CONVERSION OF BIBLIOGRAPHIC INFORMATION TO MACHINE READABLE FORM USING ON-LINE COMPUTER TERMINALS

Frederick M. BALFOUR: Information Systems Engineer, Technical Information Dissemination Bureau, State University of New York, Buffalo, New York

A description of the first six months of a project to convert to machine readable form the entire shelf list of the Libraries of the State University of New York at Buffalo. IBM DATATEXT, the on-line computer service which was used for the conversion, provided an upper- and lowercase typewriter which transmitted data to disk storage of a digital computer. Output was a magnetic tape containing bibliographic information tagged in a modified MARC I format. Typists performed all tagging at the console. All information except diacriticals and non-Roman alphabets was converted. Direct costs for the first six months were $.55 per title.

Several recent articles have reported on methods and related costs to convert library bibliographic information to machine readable form. Chapin (1) compared keypunching, paper tape, and optical character recognition. Keypunching was also described by Hammer (2), and Black (3). Buckland (4) described paper tape conversion, and Johns Hopkins University (5) reported on optical character recognition. On-line computer terminals have been proposed (6), but have hitherto not been tried in a large library.

Without attempting to discuss the various techniques, this paper presents a detailed report of converting with on-line computer terminals. It is hoped that the experiences reported here and in the cited articles will provide suitable information to a library administration considering large-scale conversion.

BACKGROUND

In 1965 a systematic program of automation was begun in the Libraries of the State University of New York at Buffalo. The general goals of the program were to improve services to patrons and streamline internal operations.

There are three general areas usually considered for automation in a library: acquisitions and accounting, the card catalog, and circulation control. An analysis of the system indicated that conversion of the card catalog to machine readable form would provide the greatest improvement in library services and operations.

The reasons for the decision were as follows. First, the University Libraries are growing rapidly; in one year the shelf list will increase by 60,000 to 100,000 titles, or about 15 to 25 per cent. Second, SUNY Buffalo is currently planning a new campus which will be completed in five to ten years. In the interim, the University will be spread over three major campus locations, with many smaller offices and departments located throughout the city, and the Libraries must provide some form of bibliographic index for each location. The conversion of the shelf list to machine readable form will allow this distribution of the bibliographic information at a very low cost per title. Finally, the project will provide experience in using magnetic tape for the handling of bibliographic information, so that when the Library of Congress' MARC Project begins to produce magnetic tapes, SUNY Buffalo will be able to utilize them immediately.

SELECTING THE CONVERSION HARDWARE

In 1966, a proposal for converting the shelf list to machine readable form (7) was presented to the Library administration. It pointed out the many improvements in patron services, the advantages to the Library staff, both professional and clerical, and the monetary savings to be realized by such a conversion. It discussed the four methods of file conversion then feasible: punched cards, optical scanners, punched paper tape, and magnetic tape-keyed data converters (as exemplified by the Mohawk Data Sciences equipment) (8). The proposal recommended using the magnetic tape-keyed data converters because of their input speed, ease of entry, and elimination of handling cards or paper tape.

During the first quarter of 1967, a fifth method of conversion was considered, an IBM product called DATATEXT (9). It required the rental of an IBM 2741 communications terminal (essentially a typewriter), a Western Electric 103a Data-Set, and a voice-grade telephone line to the nearest IBM installation, which was Cleveland, Ohio. A customer may buy time in six-hour blocks called DATATEXT Agreements. An Agreement covered a time segment from 7:00 a.m. to 1:00 p.m., or from 1:30 p.m. to 7:30 p.m., five days a week. DATATEXT provided everything that the magnetic tape converters did with some important additions. First, it had upper- and lower-case alphabet using a shift character (The Library administration had seen only the Mohawk upper-case converter). Second, the typewriter gave a typed copy which was easy to proofread. Third, corrections were much easier because of the text-editing capabilities of the on-line computer.

