SYMPOSIUM

Data Preservation through Data Archives

Jeremy J. Albright, *Inter-university Consortium for Political and Social Research*
Jared A. Lyle, *Inter-university Consortium for Political and Social Research*

Science functions best within a liberal democracy. Every hypothesis test is an expression of doubt, as it carries with it the implication that a particular presumption may be incorrect (Kruschke 1998), whereas authoritarianism punishes challenges to prescribed beliefs. Consequently, science can lead to true innovation and improvements in knowledge only when laws and social norms permit dissent.

At the same time, scientific assertions themselves must be open to challenges and verification. To ensure that a surprising finding is not the result of an intentional misrepresentation of the data or even a simple coding error, it is imperative that researchers be able to replicate the analyses of others. Although this may be difficult in the context of qualitative research (how does one replicate in-depth interviews?), careful documentation of coding choices and programming syntax can make it quite easy to verify results from quantitative studies (Long 2008). Today, the Internet facilitates the posting of data and programming code for anybody to access, and an increasing number of social scientists argue in favor of making the electronic posting of replication materials a practice as common as fully listing citations (Freese 2007; King 1995).

Nonetheless, there are certain issues that may arise when an investigator chooses to make data available to the research community that merit further discussion in our discipline. First, data collection is a time- and resource-intensive process, and precautions must be taken to ensure that the individuals who have taken on a major project retain full rights to claim credit for their findings. If data are made public prematurely, the result can be a public-goods dilemma in which free riders attempt to publish first with data that others have paid to assemble. Second, protecting the identities of study participants in a publicly released data file is essential to maintaining the credibility of research organizations and ensuring that subjects are willing to take part in future investigations. A substantial literature has developed exploring threats to the privacy of research subjects, and data collectors in our discipline should take the relevant findings into considerations. Finally, standards have developed over the last decade for storing metadata (data about the data) with the goal of providing a common language for archiving digital information. Adherence to these standards facilitates the creation of documentation that can be understood by data users without having to contact the original PIs, thereby maximizing the utility of existing data now and into the future.

The present article seeks to outline these issues by proceeding as follows. First, the benefits of archiving quantitative data are summarized. A second section describes some pitfalls that must be avoided to protect the work of data collectors and study participants. A subsequent section outlines the archiving process at one data archive, ICPSR. A listing of the benefits of relying on an established archive for data preservation concludes.

WHY ARCHIVE?

Norms of methodological transparency encourage honesty in the reporting of research results. In a worst-case scenario, pressures for career advancement, tenure, or prestige may create perverse incentives to "publish or perish" that, if not countered with some form of accountability, can easily lead researchers to misstate conclusions. Yet erroneous inferences may not even necessarily result from nefarious intentions. Simple coding errors or a flawed syntax file can produce results that the investigator believes to be correct even when they are not. Making the data and programming decisions publicly available limits the extent to which bad findings influence future research.

There are, in fact, ample examples of errors in quantitative analysis leading to—at best—ambiguity in findings. One replication of a 1986 *American Sociological Review* article led to a debate over whether four different couples in the analyzed survey sample were really having sex 88 times a month, or if the 88s in the data file were actually meant to refer to missing observations (Jasso 1985, 1986; Kahn and Udry 1986). A broader study in economics by Dewald, Thursby, and Anderson (1986) sought to replicate a year's worth of articles in the *Journal of Money, Credit, and Banking*. The principal finding was that, in the vast majority of cases, it was entirely impossible to exactly replicate the published results even with the help of the articles' original authors. This led to the adoption of more stringent requirements in journals such as the *American Economic Review* requiring that data be made available at the time of publication.

Two developments in political science have led to an increase in the extent to which authors are making their data available to other researchers. First, the Internet provides a means for authors to distribute original data directly and immediately to the entire research community. In comparative politics, for example, various "expert surveys" of party ideology are available for easy download from the respective PIs' Web sites (Benoit and Laver 2006; Hooghe et al. 2008). This has in turn made it possible to use the indicators for novel analyses beyond what the original researchers envisioned for their projects, thereby contributing to the

doi:10.1017/S104909650999076X

cumulation of knowledge. Second, Gary King (1995, 2003) has been a persuasive proponent of transparency and replication in the discipline, and journals such as *Political Analysis* now maintain Web appendices with the data and/or programming syntax necessary to reproduce reported results available to the broader research community.

