Tackling ICPSR Online Codebooks With Success

by Jackie Shieh [1]
ESRC Data Archive
University of Essex,

INTRODUCTION

The focus of cataloging the Inter-university Consortium for Political and Social Research (ICPSR) online codebooks is to provide users in a timely fashion adequate bibliographic information on VIRGO, the University of Virginia Library's computerized library system. Many catalogers today are cataloging materials that cannot be held in hand. Gathering bibliographic information for electronic formats can be a bewildering and monstrous experience. The author shares her experience on how the fear of working with computer files was reduced to a minimum with the help of the computer support department, and the sense of triumph and accomplishment she felt when patrons successfully retrieved what they needed through the online catalog!

BACKGROUND

The ICPSR is one of the world's leading repositories and data dissemination organizations for machine-readable social sciences data. ICPSR receives, processes, and distributes machine-readable data on subject matters covering over 130 countries. The content of the ICPSR archive extends across the spectrum of economic, sociological, historical, organizational, social, psychological, and political concerns. ICPSR was founded in 1962 as a partnership between the Survey Research Center at the University of Michigan and 21 universities in the United States. Currently, membership extends to over 370 colleges and universities world-wide. The University of Virginia (UVa) faculty and graduate students (undergraduates with the permission of an instructor) may order ICPSR data free of charge by contacting the Official Representative in the Social Science Data Center.

There are currently over 350 studies available at UVa. Most of the studies consist of more than one dataset, and often include a study description and a codebook. The codebooks which are in print or machine-readable format describe the location of the variables in the data record.

VIRGO, a NOTIS-based online catalog provides online access to the Library's holdings through keyword, author, title, subject, and call number searches. It also offers online access to: nine periodical indexes published by the H. W. Wilson Company; CURRENT CONTENTS (an index to recent issues of over 6500 scholarly journals); ABI/INFORM (provides citations and abstracts from business and management journals); and NEWSPAPER ABSTRACTS (indexes and abstracts articles in 28 major newspapers).

CATALOGING PROJECT FOR ONLINE CODEBOOKS

Although, Alderman Library began a cataloging project for paper codebooks in the summer of 1994, the project of cataloging the machine-readable codebooks did not take place until January 1995 with the arrival of the Original Cataloger for Electronic Resources. Before the project began for the machine-readable format, there were several issues that needed to be considered — 1) the number of titles incoming and in backlog, 2) the status of acquisition, 3) the procedure of retrieving bibliographic information, 4) the cataloging procedure, namely the procedure from getting actual text file to OCLC MARC format, and 5) the limitation for access of the materials.

Currently, the online codebooks reside in the machine named Maggie under the directory, /archive/public/icpsr at the University's Information Technology and Communication (ITC)[2]. They are accessible via all networks at the University. The files are sub-organized by the ICPSR series number from 0001-9999. Each individual series contains at least two basic files —the codebook and statistical data files.

To catalog a codebook, the following information needs to be obtained —the actual title, statement of responsibility, edition, file characteristics, physical description, series information, publication or distributor, special note and terms of availability, etc. The chief sources for the bibliographic description are taken from the title screen and table of contents.

Take for example, *CENSUS OF POPULATION, 1910. UNITED STATES: PUBLIC USE SAMPLE* (ICPSR ; no. 9166). To obtain the title proper, one can either pull up the title screen of codebook 9166 online from the sub-sub-directory of series number 9166 (*See* Figures 1 and 2), or consult the paper format reference book, ICPSR's *Guide to Resources and Services, 1994-1995* . Some of the series only have online codebooks, while others have both online and print versions. The Social Science Data Center at the University does not have the complete collection of neither online nor print version. Series titles are added as the faculty place subscription orders which result in the increase of the database, thus the ICPSR is a growing collection.

Once the bibliographic information is recorded, the title is searched against VIRGO and the national bibliographic utility, OCLC. If a record is found in VIRGO for paper

format (even if it is a different edition or release), the original cataloging for the computer file is created using the DER (derived) feature from the existing bibliographic record. If no VIRGO record is available and an OCLC record is found for a different format or edition, the OCLC printout is used to create an original record on VIRGO. All original cataloging for machine readable datafile are created locally on VIRGO. Currently, tape loading of machine-readable MARC records without field 856 from VIRGO to OCLC is not available.

