Paper 055-2022

SAS as a Tool in Data Curation: A Case Example with the Inter-university Consortium for Political and Social Research

Piotr Krzystek, ICPSR at the University of Michigan, Ann Arbor, MI

ABSTRACT

The Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan specializes in archiving social science data by utilizing various practices, such as data curation, to make its objectives possible. Data curation is the practice of taking deposited data and documentation, running various forms of quality check to “clean up” the information, and then releasing the materials online for interested individuals to download and analyze on their part. Various software tools and statistical packages are utilized in the data curation process, including SAS. The purpose of this paper is to discuss how SAS is utilized as part of the data curation process in the setting of ICPSR. The outline for this paper includes discussing what ICPSR and data curation are, the SAS procedures that are utilized with each step of the data curation process, how issues that may arise in using SAS are remedied, and the potential for SAS in benefiting and improving the curation process more in the future. The version of SAS utilized to perform the data curation tasks is 9.4, and users do not need to have prior experience with the program to edit and run code.

INTRODUCTION

The Inter-university Consortium for Political and Social Research, or ICPSR, is a membership-based organization located within the University of Michigan’s Institute for Social Research. Considered to be one of the largest archives for social science research data in the world, the organization’s 850+ participating members have access to numerous studies which individuals can perform their own analyses and research on. Between the practices of Principal Investigators (PIs) depositing their studies and users being able to download their accompanying datasets, data curators work on making sure that the information has met all appropriate standards.

Data curation is a set of practices performed with the purpose of processing data and accompanying documentation so that it is devoid of any issues (e.g., inconsistencies, out of range values) and meets certain standards; the latter mostly related to maintaining confidentiality of information related to sensitive topics and producing metadata in DDI. At ICPSR, various statistical tools and software (including processing tools developed by ICPSR) are utilized in the data curation process, with SAS® being one such program. In the process of data curation, issues can arise ranging from data not converting to other formats properly to certain types of outputs not generating. The purpose of this paper is to focus on the aspects of the data curation workflow that involve utilizing SAS, the issues that may arise in such work and solutions that are utilized in resolving these problems, and the potential for using SAS to benefit and improve the curation process in the future.

DATA CURATION AT ICPSR

The practice of data curation can be defined basically as the standardization of data. Since the specific tasks and concepts of data curation can vary between different groups that engage in this practice, the current focus will be on how data curation is performed by the Inter-university Consortium for Political and Social Research. At ICPSR, data curation is usually defined by an established workflow of tasks, certain levels of curation work, and access levels that determine how a study is released. The workflow that is usually laid out includes initial work on files for a study (e.g., data, documentation) followed by various forms of quality check before the release of a study. Each study that is worked on has differently designated levels of curation, with Level 1 studies receiving minimal work while Levels 2 and 3 include additional processing tasks. Lastly, the access level of a release determines whether a study will be made available to the public or restricted to those who receive permission to utilize its data. Any exceptions to

the workflow and levels can exist depending on specific archives or requests from PIs. The initial curation work, once a study has been assigned to a data curator to be worked on, is where the more technical usage of programs such as SAS takes place.

PROCESSING DEPOSITED DATA

The starting point of the data curation process is receiving the datasets and documentation for a study that are deposited by the PIs or their designate. Data can be deposited in various file types and formats, including SAS files in sas7bdat format. While it is a rare occurrence, sometimes data files are deposited in an stc format, which requires the following code to generate a different file type for further curation:

```
libname output "<filepath where output file will be located>";
filename input "<filepath where stc file is located>";
proc cimport library=output infile=input memtype=data;
run;
```

For studies that are deposited in SAS, codes that are used to run initial analyses contain the FREQ and CONTENTS procedures. The lines of code are as follows:

```
options nofmterr;
libname start "<filepath containing dataset>";
filename infile "<filepath specifying dataset>";

ods pdf file="<filepath name containing freq PDF output>";
anchor="dataset" uniform style=PEARL bookmarkgen=yes bookmarklist=show;
proc freq data = <data file name>;
tables _all_ / missprint;
run;
ods pdf close;

ods pdf file="<filepath name containing contents PDF output>";
anchor="dataset" uniform style=PEARL bookmarkgen=yes bookmarklist=show;
proc contents data = <data file name> varnum;
run;
ods pdf close;
```