Although it is possible to set up a personal Web site to distribute original data beyond the variables analyzed in a journal article, there are very good reasons for turning to a well-established data archive instead. For one, Web sites are rarely permanent (Altman and King 2007; Seneca 2009), and it is unclear if the data will continue to be available when the researcher switches institutions, retires, or expires. Data archives ensure long-term availability of data through standardizing formats, migrating formats as technologies change, and guarding against general degradations in data. As the NSF Blue Ribbon Task Force on Sustainable Digital Preservation and Access (2008) recently reported,

> Without ongoing maintenance, digital assets will . . . fall into disrepair, succumbing to a host of "digital diseases" that impair or limit the ability to use them: bit rot (or degradation of the object so that it is no longer readable), technological obsolescence (which means that systems no longer exist that can read the encoding in which the data are represented and stored), or even outright loss. Consequently, preventive measures must be taken to insure that the media (tape, disk, and so on) are stable and the information encoded thereon can be read.

In addition, a data archive will be very familiar with the relevant proprietary and privacy issues that accompany publicly releasing a data file. Finally, data archives catalog and describe data collections using international standards, which enable efficient retrieval through a central source and permit reuse by researchers unfamiliar with the original collection. Increased access can then foster greater inter- and intra-disciplinary connections

ISSUES TO CONSIDER WHEN ARCHIVING

Embargoes

One policy that some (mostly non-political science) journals have adopted is to *require* that data be available for others to access when a manuscript is published. Likewise, many agencies and organizations that fund scientific research demand that newly collected data be deposited in an archive for future researchers to access. The intent is to make the comprehensive documentation of one's methodological decisions an integral part of the research process and additionally discourage sloppiness in scholarship. Unfortunately, although the immediate provision of data facilitates replication and accountability, exceptions may be necessary if the data are of a restricted or proprietary nature.

Collecting novel data can be a time- and resource-intensive endeavor, and scholars willing to take on a large data-collection project are understandably protective of their intellectual property. If others not involved in gathering the observations are able to access the data prematurely, they may then be able to claim credit for findings derived from the data before the original PIs have had an adequate chance to complete their own analyses (Fowler 1995). Requiring the public release of even subsets of proprietary data can lead to free-rider problems and ultimately discourage large-scale data collection efforts by individual researchers.

Nonetheless, as norms of transparency and replication continue to develop, the discipline will need to consider how it will simultaneously protect intellectual property and facilitate open access to research findings. If journals decide to require that all data analyzed are made available, they may follow the lead of the American Economic Review *and allow for editor-approved exceptions in cases of embargoed data. At the same time, depositing data with an archive—rather than making it available on a personal Web page—adds protection insofar as the archive can restrict access to data holdings for users or institutions that have violated clearly stated embargoes.*

Indeed, there is a recent precedent that shows that these concerns are justified. The National Institutes of Health maintain databases of raw data in order to promote transparency in NIH-funded research. Typically, new data deposited in an NIH database are protected by an embargo specifying that only the PIs have a right to publish using the data within a particular time frame. However, in 2009, a researcher from the University of Washington discovered a paper online at the *Proceedings of the National Academy of Sciences* based on data she had collected and that were still under embargo (Holden 2009). Although actions were taken to punish the unauthorized use of the data, the experience demonstrates that strong regulations will be required to ensure that one scholar's hard work is not denied credit when data are archived.

Fortunately, similar issues have not arisen prominently within political science. Nonetheless, as norms of transparency and replication continue to develop, the discipline will need to consider how it will simultaneously protect intellectual property and facilitate open access to research findings. If journals decide to require that all data analyzed are made available, they may follow the lead of the *American Economic Review*

https://doi.org/10.1017/S1049096510990768 Published online by Cambridge University Press

and allow for editor-approved exceptions in cases of embargoed data. At the same time, depositing data with an archive—rather than making it available on a personal Web page—adds protection insofar as the archive can restrict access to data holdings for users or institutions that have violated clearly stated embargoes.