Text-editing can best be illustrated by describing a typical DATATEXT job. A typist working from source material produces a typewritten page; at the same time, the IBM 2741 she is using transmits the data being typed to the computer in an area called "working storage". When typing is completed, the clerk gives the appropriate command and the information is stored in an area called "permanent storage", a computer manipulation which can be compared to taking a page from the typewriter and placing it in a folder in a file cabinet. When the typist wishes to make changes to the information, she can give a command to recall it from permanent storage to working storage. She can then manipulate it in several ways. During original entry, the computer automatically assigned numbers to each line. Using these line numbers, a typist can move information within the text, can add or delete information, and can correct errors. Commands are very simple and concise; for example, it takes four keystrokes to move a new line into the text. In making a correction, the typist merely types the incorrect word and the correct word; the computer then types the complete line to show that the correction has been properly executed. (This instant replay, or on-line interaction, is a benefit unique to the on-line terminal.) After any change, the computer automatically renumbers lines and reformats the entire text. A sample of typed input is illustrated and discussed later in the article.

In April 1967, it was decided to test the DATATEXT service because of its powerful correction capability, and because it could be installed and working within three weeks. In May the console was delivered, the telephone equipment installed, and a long-distance line to Cleveland rented. A one-month test of DATATEXT proving successful, three more consoles, data sets and telephone lines were added, and the Conversion Project was fully underway.

TRAINING THE TYPISTS

The majority of the typing and proofreading staff were drawn from existing personnel in the cataloging department. Individuals chosen had a background in either catalog card typing or file maintenance, and consequently a good working knowledge of information on a catalog card. It was anticipated that with a minimum of further training, the typists could identify and tag information as they were typing it at the console. This assumption was critical to the success of the project, since the Library could not afford the professional time necessary for complete pretagging of bibliographic information.

Typists involved in the one-month test were given several hour-long training sessions on tagging before the console arrived. When the project got underway, a list of all possible tags was posted near the console, and a librarian was nearby to answer questions. After three weeks of operation, it was obvious that the typists could tag at the console, thus making this part of the test run a success.

The tagging system used was developed from the MARC I pilot project (10). Most of the original tags were retained and several additional ones designed to meet specific local needs. Tape files created were formatted according to MARC I specifications, although fixed fields were left blank. The tagging system is outlined in a reference manual prepared for typists and proofreaders (11).

Operation of an on-line console requires special training. IBM sent a DATATEXT Instructor to Buffalo on several occasions to provide typist training. For the major training session, which occurred in June, the IBM representative came for a full week. Ten typists were trained; five specialized in entering information, and five specialized in retrieving, correcting, and transmitting information. By the end of the week both groups were skilled in their respective specialities, and many typists were able to perform well in both areas. Later, typists were trained in several sessions by one of the Library's typing staff.

During the first three months, the author was near the terminals at all times to answer questions on terminal operation, to collect data for measuring and controlling performance, and to act as supervisor. A librarian was on call for questions on complex library problems, and the Programmer-Analyst was available to help solve problems regarding input format and tagging. At the end of this period, appropriate clerical staff had been trained to supervise minute-to-minute operation.

CONVERSION PROCEDURES

The general method of conversion (Figure 1) was as follows. A typist typed into "working storage" for an hour, inputting 15 to 30 shelf list cards. She instructed the computer to store this "document" in a permanent storage location on disc. She then placed the typed copy and cards in a proofreading bin, cleared working storage, and started another document.

A proofreader compared typed copy with original cards and indicated any errors. The corrected document then went to a correction typist who "retrieved" the document from permanent storage to working storage, performed the corrections, and transmitted the corrected document to magnetic tape.

The original uncorrected document was left in permanent storage overnight and deleted the following day. Documents were transmitted to tape

