Unfortunately, the bid was for the scanning only, and Library staff members would have had to retype the records for the scanner. Since the cost

SPECIAL CHARACTERS NOT PRECEDED BY A WORD SEPARATOR

Character	Standard Keypunch Keyboard	Card Code	Name
$	#	3 - 8	Dollar
@	@	4 - 8	At
>		5 - 8	Greater than
<		6 - 8	Less than
+	&	12	Plus
/	/	0 - 1	Slash
#		0-2-8	Record mark
,	,	0-3-8	Comma
=	%	0-4-8	Equals
-	-	11	Minus (below center)
#		0-6-8	Number
%		0-7-8	Percent
-		11-0	Minus (on center)
_	$	11-3-8	Underscore
*	*	11-4-8	Asterisk
:		11-5-8	Colon
;		11-6-8	Semicolon
'		11-7-8	Apostrophe
?		12-0	Question mark
.	.	12-3-8	Period
(	¤	12-4-8	Open parenthesis
)		12-5-8	Closed parenthesis
[		12-6-8	Open bracket
]		12-7-8	Closed bracket
a - z, 0 - 9		usual punch	Small letters and digits

SPECIAL CHARACTERS PRECEDED BY A WORD SEPARATOR (0-5-8)

Character	(0-5-8) plus Card Code	Name
´	1	Acute accent
`	2	Grave accent
^	3	Circumflex
:	4	Umlaut (diaeresis)
°	5	Bolle o
Ñ	6	Tilde over capital N
♭	7	Flat
♯	8	Script L
~	5-8	Tilde
♭	6-8	Cedilla
ˇ	12	Haček
/	0-1	Slash (Overprint)
ξ	11-0	Ampersand (Greek letter Xi)
"	11-7-8	Double quotation mark
A - Z	Usual Punch	Capital letters
		A black square is printed by any of the following when preceded by the word separator:
		9, 0, 3-8, 4-8, 0-2-8, 0-3-8, 0-4-8, 11,
		0-6-8, 0-7-8, 11-3-8, 11-4-8, 11-5-8,
		11-6-8, 12-0, 12-3-8, 12-4-8, 12-5-8,
		12-6-8, 12-7-8

Fig. 1. Card Codes.

based on the scanning alone was close to 30¢ a title, that bid was also rejected as being ultimately more costly.

The final choice of a keypunching service in San Francisco was made on the basis of its proximity to Santa Cruz, on the enthusiasm of the bidder for the task to be undertaken, and on a reasonable cost estimate. The service bureau completed the task in slightly more than three months. Key verification was done on an IBM/056 Verifier during this mass conversion.

The expanded character set employed is shown in Figure 1. It was chosen because it was available on an IBM/1401 computer at the Los Angeles campus of the University (UCLA). At that time, it was the only 1401 with such a printer on the West Coast. The character set had been selected by librarians at UCLA from characters offered by IBM in the summer of 1964 for the 1403 printer.

As Figure 1 shows, digits and lower-case letters are punched as usual; upper-case letters are preceded by the word-separator character (0-5-8). Punching for the special characters is described in the tables. There are two minus signs: the one printed from an 11-punch is below the center of the character; obtaining a centered minus requires a multiple punch (11-0). The underscore prints in a space by itself, just as do other characters. It requires special programming to overprint this character by suppressing paper spacing. The virgule overprint requires two columns to punch. Sharp-eyed readers will notice that the virgule appears twice in Figure 1, and it has been counted twice for the total of 101 characters. The blank has also been counted as a character, but the black square, which was not used at Santa Cruz, was not counted.

All data elements were encoded in fixed card fields; that is, the field for each type of information had a fixed length, generally 300 characters. It was not necessary, however, to use the entire field or to fill it with zeros or other codes. No terminating characters were used to separate the fields. Each type of information was included on one or more cards bearing a code which would tell the computer precisely what type of information the card(s) contained. All of this is illustrated in Table 1.