Privacy

An additional problem with making data widely available is guaranteeing the privacy of a study's participants. This is an issue that has gained a good deal of attention in recent years, as increases in computing power and improvements in matching algorithms have decreased the costs required to connect information from different databases in a manner that can compromise sensitive information (Bethlehem, Keller, and Pannekoek 1990). As a consequence, the field of statistical disclosure control (SDC) has been developed to study threats to the identification of study respondents in publicly released data files (Willenborg and de Waal 2001).

The problem is that the release of even seemingly benign indicators may be potentially exploitable by a nefarious "intruder" attempting to ascertain the identity of an individual. A single variable in isolation might not reveal much information, but a combination of indicators can yield enough unique knowledge to match a row in a data file with a specific person (for example, there are very few female public servants from small Alaskan towns who are multi-millionaires). What is more, an intruder may plausibly merge a single data file considered safe for release with a separate database that, in combination with the first, yields sufficient information to risk disclosure (Paass 1988). Further problematic from an analytic perspective is the fact that variables used in variance estimation for complex surveys—weights, strata, PSUs—can even be exploited to identify the geographic location of a respondent (Mayda, Mohl, and Tambay 1996). Thus, many large-scale surveys (including the ANES) limit the amount of design information available in public releases, despite the fact that this information is essential for making correct inferences.

Several data-collection organizations, including ICPSR and the United States Census Bureau, frequently employ methods for masking potentially identifiable records (O'Rourke 2003; Zayatz 2006). The goal of this process is to alter data files prior to public release in a manner that retains as much of the original information as possible while ensuring that no single record can be identified. Commonly used masking approaches include data swapping (interchanging values on sensitive variables among cases), microaggregation (combining similar records and assigning the average within-group score to each case), adding random noise to observations, and top or bottom coding of variables with skewed distributions (e.g., income). Some recent attention has even been given to the release of entirely simulated data—based on methods akin to multiple imputation for missing data—that mirror the multivariate distribution of the original data file but contain none of the original observations (Raghunathan, Reiter, and Rubin 2003).

SDC can be seen as a constrained optimization problem in which the threat must be minimized subject to some maximum measure of information loss (Domingo-Ferrer and Torra 2001). The literature is highly technical, with the bulk of the relevant publications appearing in informatics and statistics journals. There is software that can carry out SDC analysis and make the optimal masking decisions for the user, including the free sdcMicro package for *R* (Templ 2008). However, data collectors may prefer to rely on the experience of an established data archive to ensure that the optimal amount of data utility is retained in the presence of masking.

ARCHIVING PROCESS AT ICPSR

Data archives have supported the social science community for over a half century by allowing researchers to replicate, verify, and extend original findings (Inter-university Consortium for Political and Social Research 2009). Examples of long-standing data archives include The Roper Center for Public Opinion Research, the Inter-university Consortium for Political and Social Research (ICPSR), the Henry A. Murray Research Archive, and the Odum Institute for Research in Social Science. Since we are directly familiar with ICPSR's archiving processes, in this section we will use ICPSR as an example of what a data archive can do to enrich, preserve, and disseminate data collections for the research community. Established in 1962, ICPSR maintains a data archive of more than 500,000 social science research files.

ICPSR engages three primary steps to archive data. First, staff members attempt to understand the data collection by working with the original owner or curator of the data. Next, the data collection is deposited at ICPSR, where data curators review, validate, and process the material. Finally, the archive distributes a permanent version of the data collection on its Web site or at its facilities, with preservation copies maintained at ICPSR and replicated offsite. (See figure 1.)

