

From Preserving the Past to Preserving the Future: The
Data-PASS Project and the challenges of preserving
digital social science data

Myron P. Gutmann, University of Michigan

Mark Abrahamson, University of Connecticut

Margaret O. Adams, National Archives and Records Administration

Micah Altman, Harvard University

Caroline Arms, Library of Congress

Kenneth Bollen, University of North Carolina at Chapel Hill

Michael Carlson, National Archives and Records Administration

Jonathan Crabtree, University of North Carolina at Chapel Hill

Darrell Donakowski, University of Michigan

Gary King, Harvard University

Jared Lyle, University of Michigan

Marc Maynard, University of Connecticut

Amy Pienta, University of Michigan

Richard Rockwell, University of Connecticut

Lois Timms-Ferrara, University of Connecticut

Copeland Young, Harvard University

Paper prepared for *Library Trends* Special Issue about the NDIIPP
Program

July, 2008

Social science data pose one of the unusual challenges of digital preservation.¹ From one point of view, they represent an unqualified success: almost from the beginning of the modern era of computerized research in the 1960s, sustainable organizations have ensured the preservation of critical data used by social scientists. In fact, an earlier issue of *Library Trends* (Heim, 1982), "Data Libraries for the Social Sciences", was entirely devoted to a discussion of development and challenges related to these efforts. Nonetheless, at the beginning of the 21st century there are still significant gaps in the holdings of major data archives, despite their obvious value for conducting new research and replicating earlier research (Freese, 2007; King, 1995, 2007). Responding to the NDIIPP request for proposals, the Data-PASS Project (Data Preservation Alliance for the Social Sciences) drew on a view shared by the major U.S. data archives: they needed to cooperate fully to ensure that at-risk social science data were identified, acquired, and preserved, and that they needed to establish a future-oriented organization that would continue to collaborate on those tasks.

The Data-PASS partnership consists of four academically-based social science data archives and the Electronic and Special Media Records Services Division of the U.S. National Archives and Records Administration (NARA), supported by a

strong technical infrastructure. The academic partners are the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, the Roper Center for Public Opinion Research (Roper) at the University of Connecticut, the Howard W. Odum Institute for Research in Social Science (Odum) at the University of North Carolina - Chapel Hill, and the Henry A. Murray Research Archive at the Institute for Quantitative Social Science (IQSS) at Harvard University. Harvard has also provided technical leadership for the project, in close collaboration with the other partners.²

The Data-PASS project began as an effort to identify digital materials -- some of them now very old in digital preservation terms -- that had never been systematically archived, and to appraise them and acquire the most valuable. While we knew about some of the content that the project planned to acquire from the beginning of the project, much of it has required research to discover what exists, followed by serious negotiations to acquire the data. As the project has progressed, however, it has increasingly turned its attention from identifying and acquiring legacy and at-risk social science data to identifying ongoing and future research projects that will produce data, and devising mechanisms for ensuring that those data are preserved and shared at an appropriate moment. This article is about the project's context and history, with an

emphasis on the issues that underlay the transition from looking backward to looking forward.

Social science data may be the oldest digital resource in the world. Starting with the 1890 U.S. Census, data used for social, economic, and political research were the first materials to be converted to digital format for analysis by computer technology (Anderson, 1988). The first tabulating machines were invented for conducting the 1890 Census; the IBM Corporation grew out of this invention.

Nearly fifty years after the 1890 census, in the 1930s, a remarkable group of social scientists invented a reliable, scientifically validated means for ascertaining public opinion that has enabled citizens to have a voice about the nation's affairs (Converse, 1987; Igo, 2006). For three-quarters of a century, public opinion polls, social surveys, and other kinds of structured interviews have tracked people's values, attitudes, knowledge, and behavior - measuring and recording the cultural and social heritage of the *whole* people as can no other "machinery." Surveys have done more than predict the outcomes of elections or tell us when presidents lose popularity. They inform us about aging, health and health care, race relations, women's rights, employment, and family life - virtually the full story of the social and cultural tapestry that makes up our

nation. Further, they provide the data necessary for sound, empirically based policymaking.

Surveys have also helped us to understand national tragedies, from the attack on Pearl Harbor to the present day. In the immediate aftermath of the attacks on America of September 11, 2001, pollsters at the National Opinion Research Center (NORC) of the University of Chicago were trying to understand and cope with what had just happened. They had intellectual resources that were available to no one else, because NORC had earlier studied the country's response to yet another national tragedy. From November 26 through December 3, 1963, NORC researchers had assessed the nation's reactions to the assassination of John F. Kennedy. They had observed deep physiological and psychological pain, the diverse ways in which the nation mourned, and the shame with which many citizens assessed that event.