There are basically two ways that information can be encoded into cards. This is discussed in references (3) and (6) especially. To use a completely variable format it is necessary to have field delimiting codes. If a fixed sequence of data elements is established (e.g., author, title, publisher, etc.), one code will suffice to separate each field. Fixed sequence necessitates, however, making provision for establishing sequence for every possible data element that can occur in a catalog entry; and if one or more elements do not occur for some specific entry, then the keypunch operator must remember to add the required field-delimiting code for every data element not present, and to add the code in the proper sequence. If a number of individual codes are to be used to delimit fields, it then becomes difficult to find codes that are not otherwise used for

Table 1. Punch Card Input Format

Field ID	*Cols.*	*Comments*
Shelf Key Card, No. 000		
Call No.	1-20	As desired
Year	21-24	Year of Publication
Copy	25-26	(Blank or 01-99)
Series	27	(Either alphabetic or numeric O.K.)
Volume No.	28-30	(Blank or 001-999)
Part No.	31-32	(Blank or 01-99)
Donor No.	33-36	No. of donor on gift list (may be blank)
Date Rec'd	37-40	Month, year received at Library
Location	41-42	Alpha code designating location on campus
Type	43	0=book, 1=serial, 2=reference, 3=gov't pub., 4=see auth., 5=see subj., 6=see also subj.
Language	44-47	From 1 to 4 one-letter codes indicating lang.
Suppress	48	If "S", entry appears on shelf list only
	49-69	Unused
Card No.	70-72	Must be '000'
Accession No.	73-80	8-digit No. which sequences a batch of accessions in Call No. sequence (generally only final five digits are used)
Personal Author, No. 100-104		(Limit: 0-5 cards)*
Author	1-60	Name of author, left justified
	61-68	Unused
Special Code	69	See Table 2
Card No.	70-72	100 through 104
Accession No.	73-80	Same as Shelf Key Card
Corporate Author, No. 110-119		(Limit: 1 or 2 cards per author; 0-5 authors)*
Corporate Author	1-60	May be continued on second card
	61-67	Unused
Cont. Indicator	68	Is '-' if author is cont. on second card
Special Code	69	See Table 2
Card No.	70-72	110-119
Accession No.	73-80	Same as Shelf Key Card

* The first author to be processed by the computer is considered the main author. The main author appears on all catalogs and is represented in the title by "***".

Table 1. (Cont.)

Field ID	*Cols.*	*Comments*
Title Card, No. 200-224		(Limits: 1-5 cards per title; 1-5 titles)
Title	1-60	May be cont. on up to 4 additional cards
	61-67	Unused
Cont. Indicator	68	Is '-' if continued on next card
Special Code	69	See Table 3
Card No.	70-72	200-224
Accession No.	73-80	Same as Shelf Key Card
Publisher/Source Card, No. 300-305		(Limit: 1-2 cards per publ./source, 3 publ./sources total)
Publ./Source	1-60	May be continued on a second card
	61-67	Unused
Cont. Indicator	68	Is '-' if publ./source is cont. on next card
Special Code	69	P=publisher, S=source
Card No.	70-72	300-305
Accession No.	73-80	Same as Shelf Key Card
Collation Card, No. 400		
Collation	1-40	As desired
	41-69	Unused
Card No.	70-72	400 (1 card only)
Accession No.	73-80	Same as Shelf Key Card
Commentary Card, No. 500-509		(Limit: 1-5 cards per comment; 2 commentaries, (1 of each type)*
Commentary	1-60	May be cont. on up to 4 more cards
	61-67	Unused
Cont. Indicator	68	Is '-' if commentary cont. on next card
Special Code	69	'S' for commentary to appear on shelf list only
Card No.	70-72	500-509
Accession No.	73-80	Same as Shelf Key Card
Subject Card, No. 600-604		(Limit: 1 card per subject; 5 subjects)
Subject	1-60	As desired
	61-68	Unused
Special Code	69	May be used to indicate level of subject**
Card No.	70-72	600-604
Accession No.	73-80	Same as Shelf Key Card

* If 2 commentary entries are used, at least 1 must have an 'S' code.
** The Special Code is ignored by the system at the present time.

legitimate data. It seemed easier to input with fixed fields and simply waste a few card columns. That is, if a particular field ends anywhere in the body of the card before the final column (for example, 60), the operator simply stops and feeds in a new card. It is possible that a card may have only one character of data on it, in addition to a card-sequence number and item-type number. In practice it would seem that the loss in time in card feeding is not significant, and blank Hollerith cards are very cheap indeed.