Archival work at ICPSR begins by teaming staff members with the researcher to ensure that ICPSR understands the data that the investigator wishes to deposit and to identify any constraints on future access to the data. While ICPSR believes that all data should be made available to the largest possible audience, sometimes there are issues of intellectual property or confidentiality protection that need to be understood even before the data can be acquired. This is usually done by a researcher or data producer contacting ICPSR staff and working with a curator who has experience with the kind of data involved.

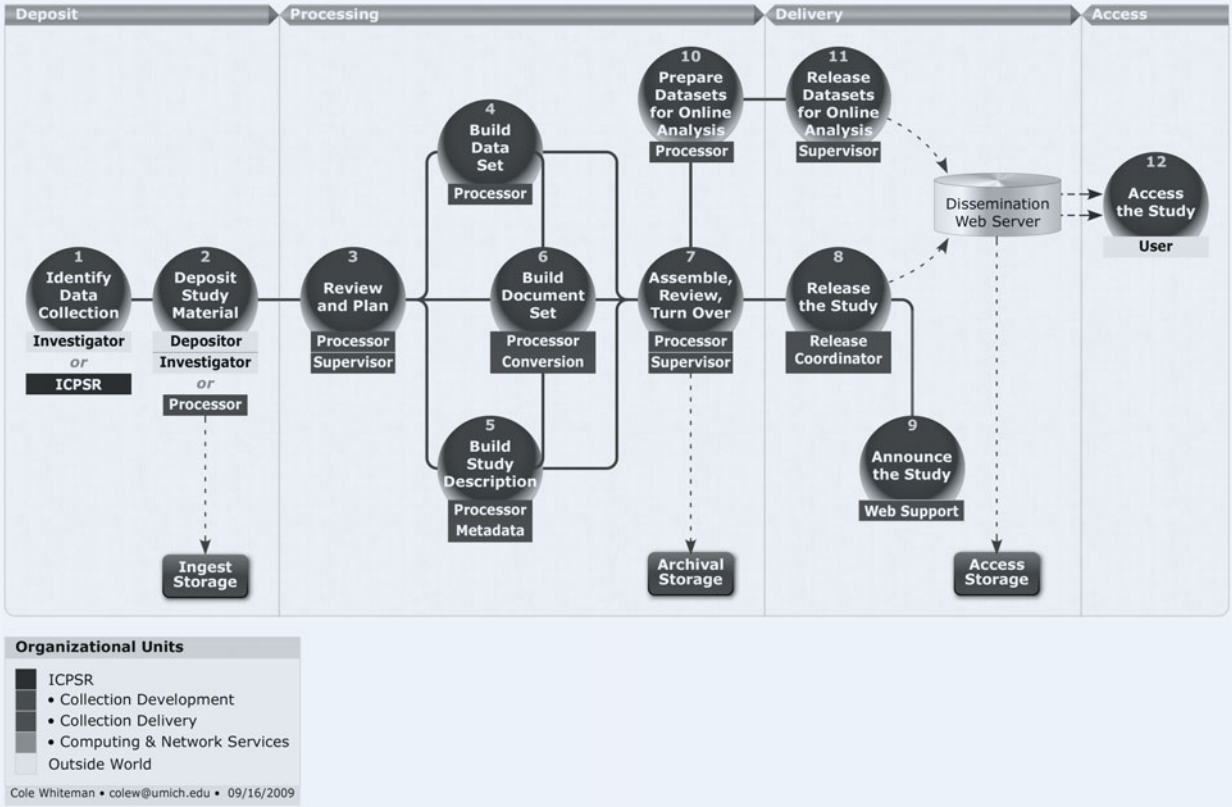
The next step in the process is the deposit itself. At ICPSR this can be done by the researcher using an online system, or it can be managed by the experienced curator already assigned to the project. The researcher or curator fills out an on-line form that builds the high-level metadata (documentation) about the study, uploads all the associated data and documentation files, and then signs an online agreement giving ICPSR permission to archive the data collection.