FILTER PROGRAM USING PERL

The initial stage of the ICPSR project required the gathering of bibliographic descriptions on individual series numbers from the ICPSR directory. That task seemed far more cumbersome than the creation of the MARC record. It became even more overwhelming when it was discovered that new titles are checked-in and constantly added to the database. Therefore, in order to manage the existing and incoming codebook files tracking uncataloged materials became the priority.

Jeff Herrin, of the Library System Office wrote and explored a filter program using PERL (<italics>*See* <end italics> Figure 3) to read all current files in the directory of /archive/public/icpsr, from file 0001 to 9999. PERL copied the first 70 lines of text files, which corresponded to '*.codebk*', then copied the readings into an output file. The program would also generated an e-mail message notifying me even if there was no successful hit after the run. The first time PERL was run we had little difficulty. A few minor adjustments were made, especially when two codebooks for the same series number were present. For instance, series 8166 had two online codebooks, i8166.codebk and i8166.codebk01.

When examining the output of each file, I found that it did contain sufficient information for creating a cataloging record— the size of the file, title proper, author, publisher and publication date information (<italics>*See*<end italics> Figure 4). Although, the filter program was a success, the Social Science Data Center cautioned us that incoming materials are being deposited to their appropriate subdirectories regularly. In response to this, a cron daemon which runs the filter program automatically is scheduled every other month[3].

The VIRGO template (<italics>*See*<end italics> Figure 5) for ICPSR online materials was created to facilitate the cataloging process. Faculty, students and staff from the University can retrieve the information as soon as the bibliographic record is created.

CONCLUSION

After implementing the PERL filter program and creating the VIRGO template (similar to constant data on OCLC), cataloging online ICPSR series became more manageable than previously perceived. Yet, tapeloading from VIRGO to OCLC still presents a challenge for the Library. OCLC currently does not allow tapeloading on machine-readable original cataloging. Since the Library is committed to information and resource sharing, it means that contributing these bibliographic records requires additional manual work. Each record must be re-created for OCLC holdings. The Library is looking forward to the day when cataloging records for machine-readable format can be transferred

Figure 1

```
archive/public/icpsr% dir
        drwxr-xr-x    4    3180  public  8192  Nov 23 09:07 0001-5999/
        drwxr-xr-x   14    3180  public  8192  Nov 23 09:07 6000-6999/
        drwxr-xr-x    9    3180  public  8192  Nov 23 09:07 7000-7250/
        drwxr-xr-x         3180  public  8192  Nov 23 09:08 7251-7500/
        drwxr-xr-x    7    3180  public  8192  Nov 23 09:08 7501-7750/
        drwxr-xr-x   10    3180  public  8192  Nov 23 09:09 7751-7999/
        drwxr-xr-x   11    3180  public  8192  Nov 23 09:09 8000-8250/
        drwxr-xr-x   15    3180  public  8192  Nov 23 09:09 8251-8500/
        drwxr-xr-x    5    3180  public  8192  Nov 23 09:09 8501-8750/
        drwxr-xr-x    7    3180  public  8192  Nov 23 09:10 8751-8999/
        drwxr-xr-x    9    3180  public  8192  Nov 23 09:10 9000-9250/
        drwxr-xr-x   12    3180  public  8192  Nov 23 09:10 9251-9500/
        drwxr-xr-x   12    3180  public  8192  Nov 23 09:10 9501-9750/
        drwxr-xr-x   18    3180  public  8192  Nov 23 09:10 9751-9999/
```

Figure 2

```
archive/public/icpsr/9000-9250/9166% dir
        -rw-r--r--   1   3180  public  408240     Nov 28 11:22   i9166.codebk
        -rw-r--r--   1   3180  public  50965936   Nov 21 11:36   i9166.odata 1
```