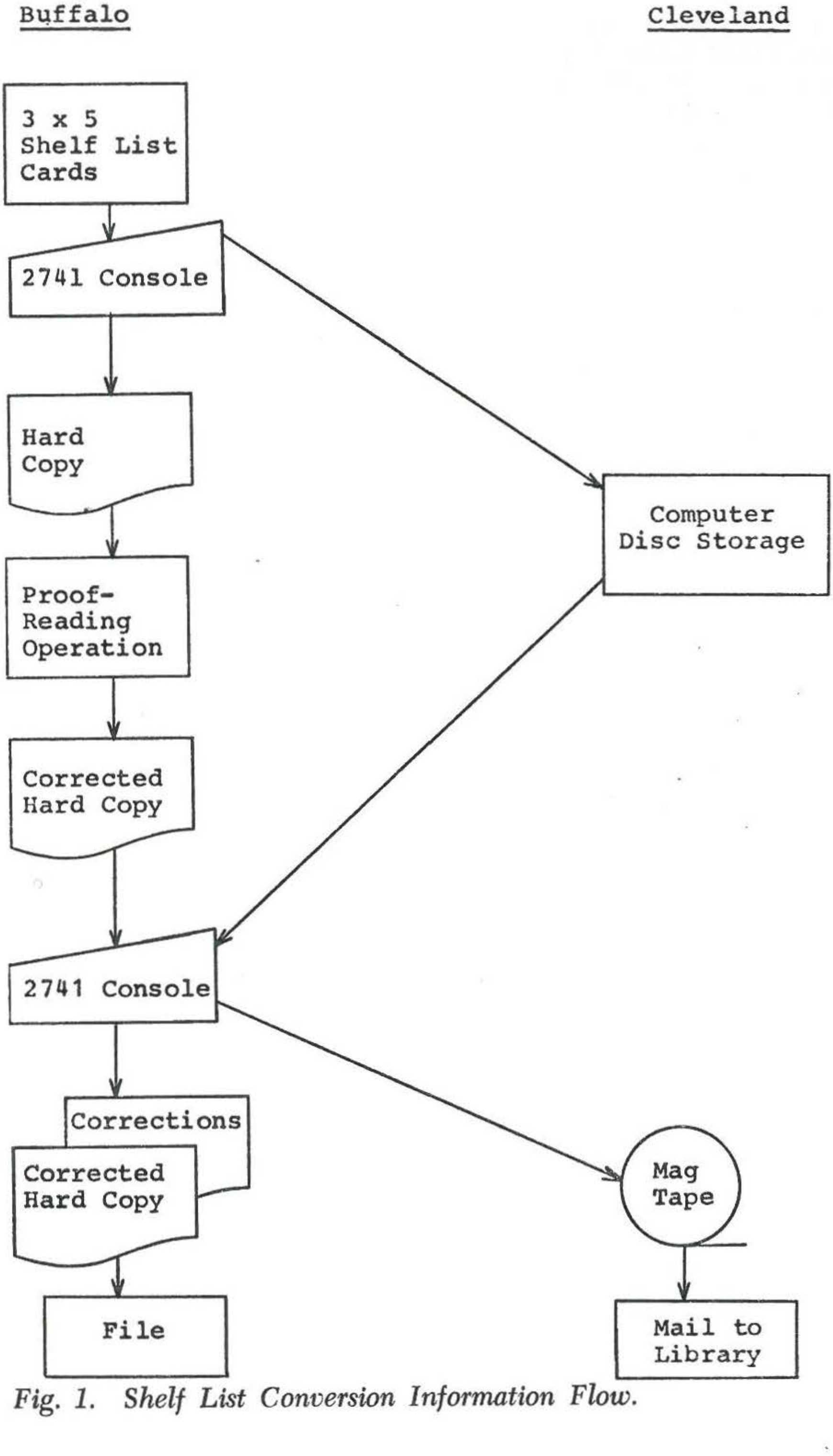

Fig. 1. Shelf List Conversion Information Flow.

for about two weeks and the accumulation returned to the library via the mails. (IBM saved all permanent storage records for one week as a security measure. If a library typist inadvertently deleted a document, it could be retrieved by the computer operator.)

Figure 2 shows a sample of typed input and subsequent correction. Line numbers, as they are stored on the disc, are included on the right margin for ease of explanation. Lines typed in capitals are computer responses to commands, the first entry being the command to clear working storage. The computer responds and then indicates that the console is in one of two general input modes. All cards are typed in "automatic" mode, for which the typist gives the appropriate command. When the computer responds the typist asks for the next line number, which is 3, and begins to input the card. In line 4, the typist makes an error and realizes it before throwing the carriage. She hits the "attention" key pro-

```
c
CLEARED

UNCONTROLLED MODE
a
AUTOMATIC MODE

n
NEXT NUMBER -- 3

90t BS2575.3.A7
10t Bible. N.T. Matthew. English. 1963. New English.          3

20t The Gospel according to Matthew=. Commemen_
                                    ntary by A.W. Argyle.  4

30a Cambridge
30b University Press
30c 1963
40t 227 p. maps. 20 cm.                                        5

50t The Dambridge Bible Commentary: New English Bible          6

70t Bible. N.T. Matthew -- Commentaries.
71t Argyle, Aubrey William, 1910-                              7

73t Title.
60z                                                            8

92t 226.207
94t 63-23728                                                   9

n
NEXT NUMBER  -- 10

6   Dambridge                 Cambridge
50t The Cambridge Bible commentary: New English Bible
```

