Using the SAS System to Generate Bibliographies from a Large Data Set

Prepared by: Mark A. Crook, OCLC Online Computer Library Center, Incorporated

Introduction
OCLC (r) Online Computer Library Center maintains the largest online database of bibliographic information in the world. This database, the Online Union Catalog, represents a unique source of information from which extensive bibliographies can be derived automatically using specially designed computer programs to retrieve, format, and print bibliographic records according to users' specifications.

In this paper I will discuss (1) the application of the SAS (r) software to problems of data retrieval and formatting specific to the OCLC internal bibliographic record format, (2) the challenges of interfacing this machine-readable output to various printing devices, and (3) the benefits and shortcomings of the computer-assisted approach to bibliography compilation.

Background
The impetus for using the SAS system to assist in bibliography compilation came as an offshoot of the OCLC Office of Research Information Services project. Information Services provides decision management services to OCLC staff based on information derived from the OCLC Online Union Catalog. To accomplish this, the Information Services team developed reusable SAS routines to do data extraction and manipulation and to produce statistics. This library of routines now permits Information Services to probe heretofore inaccessible portions of the database with relative ease and with minor modifications to existing programs. Building this library of routines was not a simple matter for two reasons: database size and record format. The OCLC Online Union Catalog comprises over 19 million bibliographic records. This makes data processing against such a file a daunting task; program efficiency is a constant watchword.

Furthermore, OCLC's internal record format is difficult to work with: variable-length records that contain both fixed and variable-length portions. The flexibility required to properly encode bibliographic information leads to a high degree of variability in the records.

These two obstacles were effectively handled by the SAS language. The code required to give complete access to all parts of the bibliographic record is only about 20 lines (BIBSAS example). In most cases, a selection decision could be made about a record without having to completely parse its contents; the record would simply be written to the output file with a PUT INFILE statement for further processing.

Processing efficiency is attained by employing random sampling datasets wherever possible and using the DATA NULL statement when doing record selection or manipulation. The combination of these two techniques leads to a significant reduction in the amount of computer resources required by most of the information requests.

However, there are some requests that require a full scan of the database. In such cases, the SAS code is streamlined to do as little record manipulation as possible and placed in the input queue to be run overnight. Usually, a complete scan of the OCLC database requires about two hours. By using these SAS routines the Information Services project was able to provide 24 hour turnaround time on an IBM (r) 3090 running MVS XA for most information requests.

Because of the success of the Information Services project, OCLC now offers an information requesting service, the Scholarly Services project, to the scholarly community. Scholarly Services is directed to the individual who is doing research either about OCLC, library and information science, or a related discipline. The most common request from participating scholars has been for a gross bibliography about a certain subject area. As a result, Information Services has concentrated on making the compilation of these bibliographies more automated and less time-consuming and labor-intensive.

Record Formatting
The Information Services staff had long provided crude printouts of bibliographic records retrieved from the Online Union Catalog for internal users. However, for external users, the bibliographies must be much more polished in terms of formatting, displaying the characters, and ordering of the items. For that reason we chose the personal computer-based Pro-Cite (tm) software package to be the target system for our records. Transferring OCLC bibliographic records to Pro-Cite involved several challenges.

First, the internal format must be converted into one that Pro-Cite accepts, a screen image format. A short SAS program (BIBPRTD) allowed us to insert the necessary control characters...
Experts' sophisticated options permit output records from this formatting data in an appropriate manner. The output records from this formatting program were written to a disk file for downloading to the PC from the IBM 3090. Following the download process, Pro-Cite's Biblio Link (tm) to OCLC software is used to download screen images to a Pro-Cite database. Pro-Cite's sophisticated options permit manipulation of the bibliographic data into a desired output form, including various citation styles, for specific output devices. A Pro-Cite user could direct the output to the screen, the printer, or a disk file. In addition, the Pro-Cite database itself is a medium of exchange for scholars.

Unfortunately, we discovered that in bypassing the OCLC Online System to generate the screen images, the special characters and diacritics remained as non-standard ASCII characters in the data (notice, the Office of LIBRARY ASSOCIATION character set). This presented a serious problem for clean display.

Again SAS programs came to the rescue. The first thing that had to be done was to either print or eliminate the extended characters. To accomplish this, a rather lengthy translation program was written in the SAS language (BIBXLT example) to convert the non-standard characters to Epson (r) graphic mode equivalents or eliminate those characters which could not be printed on the Epson printer. Because the program looked at every character, it took a substantial amount of time (usually less than 10 minutes) to complete the translation on the Pro-Cite input file of several hundred records. This printed output was acceptable, but it opened up other possibilities for printing on other devices available at OCLC.

Interfacing to Other Printing Devices

The Epson printer was slow (in letter quality mode) and could not print the entire ALA character set represented by the bibliographic records. As a result, alternative printing options were necessary to generate acceptable output.

One alternative was to print this data on the PostScript (r) laser printer. This printer offered the advantages of greater speed, an extended character set, and camera-ready output. The difficulty again was to make the printer recognize and print the special characters. Fortunately, the Office of Research licenses TeX (tm), a markup language that makes printing difficult texts a routine process. Now the challenge was to convert the special characters into TeX control sequences to get the data to print correctly. A SAS program (BIBTEX example) was written to accomplish this translation.

