Text Capture and Electronic Conversion

Background

Members of the Cataloging Directorate at the Library of Congress are experimenting with ways of doing more cataloging with fewer resources, trying to get more out of an aging computer input- update system than the system can provide, and exploiting the new bibliographic workstation (BWS) that is being installed on the catalogers' desks. To those ends, we have created a package of utilities known as Text Capture and Electronic Conversion (TCEC). These utilities allow cataloging staff to take electronic information in various formats and convert that data to useable LC MARC records which are then cut-and-pasted into the LC computer catalog.

Equipment: Bibliographic Workstation

The Library of Congress is in the process of installing the BWS on catalogers' desks to replace the aging dumb terminals currently in use. The BWS is both an input/update terminal for the Library's computer catalog files as well as a PC capable of performing many work-related tasks (WordPerfect, Internet access). All BWSs run with 16 MB RAM under OS/2 version 2.1 and TCP/IP for OS/2 version 2.0 with screen resolution of 1024 x 768 with 256 colors. The machine configurations are:

```
IBM PS/2 model 70    386     33MHz   110 MB hard drive
8514 video adapter  IBM PS/2 model 70    386     33MHz
160 MB hard drive   XGA video adapter  IBM PS/2 model 77
486DX2  66MHz   203 MB hard drive   XGA-2 video adapter
```

The BWS is an off-the-shelf personal computer that is loaded with off-the-shelf software to carry out most of its functions. To do the cataloging work, however, special software had to be written to enable use of the full ALA extended character set. Data Connections, a company in England, was hired to write the OS/2 program to enable input/update functions on the BWS as well as the extended character set. This is custom software written specifically for the Library's BWS and mainframe interaction. Also, custom keycaps were created for the Library. These keycaps replace the standard keycaps on an otherwise ordinary keyboard. These keycaps contain the normal character set as well as the ALA extended set. They are also designed to reduce glare, which was a problem with the dumb terminals. Also written for the Library was a program to enable the use of macros in cataloging. The catalogers can create approximately 100 macros for use in their work. Currently the macro capability is limited to actual keystroke types of commands. Future improvement requests to the macro capability include the ability to read the date from the PC and insert it into the

MARC record where needed for tracking purposes, the ability to print out a cataloger's macro package, the ability to chain macros together, and the ability to create even longer macros than are now possible.

Electronic CIP (E-CIP):

The Electronic CIP experiment started in April of 1993 with the idea of having publishers use the Internet (FTP) to submit manuscripts for books to be cataloged. In November 1993, the University of New Mexico Press submitted the first title to be cataloged. Since that time, the Library has received 102 manuscripts for cataloging from the 7 participating publishers-- University of New Mexico Press, University of Arizona Press, University of South Carolina Press, University of Tennessee Press, Utah State University Press, HarperCollins (adult division) and HarperCollins (children's division).

Using TCEC techniques, the cataloger supplies ISBD punctuation in the text, highlights the information to be included in one MARC field (e.g. the 245 title, subtitle, and author statement), then tells the machine what that data represents by clicking on a MARC tag button (i.e. the 245 button). The program instantly creates the appropriate MARC field in LC MARC format and displays the field in a window at the bottom of the screen. The cataloger continues and creates all of the descriptive cataloging fields possible from the manuscript text, creating most of the descriptive elements of a MARC record. The resulting record is then cut-and-pasted into an LC record creation screen and the cataloger simply presses the ENTER key to get the record into the LC system.

The program is also capable of creating a name authority record based on the manuscript and the MARC record created if the cataloger needs to make a name authority. The name authority creation routine creates the heading, source citation, and reference structure if needed.

[An expanded explanation of this as well as graphics on the procedures are available on the World Wide Web for review at http://sscd2.loc.gov/ecip/onmarc.htm (note .htm not .html as is customary).]

Internet as a source for cataloging:

The same program that is used to create the E-CIP records can be used for any type of text in electronic form from any source. Experiments and demonstrations have shown that it is possible, using OS/2 and the OS/2 Gopher client, to go out on the Internet, search an OPAC for bibliographic data, use the OS/2 window text highlighting capability to define an area of text, copy that text into memory in the OS/2 clipboard, exit the OPAC, and then retrieve the text into the E-CIP cataloging program. Since library OPACs frequently present their data in card format or with ISBD punctuation present, the cataloger simply highlights the data elements and tells the machine what they are and builds a MARC record based

on that OPAC screen image. The resulting MARC record can be used as the basis for a catalog record.

Preassigned card numbers:

The CIP Division runs two programs, the CIP program where full cataloging data is supplied and printed in the book, and the Preassigned Card Number (PCN) program where publishers can get a Library of Congress Card Number (LCCN) to print in their book. With the PCN program, an experiment is underway to enable publishers to fill out an electronic form on the World Wide Web. Publishers fill out the form, click on a SUBMIT button, and the data they supply is converted to a file that is saved on a server at LC. PCN staff then call up a program which scans this file and creates 2 products: a preliminary cataloging record and an electronic mail form letter. These products are then cut-and-pasted into the LC catalog system and the Email system and processed. Both of these products use information that was retrieved from the file. Obviously the cataloging record gives the title, publisher, ISBN, etc. but also the form letter uses data such as the requestor's name and address, the title of the book, and the LCCN which was provided to the computer by the CIP staff. These elements are stored in variables in the computer and used as often as needed for the various products being generated. This experiment dramatically shows the keystroke savings between the preliminary record and the electronic form letter that are produced.

Other efforts:

While the Electronic CIP project is the most visible project that uses TCEC techniques, there are other efforts underway that involve TCEC developments. Two efforts involve name authority records. The British Library is submitting name authority records to LC on paper. These authorities are scanned and converted to electronic format. After this, they are converted into LC name authorities by character recognition routines and then are cut-and-pasted into the name authority file. This procedure eliminates the re-keying of the data and is more accurate.

Another name authority project involves LC's Overseas Offices. Some of the overseas offices create name authorities on a PC-based cataloging package called Minaret. Minaret creates MARC records in U.S. MARC format. These records are exported to disk in MARC format, sent to Washington, read into a program and converted into LC MARC format records. Once again the person creating the name authority uses a cut-and-paste operation to copy the record into the LC authority file.

Both of these projects have shown that electronic processing can indeed speed up the process of getting name authorities into the LC system as opposed to re-keying the records. Particularly with the overseas office NARs, if the inputter is

unfamiliar with the language involved, these techniques can provide a more accurate record than re-keying.

Other projects are still in the experimental stages but also rely on programs that manipulate electronic data and produce an LC MARC record which can be cut-and-pasted into the LC system. All of these projects reduce the keystrokes required to complete cataloging tasks and also save time. They are also very accurate, assuming the original text was accurate.

David Williamson

Senior Descriptive Cataloger

Romance Languages Team

Social Sciences Cataloging Division

Library of Congress

Washington, D.C. 20540

(202) 707-6424

dawi@loc.gov or

williams@mail.loc.gov

Team OS/2