Fig. 2. Sample Input and Correction of One Shelf List Card.

ducing the underscore, rolls the platen down, back spaces, and retypes the correct word. The computer then corrects the error. In line 6 the typist misspells "Cambridge", but does not realize it before throwing the carriage. The correction is shown at the bottom although the input typist could not have performed it herself; it would have gone through proofreading and back to the correction typist. The correction is made by typing the line number, in this case "6", the incorrect word, "Dambridge", tab, and the correct word. The computer responds by typing out the complete line showing the correction.

Except for a brief period, the shelf list was converted in alphabetic order, and by December 1 shelf list drawers through the E's were completed. Early in the project, some of the literature classification, P and PQ, was converted. Foreign languages in the PQ's gave no particular problems, and typing rates did not drop.

All cards were converted in shelf list order except for those having non-western alphabets. When possible, these were transliterated and entered. Otherwise their input was delayed. Since the 2741 console has no diacritical marks, these were left out; however each card having them was entered and given a special tag to permit retrieval at a later date when diacritical marks could be added by special coding such as used by MARC.

Conversion consoles and shelf list were in the same building. Each day, several inches of cards were removed from the drawer being processed and a marker inserted indicating where the cards had gone. In general operation, cards were returned and refiled in less than a day so that inconvenience to staff was minimal. As a card was proofread, it was marked on the back with a "C" and the upper right hand corner received a very small notch with a McBee punch. Thus, newly cataloged cards filed with cards already converted are recognizable by the unnotched corner.

COSTS

Table 1 gives a statistical summary of the conversion project from July 31 through December 1, 1967. The term "L.C. card" refers to a complete bibliographic entry for a title and may include more than one physical card, or may include writing on the back of a card. Input and correction functions are reported separately and then totaled to give a realistic input rate per hour for corrected cards. Supervisor cost reflects wages of clerical supervisors only. Those of the Programmer-Analyst, the Librarian and the Systems Analyst assigned to the project are not included.

A breakdown of monthly equipment costs per console is given in Table 2. Installation costs were $150 for each terminal, and $50 for each leased telephone line. When the project operated four consoles, the monthly equipment cost was $4,472.

Table 1. Conversion Project Statistics (July 31-Dec. 1, 1967)

Input, Proofreading and Correction		
Total L.C. Cards Input		49,348
Typist Hours Input	3,035	
Typist Hours Correcting	492	
Total Typist Hours		3,527
Proofreading Hours		1,235
Number of Errors per L.C. Card		.42
L.C. Card Input Rate per Hour		16.3
L.C. Card Correction Rate per Hour		100
Overall Conversion Rate (Input & Correction) Cards per Hour		14
Proofreading Rate, Cards per Hour		40
Costs		
Labor Cost @ $1.75 per Hour		$ 8,078.00
Equipment and Supervisors		18,995.00
Total Cost		$27,073.00
Cost per Card Converted		$0.55
Utilization of Console Time		
Hours Typed	3,381	81.4%
Hours Consoles Down	245	5.9%
Hours Computer Down	91	2.2%
Hours Lost Time	438	10.5%
	4,155	100.0%

Table 2. Monthly Operational Costs per Terminal

IBM 2741 Communications Terminal	$ 85.00
Western Electric 103a Data Set	27.50
24-hour voice-grade lease line to Cleveland plus local telephone costs	385.50
2 DATATEXT Agreements @ $310.	620.00
TOTAL	$1118.00

"Hours Typed" is time that consoles were actually being used to input or correct cards. This is slightly less than "Typist Hours Worked" because some correction has been delayed, but it was included in hours worked to give true representation of input rates. "Hours Consoles Down" reflects time lost due to console breakdown. During the early part of the period, two consoles were failing often. However, as operating problems were solved, console down-time dropped far below the average 5.9 per cent shown. "Hours Computer Down" was also greater during early weeks of the project. However, for each hour down, IBM credited the Library with $12.00 ($3.00 per terminal for four terminals).