Training for the mass conversion effort at the keypunch service in San Francisco proved relatively easy. An operator's guide was produced showing the codes and conventions for each data element found on the typical catalog card. Only the shelf list card was used for the conversion. Tables 2 and 3 show the various elements that were coded. There were two forms of shelf list cards, as mentioned above: Library of Congress cards and cards produced by the new campuses' program at San Diego. On Library of Congress cards everything was encoded except Roman numeral pagination and size information, and the information at the bottom of the card: the call number used in the Library of Congress itself, the Dewey number, the LC card number, and the name of the originating library if any. On the home-made cards, Roman numeral pagination and size were not encoded. Everything in Roman characters was punched. Cards with only a small amount of information in Roman characters had a legend punched, "for complete entry see shelf list." It is simply not possible in a short article to give all the fine points of conversion. Rules for all contingencies were devised and most proved easy to follow. Twenty operators, working in two shifts of ten each, converted the 55,000 titles that existed in June, 1965, in about three months' working time. All data elements to be used later for sorting purposes on the computer were key verified, but for the first month of the conversion the entire record for each title was verified.

Beginning in December, 1965, the Library at Santa Cruz began keypunching operations. After a training period of a week and operational experience of four months, the local operators achieved a rate of 7,000 to 8,000 keystrokes per hour, with a net error rate of only 12 errors in approximately 24,000 keystrokes; That is, the operators recognized a number of errors and corrected them at the time of initial punching. The 12 remaining errors should be caught during proofreading, which we substituted for key verification in the ongoing production system. It was felt desirable to combine proofreading for transcription accuracy with the typical library practice known as "revision," which implies that the catalog copy be reviewed for content as well as accuracy. This is true even for text taken from Library of Congress catalog copy. Elements such as the form of entry, the form of series note if any, number of subject headings and form, etc., are all reviewed by a cataloger other than the one who initially prepared the copy. Proofreading and revision was done from a

Table 2. Special Codes for Authors

Code	*Meaning*	*Notation*
Type 1:	To create added notation on author catalog:	
J	Joint author	Joint Auth.
C	Compiler	Comp.
E	Editor	Ed.
G	Joint editor	Joint Ed.
I	Illustrator	Illus.
P	Publisher	Publ.
T	Translator	Trans.
Type 2:	To specify a substitute sort key:	
X	Use this author as a substitute sort key for previous author. Previous author will appear on appropriate catalog but this author will not.	

Table 3. Special Codes for Titles

Code	*Meaning*
X	Suppress listing this title in title catalog
T	Title is a transliterated title
S	Title is a series title
P	Partial title
D	Standard Title or conventional title

In all cases, the first title encountered when processing a given entry will be the only title which appears in the author and subject catalogs.

printout on a line printer having only 64 characters available in the character set. This number of characters suffices, however, since there are only 64 usable card codes, i.e., the pattern of holes in each column of a Hollerith card. There are only 64 valid combinations which can be read by the computer equipment. As illustrated in Figure 1, some characters with diacriticals require three punched columns to produce one character in the ultimate printout.

RESULTS

For the mass conversion which took approximately three months during the summer of 1965, the total cost was slightly less than $34,000, or approximately 60¢ per title. In a discussion of the project, after its conclusion, with the two supervisors of the service bureau in San Francisco, it was agreed that in all likelihood the service bureau operators had just

reached peak efficiency about the time the project terminated. That is, had the project continued, the cost per title would have decreased. From work records maintained by the service bureau, it was apparent that the first two months of the project was a learning period, as the output of the operators rose continually during that period of time. During the last month the productivity curve leveled off considerably.