Once the deposit is finalized, a curator (called a data processor at ICPSR) verifies the data and builds the final documentation. This process starts with a confidentiality review, designed to make sure that there are no direct identifiers (e.g., name, social security number, telephone number, etc.) in the data file. It continues with a check to make sure that the data

https://doi.org/10.1017/S1049096510990768 Published online by Cambridge University Press

Figure 1

The ICPSR Pipeline Process: How ICPSR Acquires, Archives, and Disseminates a Typical Study



match the documentation. If inconsistencies, errors, or confidentiality risks are found, the data processor works with the original researcher to correct the issue. The process concludes with the construction of a final, permanent version of the data, alternative versions (such as in SAS, Stata, SPSS, or tab-delimited files), and final documentation. The final processing steps can also include the creation of versions for online analysis.

Concurrent with the data-processing work, the data curator builds documentation using Data Documentation Initiative (DDI) markup, an international standard for documenting social science research (Vardigan, Heus, and Thomas 2008). DDI is a non-proprietary XML-based approach to metadata that is appropriate for long-term preservation and enables researchers to understand and exchange data. The DDI metadata specification originated at ICPSR and is now the project of an alliance of about 25 institutions in North America and Europe.

Once the processing and metadata work is done, the data collection is carefully checked by a supervisor. The dissemination-ready data are then entered into the ICPSR catalog and made available on the Web. In some cases data cannot be distributed via the Web, so ICPSR has systems for limited distribution that require data users to sign a contract or come to ICPSR to use the data. After data are released, ICPSR provides support for downloading and using data.

All data are preserved and replicated in a variety of locations. Copies of ICPSR data for distribution are replicated in the "cloud," as well as at computer installations at the University of California, San Diego, and other locations. ICPSR maintains multiple copies in Ann Arbor as well. Over time and as needed, data are transformed so that formats and technologies are kept up-to-date, and file integrity is maintained. ICPSR has made it a priority to demonstrate compliance with prevailing standards and practices of the digital preservation community, such as the NASA-produced international standard "Reference Model for an Open Archival Information System (OAIS)" (Consultative Committee for Space Data Systems 2002).

CONCLUSION

To conclude, we note that there are several advantages to relying on a longstanding professional data archive, such as ICPSR, for making original data available to other potential users. These include:

1. A personal Web page on a university or private server will not be permanent, as jobs change, servers crash, and people retire, die, or forget.

https://doi.org/10.1017/S1049096510990768 Published online by Cambridge University Press

2. A data archive can enforce embargoes as well as punish the use of embargoed data by denying violators access to other data for a period of time (as NIH did in the example described earlier).
3. Data archives are familiar with the SDC literature and can work with PIs to balance the trade-off between data utility and the protection of study participants.
4. Data archives are familiar with international standards for creating and storing metadata, which greatly enhances the usability, interoperability, and exchange of data.
5. Archives often have specialized tools available to facilitate analyzing variables either with a user's preferred software package or directly online.
6. The reputation of a given archive provides legitimacy to a public release of a PI's data.
7. Archives can provide general user support and other data-maintenance services, freeing the researcher from routine and otherwise time-consuming tasks.
8. Data archives can provide a third-party system for ensuring the integrity and accessibility of research data for the future (see Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age 2009). ■

NOTES

We thank Cole Whitman for supplying the figure mapping the ICPSR Pipeline Process.

REFERENCES

Altman, Micah, and Gary King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data." *D-Lib Magazine* 13 (March/April). http://gking.harvard.edu/files/abs/cite-abs.shtml.

Benoit, Kenneth, and Michael Laver. 2006. *Party Policy in Modern Democracies.* London: Routledge.

Bethlehem, Jelke G., Wouter J. Keller, and Jeroen Pannekoek. 1990. "Disclosure Control of Microdata." *Journal of the American Statistical Association* 85 (409): 38–45.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2008. "Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation." http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf.

Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age.* National Academy of Sciences. Washington, D.C.: National Academies Press. http://books.nap.edu/catalog.php?record_id=12615&utm_medium=etmail&utm_source=National%20Academies%20Press&utm_campaign=NAP+mail+new+08.05.09&utm_content=Downloader&utm_term=.