These procedures help to list out all variables and their respective values and are commonly performed in order to review and detect potentially sensitive information. Sometimes, however, if a dataset has too many variables or observations, an error will generate stating that the FREQ procedure could not run due to insufficient memory. Solutions to overcoming this issue would include editing the code so that it would resemble the following:

```
ods pdf file="<filepath name containing contents PDF output>";
anchor="dataset" uniform style=PEARL bookmarkgen=yes bookmarklist=show;
proc freq data = <data file name> (drop=<UniqueID>);
tables _all_ / missprint;
run;
proc freq data = <data file name>;
tables <UniqueID> / missprint;
run;
ods pdf close;
```

In this circumstance, a variable with almost guaranteed individual instances of values, such as unique IDs, are dropped in the first line of the FREQ code while another which only specifies that variable is listed. This change is often used to work around the memory error.

On occasion, PIs will deposit SAS code files and request curators to utilize them for initial processing. In the most common situation, the code will contain syntax in order to create a format sas7bcat file that is used to incorporate value labels. Such code usually contains the following lines:

```
libname format "<filepath where sas7bcat file will be produced>";

proc format library = format;
    Value VARIABLE
          1 = "Value Label"
          2 = "Value Label"
          3 = "Value Label";
run;
```

The format file that is produced is then imported into other processing files, usually an SPSS syntax file, for further curation. Aside from these format code files, some PIs will deposit code that are to be released as secondary data analyses for users; usually with accompanying user guide documentation. This code may be used to produce summary variables or replicate published findings. These types of code are treated as supplemental files and released as they are provided to ICPSR, usually as part of a compressed zip package.

Once the initial SPSS syntax processing history file is put together for a study, a subsequently produced input dataset is processed via an internally-produced batch processing system called Hermes. The running of Hermes, along with other scripts in the data curation process, is performed via UNIX in a PuTTy interface. The Hermes script generates an entire statistical suite package which includes copies of the curated dataset in different versions of major statistical programs, including SAS. Compared to other statistical packages, SAS files are least likely to face issues in the conversion process such as dealing with datasets that contain variables with large format lengths and value labels incorporated into inappropriate variable types.

SELF-QC

Once the initial data and documentation has been produced and compiled via Hermes, several procedures that belong to the phase of "Self Quality Check", or Self-QC, are performed before the study is reviewed by other curators for eventual release. SAS procedures and codes that are utilized in Self-QC can vary depending on the size of datasets that are produced for a study.

For standard datasets that are not large in terms of file size (moderate number of observations or variables), a standard code consisting of elaborate macro commands is used to generate outputs. While the majority of the code is left unchanged when editing a template per study and dataset, the following lines which specify file paths and datasets are amended:

```
%let project = S#####-####;

%let work_folder = (<current file path>);

%let original_folder = (<file path containing deposited dataset>);
%let original_file = (<deposited dataset file>);

%let hermes_folder = (<file containing processed dataset>);
%let hermes_file = (<processed dataset>);
```

Aside from the macro commands, procedures that are utilized in the code include PRINTTO, CATALOG, IMPORT, FORMAT, DATASETS, SQL, REPORT, COMPARE, APPEND, DOCUMENT, TEMPLATE, REGISTRY, and COPY. The outputs generated via this code, which are in html format, include differences between the deposited and post-Hermes datasets such as the number of variables and the differences in variable types, lengths, and labels. It should be noted that only variables that have flagged

differences are displayed in the output; as variables with the exact same information between datasets are not included due to redundancy. This output allows the curator to review required changes to a dataset such as changing variable names and cases of labels.