Originally, this program ran on a personal computer and the output was transferred to a Sun (r) system where TeX and the PostScript device resided. When OCLC purchased the SAS System for the Sun workstations, the code was immediately ported and run directly on the Sun system with a ten-fold increase in processing speed. Output from the translation program was then input to TeX which created a device independent (dvi) file which was routed to the PostScript printer. The resultant output was high quality, but quantity and speed of output were insufficient for large volume applications.

Since OCLC also owns a Xerox (r) 8700 Electronic Printing System, which is rated at 80 pages per minute and can print the entire extended character set, it became a logical output device for the data. Instead of doing lengthy translations of the data on the personal computer, the Pro-Cite output file was uploaded to the 3090 for eventual printing on the Xerox 8700. This time there were no translation problems because the special characters were included in one of the fonts on the Xerox 8700. However, overstriking diacritics and page formatting were necessary to generate quality copy. One more time a SAS program (PCXLT example) efficiently solved the problem and permitted high-quality, multiple-copy, large-volume output.

As you can see, SAS programs were therefore instrumental in every facet of the Sytem development. Now the generation of a large bibliography can be completed in about two hours after the initial data selection is completed overnight. Compare this to the labor-intensive process of collecting bibliographic citations in the traditional manner.

Benefits and Shortcomings

In his classic work, On Compiling an Annotated Bibliography, James L. Harner observes that the compilation of a bibliography is a process of "ordering the mass of scholarship on an author or subject" (1985, 1). Harner goes further to suggest that many scholars might consider "an annotated bibliography is second-class scholarship (or worse, merely inspired clerical drudgery)" (1985, 2).

If we accept Harner's statements, adding computer assistance to the compilation work of a bibliography facilitates the ordering of the mass of scholarship and reduces the amount of drudgery involved in the process. Indeed, computer assistance in searching and ordering the data permits a bibliographer to be "organized, accurate, thorough, and meticulous in the data"
gathering phase of bibliography
development (Harner 1985, 4).

The fact that the bibliographic
citations are entered in machine-
readable form permits subsequent
manipulation and output on various
printing devices. Using application
systems specific to bibliography
generation permits further flexibility
in storing, ordering, formatting, and
outputting the compiled information.

These benefits allow the scholar or
bibliographer to spend more time doing
the necessary linking and investigation
of additional leads and less time
searching for items. In other words,
this form of computer assistance removes
the clerical aspect and allows the
bibliographer to do more intelligent
research and offer more thoughtful
annotations.

There are obvious shortcomings,
however, to the computer-assisted
approach: omission of items not included
in machine-readable databases, incomplete
retrievals of related items, and total
reliance on a computer generated process.

In spite of the pervasiveness of
machine-readable databases, there are
still many items that are not accessible
by computer. This problem is especially
acute in research in the humanities
where older and rare materials are
important, but difficult to locate due to
relative disinterest in entering
their contents or description online.

Even the 19 million bibliographic entries
in the OCLC database represent only a
fraction of the items held in member
libraries. Pertinent items not entered
in the database are missed by this
method of bibliography compilation.

While omission of items is certainly
possible when compiling a bibliography
by hand, the fact that one uses a
computer does not guarantee completeness.

Even if the user exhausts the keywords
believed to be appropriate to the
subject, he cannot be guaranteed of the
retrieval of every scrap of information.
Spelling errors, subjective decisions
about keywords and headings, and non-
indexed items all contribute to the
doubt surrounding the exhaustiveness of
a given bibliography.

Finally, the most insidious of the
shortcomings of this technique is the
complete reliance on the computerized
process. In The Cult of Information,
Theodore Roszak (1986) warns that we
should not sublimate our intuitions,
intellects, and ideas to the impression
that a computer can do everything for
us. More information, even in a biblio-
graphy, is not knowledge. In this case
the compiled data must be a point of
departure for tough evaluation of the
cited items and insightful annotation.

Conclusion

The SAS language provides an
excellent tool for doing data retrieval
and manipulation, even in large data-
bases. The flexibility of the SAS
language and the relative ease of using
it to generate general and special
purpose routines facilitate interface
to other personal computer application
software and output devices.

Computer-assisted bibliography
compilation is a timely idea. Harner
(1985, 27) predicted that "it will not
be many years before most bibliographies
will be published electronically on
either disks or tape or through
commercial vendors..." Today, it is
possible to produce computer-generated
bibliographies that can be distributed
either in machine-readable or paper
format. Indeed, OCLC's forthcoming
reference system will permit end-users
to generate such listings from an
online session. However, it must be
realized by the recipients of such
listings that they are only the first
step in the research and compilation
of a scholarly, annotated bibliography.

References

Harner, James L. 1985. On compiling
an annotated bibliography. New York:
Modern Language Association of America.

Roszak, Theodore. 1986. The cult

SAS is a registered trademark of SAS
Institute, Inc.

Pro-Cite and Biblio-Link are trademarks
of Personal Bibliographic Software, Inc.

TeX is a trademark of the American
Mathematical Society.

PostScript is a registered trademark of
Adobe Systems, Inc.

Sun is a registered trademark of Sun
Microsystems, Inc.

IBM is a registered trademark of Inter-
national Business Machines Corporation.

Xerox is a registered trademark of
Xerox Corporation.

Epson is a registered trademark of
Seiko Epson Corporation.