A Common Platform for Data Dissemination

Dan A. Black, University of Chicago
Myron Gutmann, University of Colorado, Boulder

Prepared for discussion at the fourth in-person meeting of the National Academies' Standing Committee on the Future of NSF-Supported Social Science Surveys

Thoughts and observations included in the paper are solely those of the authors, and are not necessarily endorsed or verified as accurate by the National Academies of Sciences, Engineering, and Medicine.

When we were graduate students and assistant professors, using data sets involved learning the intricacies of reading tapes and job control language, but learning the content of a survey was an easy task. One sat down with the survey instrument and discovered what questions had been asked. After selecting the desired variables, the "real work" of wrestling with tapes and job control language began. As computational power skyrocketed and costs plummeted, we moved data sets off tapes and mainframe computers that filled huge rooms in university computer centers and put them on electronic drives on laptops.

The reduction in computing costs, however, also unleashed another transformation in the use of social science data: Computer Assisted Interviewing (CAI). The use of computers in the interview process greatly facilitated interviews. In the days of paper and pencil surveys, we relied on interviewers to be able to adjust the interview in real time to account for differences in the patterns of interviews that are based on respondents' answers to previous questions. These "skip patterns" were often quite confusing and resulted in unintended "skips" that degraded the quality of the data. The use of CAI made these unintentional skips much less common. In addition, because computers could easily handle complex skip patterns survey designers were now free to implement highly complex instrument designs. Indeed, data sets designed after the advent of CAI such as the National Longitudinal Survey of Youth, 1997 Cohort, regularly have interviews that produce more new variables than the number of respondents interviewed.

Of course, no respondent comes close to answering the full set of questions on the NLSY 1997. Indeed, if one were to print out the html version of the instrument the document is over 1,700 pages long. The day when one would quickly peruse a survey instrument to learn the content of a survey has vanished. The complexity of data sets, particularly panel data sets, increased dramatically with the expansion of CAI.

In many ways, social science surveys have not kept pace with these changes. Data dissemination and extraction tools created for many survey data sets often have the look and feel of 1990s technology from the era of data tapes that preceded data access via the World Wide Web. Most of these technologies were developed before the explosion of complexity that accompanied the increased use of CAI. These technologies failed to keep pace with the rapid development of database management systems that computer science developed and were heavily utilized in the business community where data sets were inherently more complex. They also failed to keep pace with approaches to data dissemination and preservation that have developed over the past twenty years.

This is a long-winded introduction to an important issue that the Standing Committee considered during its meetings. Committee members asked whether it was possible to improve the ways that the three NSF-sponsored surveys (ANES, GSS, and PSID) managed and disseminated their data, and whether it was possible to do a better job of ensuring the long-term preservation of the data. In doing so, members asked whether the solution to improved dissemination and preservation would be to turn those tasks over to specialist organizations, and whether it would be best to consolidate those tasks for the three surveys into a small number of separate awards. This document presents our views on the matter, informed by these discussions.

Fortunately, there are excellent models to inform us. For social science data, the development of the Integrated Public Use Microdata Series (IPUMS) program at the Minnesota Population Center developed with the support of the National Institutes of Health and the National Science Foundation represented an important advance for social science research. Steven Ruggles, who provided testimony to the Standing Committee, played a key role in the development of the IPUMS program and the program changed the way in which a younger generation of social science researchers expected their data to be delivered. These publicly available census data, based on the Census Bureau's public use microdata samples (PUMS), have been integrated to make the work of researchers easier, and shared via a web interface that is very efficient and produces versions of data for many popular statistical programs. It is literally possible to select data after lunch and be doing the analysis the same afternoon. This represents a dramatic cost savings for research projects, reducing the time spent by researchers and their assistants preparing data files.

In his testimony to the Standing Committee, Professor Ruggles reviewed the data dissemination approaches of the three NSF-funded surveys. The Michigan-based surveys (PSID and ANES) are considerably more difficult to use than, say, the data in the IPUMS system. This is not to fault the University of Michigan; the creation of modern data dissemination systems is quite expensive. Their antiquated systems disadvantage these data for two reasons. First, the use of the data requires considerably more resources and time for the creation of data sets for analysis. This will of course limit the use of the data. Second, new users who have been accustomed to the use of modern data extraction systems will find this old technology unfamiliar and painfully difficult to use. This may prevent the younger generations of scholars from using the data, which may ultimately threaten the viability of the surveys. Professor Ruggles was more complementary about NORC's new data extraction system for the General Social Survey (GSS) (with the caveat that Ruggles contributed to the development of the updated GSS system and that one of the two authors of this piece has a close relationship to NORC). The GSS, however, is a relatively simple survey when compared to the PSID, and whether a system similar to the NORC system could be adapted to other studies -- especially longitudinal data -- remains to be seen.

What we think is clear is that NSF needs to give considerable thought to the possibility of updating the data dissemination functions of the three surveys. It is difficult to sustain the existing system where the dissemination activities are controlled by the survey managers themselves, given their limited expertise. That's a likely transition that NSF would need to manage, with a variety of potential outcomes, including separate dissemination contractors or a single contract.

Data dissemination is only one missing infrastructure component; the other is serious attention to long-term preservation. Professor Ruggles mentioned the necessity for both the GSS and PSID to update and document their archival systems. We emphasize this point because we worry that much of the original sample identifying information from the early years of the ANES and GSS surveys may be lost because of the failure to adequately archive information. This could be a great loss to the scientific community. For instance, if one wished to examine the intergenerational transmission of voting patterns or attitudes in the early years of the ANES or the GSS, the availability of adequate documentation would be needed to generate a sample of respondents' children and even grandchildren that could represent an important contribution to science. Only imagination (and resources) limits the creative ways in which these data could potentially be used. Dealing with this inadequacy would require that NSF find a way to ensure preservation, which we believe would again involve consideration of various contracting arrangements, either individually or centrally.

Infrastructure investments such as these often lack the allure of supporting exciting, new research projects. To our mind, however, the investments have immense returns. The IPUMS project at University of Minnesota's Population Center has probably done more to promote the access and proper use of the Census PUMS than anything else. The result has transformed how social scientists use Census data, both in their teaching and their research. Investments in the "Big Three" represent investment in terms of today's dollars of hundreds of millions of public monies. Preserving this investment through the proper archival of the data and encouraging their use with a modern data extraction system is an important means of preserving this investment.