Table 4 shows the cost of the production operation established in Santa Cruz. The costs used are somewhat arbitrary. For example, the "keypunch operator" classification at the University had six steps. In producing an average cost should the actual rate being earned by the keypunch operators be used, or the beginning rate, or some other? The amount used in Table 4 represents an average of the pay being received by the 2.3 full-time equivalent operators, rounded upward to an even amount. Hollerith card costs can also vary slightly. Table 4 uses $1.00 per thousand as a reasonable price and one which could probably be obtained anywhere in the country. Costs are based on rates obtaining in February, 1967.

Table 4. Cost Per Title to Produce Machine Readable Catalog Data

Keypunch rental, $65.00/mo. (one-shift operation)	$ .026
Keypunch operator, $2.10/hr.+20% overhead	.168
Blank Hollerith cards, $1.00/thousand	.009
Machine listing of cards for proofreading (printing at 390 cards per minute)	.002
Proofreading, $5.00/hr., 120 titles/hr.	.042
Correction of errors	.020
Total	$ .267

DISCUSSION

There are, of course, hidden costs in the ongoing production at Santa Cruz that are difficult to fix because the University does not charge for them. For example, there is the cost of space occupied by keypunch operators and by equipment, the cost of air conditioning and electrical supply, the cost of adding internal partitions, doors, etc. Spread over a yearly total of some 30,000 new titles, these unknown costs and the additional costs of supervision could not be very great per title, assuming that the rate of keypunching production remains relatively constant. However, labor costs may prove to be a key factor in some geographic areas. In Santa Cruz good keypunch operators were available at a reasonable cost, but in large metropolitan areas this may not be true. Since the operator cost is over 60 percent of the total per title, it obviously can be a critical factor.

What happens after catalog copy is converted to machine readable form by punch cards or any other method, depends on computer equipment available and the programs written to process the catalog card data. This is an easy statement to make, yet the road to the production of either catalog cards or book-form catalogs is not easy. Even the data themselves can cause problems. For example the reader will note in Figure 1 that the character used to indicate an up-shift for capital letters has a card code of 0-5-8 fixed by the manufacturer, IBM, and necessary to print with the expanded character set chain. In retrospect it would have been better to convert the final code configurations as part of a computer processing step than to punch them from the beginning. The 0-5-8 shift code is a special code used within the 1400 series computers, and is known as a *word mark* or *word separator*. In normal operation of the 1401 computer these marks are used to delimit fields within the memory of the machine. Certain program commands use these marks to detect when the beginning or the end of a field has been reached. Use of such a code in the data can raise havoc with a program unless the programmer is constantly alert to the problem and takes great pains to circumvent it. Some other code such as the $ sign might have been used and then converted, prior to the final run on the 1401, as part of the computer processing to the code needed for printing purposes.

While the number of articles on catalog conversion has not yet been overwhelming, it is apparent that there is a great deal of interest in the field. One might ask: "Should every library proceed to convert its own catalog?" DeGennaro has addressed this problem (11) and the reader is referred to that discussion. Perhaps the question has no ideal answer. It does seem unfortunate, however, that the pre-1955 National Union Catalog is not to be published from a record that is machine readable.

It would seem possible, however, that in the future methods could be devised to use machine readable records produced by the larger libraries and some procedure whereby the smaller libraries could check their holdings against those of the larger libraries. By some fairly simple method, a subset of the master machine records could be selected for use in the catalogs of smaller libraries.

The Santa Cruz project began before the Library of Congress had announced results of preliminary plans for the MARC project (12). To a certain extent the catalog record at Santa Cruz could be converted into the MARC format, although MARC goes far deeper in coding discrete elements of data within the catalog record than does Santa Cruz. However, to the extent that discrete data elements are encoded and identified properly in the machine record, any catalog format can be transformed into any other catalog format by the computer. The key is the proper identification of each data element.