Figure 3

```
#!/usr/bin/perl

$user="userid@virginia.edu";
$base="/lv2/users/userid/icpsr";
$tmpfile="$base/invent";
$pages="$base/pages";
$lastscan="$base/lastscan";

open(NEW, "find /archive/public/icpsr/*/* -name *codebk* -newer $lastscan -type
f -printl");
open(TMP,"> $tmpfile") || die "Can't open tmp file!\n";
print TMP "Here's the new ICPSR codebooks:\n\n";

while($file=<NEW>) {
        print TMP $file;
        chop $file;
        $number = $file;
        $number =~ s/.*i([0-9]*)\.*cod.*/$1/;
        $edition = $file;
        $edition =~ s/.*\.codebk(.*)/$1;
        $size = -s $file;
        open(CODEPG,"> $pages/$number.$edition") || die "Fail to open!\n";
        printf CODEPG "Size: %s bytes\n", $size;
        chop $file;
        $number = $file;
        $number =~ s/.*i([0-9]*)\.*cod.*/$1/;
        $edition = $file;
        $edition =~ s/.*\.codebk(.*)/$1;
        $size = -s $file;
        open(CODEPG,"> $pages/$number.$edition") || die "Fail to open!\n";
        printf CODEPG "Size: %s bytes\n", $size;
        open(CODEBK,$file);
        while(<CODEBK>) {
                print CODEPG;
                last if $. > 80;
        }
        close(CODEBK);
        close(CODEPG);
}
`touch $lastscan`;
printf TMP "\nThe first pages are in $pages\n";
close(TMP);
close(NEW);
system("mailx -s \"New ICPSR items\" $user < $tmpfile");
system("rm -f $tmpfile");
```

figure 4

Size: 408240 bytes
1

CENSUS OF POPULATION, 1910 ^MUNITED STATESY:

PUBLIC USE SAMPLE

(ICPSR 9166)

Principal Investigator

Samuel H. Preston
University of Pennsylvania

First ICPSR Edition
Spring, 1989

Inter-university Consortium for
Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106

1

BIBLIOGRAPHIC CITATION, ACKNOWLEDGMENT OF ASSISTANCE
AND DATA DISCLAIMER

All manuscripts utilizing data made available through the Consortium should acknowledge that fact as well as identify the original collector of the data. In order to get such source acknowledgment listed in social science bibliographic utilities, it is necessary to present them in the form of a footnote or a reference. The bibliographic citation for this data collection is:

Preston, Samuel H. CENSUS OF POPULATION, 1910 ^MUNITED STATESY: PUBLIC USE SAMPLE ^Mcomputer fileY. Philadelphia, PA.: University of Pennsylvania. Population Studies Center, 1989 ^MproducerY. Ann Arbor, MI.: Inter-university

(END)

Figure 5

```
UL- ALK8984 FMT D RT m BL m DT 01/04/95 R/DT 01/25/95 STAT nn E/L   DCF a D/S D
 SRC d PLACE miu LANG eng MOD   T/AUD   D/CODE ? S/STAT ? DT/1 ???? DT/2
 DF/TYP d MACH   FREQ   REG   GOVT

 040:  :  a VA@  c VA@
 049:  :  a VA@@
 090/1:  :  a H62  b .I25 no.
 100:1 :  a <author>
 245:10:  a <title>
 256:  :  a Computer data (1 file : ca.  kilobytes).
 260:  :  a  Ann Arbor, Mich. :  b Inter-university Consortium for Political and Social Research,  c
<year>.
 490/1:1 :  a ICPSR ;
 500/1:  :  a Codebook to accompany related data tape.
 516/2:  :  a Text.
 516/3:  :  a <Numeric (Summary statistics).>
 520/4:  :  a <Optional.>
 500/5:  :  a <Also available in paper format.>
 580/6:  :  a Issued also in paper format, titled:
 537:  :  a Hard copy documentation (year) transformed into machine-readable text utilizing
Optical Character Recognition (OCR) Scanning, date.
 650/1: 0:  a <subject>
 700/1:10:  a <personal author>
 710/2:21:  a Inter-university Consortium for Political and Social Research.
 710/3:21:  a <corporate author>
 830/1: 0:  a ICPSR (Series) ;  v
 856/2:7 :  m Social Science Data Cender and ICPSR services, (804) 982-2630  u gopher://
gopher.lib.virginia.edu:70/11/socsci/icpsr  2 gopher.
```

successfully via either INTERNET FTP or tapeloading. The less editing required on one record, the more reliable the information remains.

1. Paper presented at IASSIST 95 Quebec City, Quebec. Jackie Shieh is Original Cataloger for Electronic Resources at Alderman Library, University of Virginia Library, e-mail ejs7y@Virginia.edu

2 The ITC is the equivalent of Computer Center in other institutions.

3. In UNIX, the cron daemon runs shell commands at specified dates and times. Regularly scheduled commands can be specified according to instructions contained in the crontab files. The cron daemon examines crontab files and at command files only when the cron daemon is initialized.

4. Tapeloading is available for the Internet Cataloging Project participating libraries under certain guidelines, Erik Jul's *Building a Catalog of Internet-Accessible Materials: Project Overview.* URL:http://www.oclc.org/oclc/man/catproj/overview.htm.