In 2001, the researchers again found pain and anger, but also resilience, renewed national pride, and confidence in institutions. Their difficult path to making this comparison of how the nation responded to two very different tragedies helped us justify the Data-PASS project. By September 13, 2001, NORC researchers were in the field asking many of the questions that they had asked in 1963. It had proved easy to find those questions: they existed in "hard copy" in filing cabinets. What

proved hard to find and then to use were the digital data from that 1963 study. NORC first inquired whether it had followed its now-standard practice of depositing the dataset at ICPSR or Roper. It had not done so. The next step was to turn to NORC's internal digital archives. The tapes from the assassination study existed and were usable, but they contained only a subset of the data. Where was the complete data set? Without it, the researchers could not make the desired comparisons. NORC's archival staff was confident that the data still existed on punched cards. NORC's onsite card collection did not have the cards, but there was a collection of boxes of punched cards at a private storage facility. The boxes were retrieved, though not without some difficulty. Then it proved necessary to find a way to read the cards and handle an obsolete data format. The data were retrieved; preservation had worked -- but only by a thread (Smith & Forstrom, 2001).

This kind of story could be told by almost every archivist. The nation's digital heritage is fragile and growing more so. The problem of its preservation involves far more than technical remedies. It involves capturing knowledge that resides in aging minds, collating printed information that is now scattered throughout organizations, retrieving and reading digital media of many kinds written in many different formats, and dealing with obsolescent equipment and software.

The Landscape of Social Science Research Data

We begin our detailed discussion by talking about social science research data, who collects and owns them, and some of the challenges posed by that landscape. Two of the many kinds of digital content used for social science research have been of particular interest to the Data-PASS project. Both represent information given by an individual, either to a human interviewer or self-reported on a survey form (more recently electronically), or observed and interpreted by an outsider.

A large proportion of the information used by social scientists comes from responses to questions that are closed ended, either asking for specific factual information ("How old are you?") or eliciting categorized responses to questions of opinion ("Which of these candidates do you think you will vote for in the coming presidential election?").³ The common characteristic of these surveys is that their information can be readily converted to coded values and made digital. Some of these data, including the census, have their origins in administrative records that require universal coverage. Others have since the 1930s made use of sampling techniques that allow relatively small sample surveys to represent a larger population, whether local, regional, or national.

Not all research questions are suited to categorical responses. Often the exact words the survey respondent used to

discuss her life, attitudes, or experiences are important. These qualitative survey responses require that the full text of the interview be made available for study and analysis, something for which analytic software is now available. The Data-PASS partnership has also collected video and audio records of qualitative responses, where such multimedia records can serve to deepen our understanding of the responses and respondents. While many of these qualitative studies are the product of randomized samples of respondents, other researchers may be more directed in their interviewing strategy, choosing key actors or selecting representative lives to chronicle. Finally, social science data increasingly goes beyond surveys, and the Data-PASS partnership is actively engaged in collecting, disseminating and preserving data from social science experiments, administrative records, observational data, process-produced data, simulations, and other forms of research-related digital content.

These materials have diverse origins. The largest producer of social science content in the United States is the Federal government. Many Federal surveys are conducted by government agencies (especially the Bureau of the Census), while other Federal activities are conducted under contract by others. At the opposite end of the spectrum from Federal activities are those of private survey and marketing firms, which have developed since the 1930s to gauge public opinion and to

evaluate the demand for products, and research organizations that grew up after World War II to meet a demand for research (examples are the RAND Corporation, spun off from Douglas Aircraft, and the Survey Research Center at Michigan, which grew out of USDA survey efforts (both in 1948), and RTI International, set up in 1958). In the middle of the scale are academic researchers, who design research projects supported with local funds, corporate or foundation support, or government grants. Some researchers manage their own data collection, but many large-scale projects are done under contract by local or national-level survey organizations, including NORC, the Survey Research Center at Michigan, RAND, RTI, and Westat.