In the situation where datasets are too big to run and thus cause the macro code to fail due to memory issues, a simpler code consisting of only the IMPORT and COMPARE procedures is utilized. This code consists of the following lines:

```
proc import out = work.prehermes datafile='<pre-hermes_dataset>' dbms=SAV
replace;
run;

proc import out = work.posthermes datafile='<post-hermes_dataset>' dbms=SAV
replace;
run;

proc compare base=prehermes compare=posthermes maxprint=(600,10000);
run;
```

The actual running of the code is performed not in SAS itself but via a SAS script in UNIX (sas <name of file>.sas), as attempting to run the code in SAS would cause the same memory issues that can be found with the more elaborate macro code. Running the script on this code produces lst and log file outputs. Being a simpler code containing only the IMPORT and COMPARE procedures, the information that is generated is simply the number of variables and observations between the deposited and post-Hermes datasets. If a discrepancy between the two occurs, it would usually mean that variables in the deposited dataset were either dropped by accident (which would require making pre-Hermes revisions) or at the request of either a PI or by a study's series standards which deviate from the establish Curation Level tasks.

If the outputs determine that there are no issues with the data and overall study, then the Self-QC stage can be considered complete, and the study is submitted to be reviewed by other curators. During these additional QC reviews, other curators will utilize the same SAS codes to review themselves and determine if there are no discrepancies in the output. Once all information and outputs are verified and a study is declared to have passed these QC reviews, it is eventually released by the curator and made available on the ICPSR website. For publicly available studies, users who select the option of downloading a study's dataset in SAS will receive the data in either a Cport Transport File (stc) format or a SAS setup file with the data in text format.

FUTURE IMPLICATIONS OF SAS

SAS software has presented itself as a useful program in ICPSR's data curation workflow; and it also has the potential for enhancing practices. While SAS is used for some tasks in the early stages of the curation process, other statistical packages such as SPSS are currently prioritized in the full curation process. Given that there are varying levels of curation work that can accompany studies, there is the possibility that pre-Hermes work such as those done on Level 1 studies can be performed by using SAS alone. For example, some Level 1 studies only need to have their file type converted to an SPSS format that can be read by the Hermes script. The issue with this task is that while SAS can convert a data file into file types other than SAS, only SPSS is currently acceptable to be read by the Hermes script. The solution in this circumstance would be for the script to be modified so that it will be able to read input files in other formats such as SAS.

Aside from the minimum required procedures for putting together a processing history file, other additions to the code can be made to further process a dataset in SAS. The creation of value labels for variables, as mentioned previously with the format files, is a starting point. Other methods of editing the datasets can include incorporation variable labels and reordering variables. SAS procedures that would accompany these additional tasks would include CONTENTS or SQL to reorder variables and LABEL

statements to create variable labels. All these tasks, usually done with Level 2 and 3 curation assignments, are performed to edit the datasets in order to have them match PI-deposited documentation (questionnaires, codebooks, etc.).

CONCLUSION

There is no doubt that SAS, in its current usage, is both important and essential with data curation tasks. Although it is not the main statistical program used with curation work, the nature of SAS such as its comprehensible syntax language can offer data curators an opportunity to learn the program. Once this notion becomes realized, it can give data curators a sense of liberty in that they have options for preference in whichever program to conduct their work. In addition, it can help curators expand on the program's potential in ways that have not been discussed before. As the Inter-university Consortium for Political and Social Research is a data archive, it is important to focus on the programs that put together the data themselves such as SAS.

ACKNOWLEDGMENTS

The author would like to thank the Curation Unit at ICPSR for their work into putting together the SAS codes mentioned in this paper and Dr. Amy Pienta for providing helpful feedback on earlier drafts of this paper.

RECOMMENDED READING

- Hawthorne, S. An Insider's Take on Data Curation: Context, Quality, and Efficiency. Journal of eScience Librarianship 2021;10(3): e1200. https://doi.org/10.7191/jeslib.2021.1200. Retrieved from https://escholarship.umassmed.edu/jeslib/vol10/iss3/1

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Piotr Krzystek
ICPSR at the University of Michigan
piotrk@umich.edu