ACKNOWLEDGMENT

The author wishes to thank his colleagues at System Development Corporation, in particular Mrs. Ann Luke, for helpful comments on this paper.

REFERENCES

1. The Johns Hopkins University. The Milton S. Eisenhower Library: *Progress Report on an Operations Research and Systems Engineering Study of a University Library* (Baltimore: Johns Hopkins, 1965).
2. Southern Illinois University. Office of Systems and Procedures: *An Automated Circulation Control System for the Delyte W. Morris Library; the System and Its Progress in Brief* (Carbondale, Ill.: Southern Illinois University, 1963).
3. Cartwright, Kelley L.; Shoffner, Ralph M.: *Catalogs in Book Form ...* (Berkeley: Institute of Library Research, 1967).
4. International Business Machines. Federal Systems Division: *Report on a Pilot Project for Converting the Pre-1952 National Union Catalog to a Machine Readable Record* (Rockville, Maryland: IBM, 1965).
5. Buckland, L. F.: *Recording of Library of Congress Bibliographical Data in Machine Form* rev. ed. (Washington, D. C.: Council on Library Resources, 1965).
6. Hammer, Donald P.: "Problems in the Conversion of Bibliographical Data—A Keypunching Experiment," *American Documentation*, 19 (January 1968), 12-17.
7. Kilgour, Frederick G.: "Development of Computerization of Catalogs in Medical and Scientific Libraries," *Clinic on Library Applications of Data Processing, University of Illinois, 2nd, 1964, Proceedings* (Champaign: Illini Union Bookstore, 1965), p. 25-35.
8. Patrick, Robert L.; Black, Donald V.: "Index Files; Their Loading and Organization for Use," *Libraries and Automation; Proceedings of the Conference on Libraries and Automation, Airlie Foundation, Warrenton, Virginia, May 26-30,1963* (Washington, D. C.: Library of Congress, 1964), p. 29-53.
9. Chapin, Richard E.; Pretzer, Dale H.: "Comparative Costs of Converting Shelf List Records to Machine-Readable Form," *Journal of Library Automation*, 1 (March 1968), 66-74.
10. Voigt, Melvin J.; Treyz, Joseph H.: "New Campuses Program (UCSD, UCI, and UCSC)," *Library Journal*, 90 (May 15, 1965), p. 2204-08.
11. DeGennaro, R. A.: "A Strategy for the Conversion of Research Library Catalogs to Machine Readable Form," *College & Research Libraries*, 28 (July 1967), p. 253-257.
12. U. S. Library of Congress, Information Systems Office: *A Preliminary Report on the MARC (Machine-Readable Catalog) Pilot Project* (Washington, D. C.; Library of Congress, 1966).

COSTS OF LIBRARY CATALOG CARDS PRODUCED BY COMPUTER

Frederick G. KILGOUR: Ohio College Library Center, Columbus, Ohio

Production costs of 79,831 cards are analyzed. Cards were produced by four variants of the Columbia-Harvard-Yale procedure employing an IBM 870 Document Writer and an IBM 1401 computer. Costs per card ranged from 8.8 to 9.8 cents for completed cards.

Early in September, 1964, the Yale Medical Library put into routine operation the Columbia-Harvard-Yale computerized technique for catalog card manufacture (1), and during the following three years Yale produced over 87,000 cards. The principal objective of the CHY project was an on-line, computerized, bibliographic information retrieval system. However, the route selected for attaining the objective included manufacture of cards from machine readable data to keep up the manual catalog while machine readable records were being inexpensively accumulated for computerized subject retrieval. Catalog cards were only one product of the system, but their production was designed to be as efficient as possible within constraints of the system. Nevertheless, this paper will examine CHY card production costs as though this segment of the system were an isolated procedure, yielding but one product, as is the case in classical library procedures. Costing will disregard other benefits, such as accession lists and machine readable data produced for little, or no, additional expense.

The Columbia Medical Library and Harvard Medical Library also installed IBM 870 Document Writers and tested the programs for card production, but neither library routinely produced cards. However, Co-