One of the main reasons to begin the Data-PASS project was our knowledge that less than half of the digital social science research content created since the revolution in sample surveys that took place in the 1930s has been preserved at a professionally-managed archival institution (Pienta, Gutmann, Hoelter, Lyle, & Donakowski, 2008). There are a variety of very good reasons for this lack of attention to preservation. Some individual researchers have been reluctant to deposit their data in archives, either because they wanted to avoid sharing it with potential competitors, they lacked the time or expertise to prepare the metadata required for effective sharing, or they did not recognize its long term value. Institutional data producers

may have been under contractual obligations with those who paid for data collection to protect proprietary information. And as we saw with the story of NORC's survey on the Kennedy Assassination, some data just fell through the cracks.

Part of the challenge of social science data preservation has come from the question of who owns these data. Private businesses and university-based researchers have assumed until recently that the data they generated were their property and that they had limited obligations to share their data with others, or to ensure its preservation. This is different from the situation for Federal data producers, who are bound by law to preserve records that NARA determines to be of continuing value. This requirement extends to some social science data produced by Federal contractors, but it does not include data from Federal grantees. In most instances, those data are the property of the University or Private Research Organization that received the government grant that supported the data collection activity.

In very recent times, major Federal supporters of social science research, including the National Institutes of Health and the National Science Foundation, have announced policies that encourage and in some cases require grantees to share their data.⁴ These efforts are the consequence, in part, of many years of advocacy from within the social science community for

policies that require data sharing (National Research Council, 1985, 2003, 2004, 2005). The impact of these requirements has yet to be seen, but they are not yet strong enough to ensure the preservation of digital social science content. The NIH rules, for example, only apply to their largest grants, those with direct costs that exceed \$500,000 in a single year. Most research costs less than that, and therefore does not fall under the obligatory data sharing requirement. In fact, the data sharing rules may create more problems than they solve, because they can lead to a proliferation of Web sites for self-dissemination by researchers, with ineffective long-term preservation.

Thus, at the time our project began there remained a huge quantity of digital social science research content that had not been archived and would not have been without aggressive activities of the sort that we have taken. The good news is that these materials left trails that we can follow, in the form of press releases, public grant announcements, and publications describing research. Those are the paths that we have followed.

Identifying and Acquiring Digital Content

The main task of the Data-PASS project has been the identification, appraisal, acquisition, processing, and preservation of important social science research data. As in

many projects, the project team understood how to do some of this when we started, and we have learned a great deal more as the work progressed. What we understood from the very beginning was that a few major categories of data were most important to us. These categories overlap, and they ignore some important data, but they include most of what interested us:

- Surveys and administrative data collected by or for the U.S. Government.
- Public opinion polls conducted by well-established institutions at the state, national, or international level.
- Research data collections supported by the most important national agencies that fund social science in the United States, specifically the National Institutes of Health and the National Science Foundation. Most of this research has been done by or on behalf of colleges or universities by their faculties and others.
- Research data collected by a group of non-governmental (either non-profit or for-profit) organizations that we call "Private Research Organizations." These organizations generally perform this work under contract for governments, universities, private businesses, and other non-governmental organizations.

- Research data created in the process of preparing academic publications, some of it original and some of it derived from other data, either publicly available or not.

From the very beginning, the Data-PASS partnership constituted itself as a collaborative enterprise that divided up the tasks associated with locating these data and ensuring their evaluation and preservation. In broad terms, this is how the project divided up the work to be done:

Harvard IQSS

Harvard University's IQSS and Murray Research Archive have developed rapidly during the Data-PASS project. The original goal for this project was for the Murray Research Archive to collect major university-housed longitudinal studies, but they have succeeded in doing much more. Early in the history of the project the Murray became part of the IQSS, and transformed its collection development policy, leading to a four-fold increase in the annual rate of acquisitions.⁵ Among the most important areas of research where the IQSS has made acquisitions are longitudinal studies of personality development, plus important data about economics, food policy, global inequality, the correlates of war, and systematic replication of data collections associated with research journals.

Harvard IQSS, through its technology group, the Harvard-MIT Data Center, also provided technical leadership and cyberinfrastructure including the shared catalog of the partners' entire holdings, metadata interoperability standards, and schemas for archival replication. This used and built upon research, software, and standards developed at IQSS. This includes the Virtual Data Center (VDC), which was used to create the shared catalog during the first phase of the Data-PASS project; the Dataverse Network which superseded the VDC; universal numeric fingerprint (UNF) algorithms, which the Harvard-MIT Data Center uses for data verification; a uniform standard for citations to data supported in the shared catalog (Altman, in press-a; Altman et al., 2001; Altman & King, 2007; King, 2007).