Consultative Committee for Space Data Systems. 2002. "Reference Model for an Open Archival Information System (OAIS)." http://public.ccsds.org/publications/archive/650x0b1.pdf.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit, and Banking Project." *American Economic Review* 76: 587–603.

Domingo-Ferrer, Josep, and Vicenc Torra. 2001. "Disclosure Control Methods and Information Loss for Microdata." In *Confidentiality, Disclosure, and Data Access: Theory and Practical Application for Statistical Agencies*, ed. Pat Doyle, Julia J. Lane, Jules J. M. Theeuwes, and Laura M. Zayatz. New York: Elsevier Science Pub Co.

Fowler, Linda L. 1995. "Replication as Regulation." *PS: Political Science and Politics* 28 (3): 478–81.

Freese, Jeremy. 2007. "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods Research* 36 (2): 153–72.

Holden, Constance. 2009. "Paper Retracted Following Genome Data Breach." *Science* 325: 1486–487.

Hooghe, Liesbet, Ryan Bakker, Anna Brigevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen, and Milada Vachudova. 2008. "Reliability and Validity of Measuring Party Positions: The Chapel Hill Expert Surveys of 2002 and 2006." Unpublished Manuscript. Chapel Hill, NC.

Inter-university Consortium for Political and Social Research. 2009. *Guide to Social Science Data Preparation and Archiving: Best Practices Throughout the Data Life Cycle.* 4th ed.. Ann Arbor, MI. http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/dataprep.pdf.

Jasso, Guillermina. 1985. "Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences." *American Sociological Review* 50: 224–41.

———. 1986. "Is it Outlier Deletion or Is it Sample Truncation? Notes on Science and Sexuality." *American Sociological Review* 51: 738–42.

Kahn, Joan R., and J. Richard Udry. 1986. "Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions." *American Sociological Review* 51: 734–37.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28: 444–452.

———. 2003. "The Future of Replication." *International Studies Perspectives* 4: 100–05.

Kruschke, John. 1998. "Teaching Statistics as an Expression of Liberty." Speech at the 31st Annual Meeting of the Society for Mathematical Psychology, August 8.

Long, J. Scott. 2008. *The Workflow of Data Analysis Using Stata.* College Station, TX: Stata Press.

Mayda, Jackey E., Christopher Mohl, and Jean-Louis Tambay. 1996. "Variance Estimation and Confidentiality: They Are Related!" *Proceedings of the Survey Methods Section, Statistical Society of Canada* 135–141. http://www.ssc.ca/survey/documents/SSC1996_J_Mayda.pdf.

O'Rourke, JoAnne McFarland. 2003. "Disclosure Analysis at ICPSR." *ICPSR Bulletin* (Fall): 3–10.

Paass, Gerhard. 1988. "Disclosure Risk and Disclosure Avoidance for Microdata." *Journal of Business and Economic Statistics* 6 (4): 487–500.

Park, Inho. 2008. "PSU Masking and Variance Estimation in Complex Surveys." *Survey Methodology* 34 (2): 183–94.

Raghunathan, Trivellore E., Jerry P. Reiter, and Donald B. Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19 (1): 1–16.

Seneca, Tracy. 2009. "The Web-at-Risk at Three: Overview of an NDIIPP Web Archiving Initiative." *Library Trends* 57 (3): 427–41.

Templ, Matthias. 2008. "Statistical Disclosure Control for Microdata Using the R-Package sdcMicro." *Transactions on Data Privacy* 1: 67–85.

Vardigan, Mary, Pascal Heus, and Wendy Thomas. 2008. "Data Documentation Initiative: Toward a Standard for the Social Sciences." *The International Journal of Digital Curation* 1 (3): 107–13. http://www.ijdc.net/index.php/ijdc/article/view/66/66.

Willenborg, Leon, and Ton de Waal. 2001. *Elements of Statistical Disclosure Control.* New York: Springer-Verlag, Inc.

Zayatz, Laura. 2006. "Disclosure Avoidance Practices and Research at the US Census Bureau: An Update." *US Census Bureau Research Report Series.* http://www.census.gov/srd/papers/pdf/rrs2005-06.pdf.