"Hours Lost Time" reflects periods when a working console could not be manned because of personnel breaks or operator absence. All times are given in console-hours, four consoles operating for one hour being recorded as four hours.

The error rate of .42 errors per card is very low. Allowing 350 characters per shelf list card, typists were making one error for every 830 keystrokes. This translates to about 3 errors per typewritten page of 50 characters per line, 50 lines per page. The Office of Secretarial Studies of SUNY at Buffalo indicates that this rate is well within the tolerance for "normal" typing, as in a typing pool. When it is considered that typists were tagging and inputting complicated bibliographic information, rate of accuracy was commendably high.

Typists used in the project included the lowest salary grade of civil-service typists, part-time hourly workers, and students. An acceptable input rate for civil service typists was 18 cards per hour, which is equivalent to 21 5-character words per minute. The faster typists, at 26 cards per hour, were typing at 30 words per minute. Again, let it be mentioned that the material was complex and that typists were required to tag each piece of information.

CONCLUSIONS

Several points can be made about converting with DATATEXT. It was easy to implement and received excellent support from IBM. The IBM Information Marketing staff in Cleveland provided constant assistance during the early part of the installation and visited often once the project was successfully underway. IBM sent the DATATEXT instructor as often as needed and provided free computer time during teaching sessions.

The four long-distance telephone lines and Data Sets proved reliable. There was only one instance during the period when a line was inoperable and it was repaired in three hours. The liaison and support from New York Bell Telephone was very good.

DATATEXT costs would have been lower had the IBM installation been nearer. Cleveland is 173 miles from Buffalo giving a 24-hour lease-line cost of $342 per month. (DATATEXT service will soon include a uniform long-distance-lines cost.)

Verification or correction on DATATEXT does not require human retyping of each line of entry. Only the word in error and its replacement need be typed; the console then types the corrected line to show that the error was deleted and the replacement inserted. Consequently correction costs are low and corrections accurate.

Average rates and costs given in Table I reflect learning during the first six months of the project. Towards the end of the reported period, rates were improving and costs decreasing. Since December 1967, the project has added three more consoles and uses a DATATEXT service provided by a campus computer. Costs have dropped below $.45 per card, a figure which will increase somewhat when diacriticals are added. Potentially cost per title for complete conversion is under $.50.

REFERENCES

1. 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.
2. Hammer, Donald P.: "Problems in the Conversion of Bibliographic Data — A Keypunching Experiment," *American Documentation*, 19 (January 1968), 12-17.
3. Black, Donald V.: "Creation of Computer Input in an Expanded Character Set," *Journal of Library Automation*, 1 (June 1968), 110-120.
4. Buckland, L. F.: *Recording of Library of Congress Bibliographical Data in Machine Readable Form* (Rev. ed.; Washington, D.C.: Council on Library Resources, 1965).
5. The Johns Hopkins University. Milton S. Eisenhower Library: *Progress Report on an Operations Research and Systems Engineering Study of a University Library* (Baltimore: Johns Hopkins, 1965).
6. International Business Machines Corporation. Federal Systems Division: *Report of a Pilot Project for Converting the Pre-1952 National Union Catalog to a Machine Readable Record* (Rockville, Maryland: IBM, 1965).
7. Lazorick, Gerald J.; Herling, John; Atkinson, Hugh: *Conversion of Shelf List Bibliographic Information to Machine Readable Form and Production of Book Indexes to Shelf List* (Buffalo, N.Y.: State University of New York at Buffalo, Technical Information Dissemination Bureau, 1966).
8. Mohawk Data Sciences Corp.: *DATAGRAM No. 35, 1181 TWK Correspondence Data-Recorder,* (Herkimer, N.Y., Mohawk Data Sciences Corp., 1967).
9. International Business Machines Corporation: *DATATEXT Operators Instruction Guide,* Form # J20-0010-1 (IBM, White Plains, N.Y., 1967).
10. 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).
11. Michael M. Coffey: *Reference Manual for Typists and Proofreaders. SUNYAB Shelf List Conversion Project* (Buffalo, N.Y.: SUNY at Buffalo, Technical Information Dissemination Bureau, 1968).