ICPSR

ICPSR has focused its collection development activities around the area of federally funded research data, mostly collected by university-based researchers. The goal has been to identify important investigator initiated social science data collections that are at risk of being lost. To accomplish this task, ICPSR created a database describing all of the historic and recent grant awards made by NIH and NSF. From this database ICPSR selected records for projects whose titles and abstracts

suggested that the investigator would be collecting original social science research data. Based on these 11,265 awards, ICPSR identified and contacted researchers, in the process sending email to 6,565 investigators and reaching about half of them. More than 1,800 investigators acknowledged that they had produced data. This activity has produced important results, most notably that only about 20% of research data have been previously archived, and that only about half of the research where data had not been archived was still available for long-term preservation (Pienta et al., 2008). Most importantly for the Data-PASS project, this effort has produced significant data acquisitions that would not have arisen otherwise, but those acquisitions have required considerable effort because of the need to work with a diverse group of data producers, many of whom having data that they have not worked with in a number of years.

ICPSR currently is working to acquire over 400 of these NIH and NSF sponsored data collections. These data are drawn from a wide variety of disciplines and research topics ranging from numerous empirical observations of American family life and family dynamics across the latter part of the 20th century (see Mortimer, Finch, Shanahan, and Ryu (1992) for an example) to one of a kind explorations of the impact of historical events and natural disasters (e.g. Hurricane Andrew) on human lives (see

Lanza-Kaduce, Dunham, Akers, and Cronwell (1998) for an example). More than 50 data collections have been committed for deposit at ICPSR and ICPSR staff continue to work with researchers and data producers to ensure that as many as possible of the overall number will be preserved and shared.

Equally impressive as the wide range of disciplines and topics covered by the NIH and NSF awards are the diversity of data formats and storage media originally used to collect and store the data. While approximately 30 percent of the NIH and NSF data collections that ICPSR contacted were stored in SPSS and Excel formats, others were captured in less common formats, such as StatMost⁶. Likewise, although conventional hard disk drives appear to be the storage media of choice for social science data collections (35 percent of data respondents), other less user-friendly storage media also lurk: examples of those ICPSR staff discovered are McBee cards and DAT tape (Pienta et al., 2007). ICPSR recently started processing acquired NIH and NSF data on punched cards and magnetic tapes, and has worked to restore data saved in early versions of SPSS and EBCDIC formats. These examples of diverse, proprietary, and obsolete formats and media highlight the challenges and opportunities ICPSR faces, as well as emphasize the need to archive data early in the research lifecycle, when both equipment and staff know-how are readily available to convert data into preservation formats.

Odum

The Odum Institute has unique strengths in its archive of the Louis Harris Polls, state polls data, and studies that document southern life or collected by southern researchers. Odum's collection development for Data-PASS emphasized expanding their collection of the Harris Polls, collecting state-level polls from the National Network of State Polls (NNSP), and Private Research Organizations. The Odum Institute archive has been the home of the Louis Harris Data Center since the early 1960's, but their ability to provide a complete series diminished after a change of ownership at Harris in the late 1990s. During the first year of the Data-PASS project Odum reestablished contact with Harris Interactive, leading to an up-to-date series that is now available.

Odum is also the archival home of the National Network of State Polls (NNSP), a confederation of organizations that conduct state-level public opinion surveys. The network consists of more than 50 members from 38 states, but turnover is common and coordination can be difficult. Odum's second focus was to revive the NNSP and begin to fill gaps in its collection and ensure that future surveys and polls would be archived. These efforts have yielded dozens of successful acquisitions from new and existing members. One of the challenges of this activity is the small size of many NNSP members, and the difficulty this

causes for them to devote consistent effort to archiving their data. Future efforts to improve this process are described later in this article.

Odum's third focus was private research organizations (PROs), using an existing relationship with RTI International as a starting point. This activity has been a major challenge for the project. At the outset, Odum staff intended to approach PROs as they do other data producing institutions, but this method did not prove to be successful. PROs move quickly from project to project, and often have little or no incentive to return to a project to archive the data after it is completed. Moreover, the nature of the contracts that the PROs have with their sponsors makes it unclear whether the PRO may actually archive the data. To address these challenges Odum revised its approach and began targeting the organizations that fund the data collection activities of PROs. This too is an area of future activity that is described later in this article.

Finally, the Odum Institute worked closely with the IQSS at Harvard in installing and testing the common data archive catalog first in the Virtual Data Center and then in the DataVerse system. By working together they were able to debug problems and identify areas to add functionality in early versions of these systems. One of the key additions to the

system containing the new common catalog was the ability to search at the question level.

Roper

The Roper Center has pursued three major collection development tasks: locating and acquiring public opinion polls that are identified in Roper's iPOLL database of polls but not archived; working with NARA to complete a collection of international surveys conducted by the United States Information Agency (USIA) and partially archived at NARA, partially at Roper, and partially at locations to be discovered; and acquiring data from NORC at the University of Chicago, a PRO.

The USIA data includes more than 2,000 surveys about U.S. foreign and defense policy conducted in dozens of countries by the USIA Office of Research from 1952 through 1999. Substantial, but partial, collections of these survey files have been at NARA, Roper, the State Department (currently housing the former USIA Office of Research) and assorted academic research centers and libraries. The most comprehensive of these collections is preserved at NARA, with data spanning the mid-1950s to 1999. The Roper Center had been the initial repository for the USIA collections from 1952 to the early 1970s and again archived data from the 1990s. The overlap at the two institutions of the data from the first four decades was minimal, while both held

substantial collections of data collected during the 1990s. The goal of the Data-PASS project was to assure that both institutions preserve a complete and accessible collection, a task that is now nearly complete and will be finished by the time the funded Data-PASS project is finished in 2009.

Another component of the Roper Center's work on this project involves a subset of studies (called TRACES) identified from the Center's iPOLL database of surveys conducted in the US. These are opinion surveys conducted by private research organizations that have been identified and evaluated as having value to social science research. For these studies, Roper had previously collected only tabular topline results, and not datasets of individual responses. To accomplish this, the Center made contact and sought out data from various sources. Two of the largest collections of this type chosen for inclusion in the Data-PASS project are those of the Public Agenda Foundation and the American Association of Retired Persons (AARP). Both organizations have sporadically deposited their data with the Roper Center in the past, but never their complete series. Roper now has a much more complete collection of their surveys.

The National Opinion Research Center (NORC) was established in 1941 at the University of Denver and now operates within the academic setting of the University of Chicago. During the 1950s and 1960s in particular, NORC conducted dozens of pioneering

surveys that have become classic studies of health care, social stratification, and education, among many topics (and including the JFK assassination study mentioned earlier). Many of these studies have become important parts of the social science literature, but the datasets were never archived. As has been the case for the Odum Institute and RTI, working with a PRO presents challenges related to the availability of PRO staff and the question of who owns the data. In the case of NORC, the Roper Center identified an initial group of thirty legacy studies that seem promising, and the two organizations are now at work on getting permission from the original sponsor to transfer custody of the data, to NORC, and then for the Roper Center to acquire them.

NARA

NARA is the last of the partners and because it is a federal agency it received no direct project funding. As the archives for the nation's federal records, its statutory responsibilities for identifying and appraising all federal records, including federal data records, and then for accessioning, preserving, and providing access to federal records appraised as having long-term value, complements the work of its academic DataPASS partners. Early in the Data-PASS project the appraisal expertise of NARA archivists contributed

to the development of the Data-PASS appraisal policy. NARA's legal mandates made possible its previously described collaboration with the Roper Center and the Department of State to assure identification and preservation of as many files as are extant from USIA's almost half-century-long international public opinion survey program. Going forward the same legal authorities will support all NARA efforts related to identifying, accessioning, and providing access to federal data of value for social science and other research.

A major accomplishment of Data-PASS has been the development of an open and interoperable shared catalog of digital data sources. Because NARA is a Data-PASS partner, series descriptions for all of its accessioned data holdings are periodically harvested for the catalog. As a result the federal data preserved by NARA are now more widely known to the social science research community.

Achieving these goals required a concerted effort on the part of all five partners, sustained and made successful by the distributed project structure, where each institution had its own list of possible sources for data and its own mechanisms for locating them. This work was possible because of the creation of an Operations Committee that has met by telephone every two

weeks since the beginning of the project and made important day-to-day decisions, with the full participation of staff from the Library of Congress. These decisions were facilitated by a set of formally-accepted guidelines, practices, and procedures (Altman et al., in press). These begin with the governing Articles of Collaboration and for the academic partners, a shared Deposit Agreement to be completed by data depositors. Others are formal standards for content selection, appraisal, acquisition, metadata, confidentiality protection, data security, and guidance for handling fragile materials.⁷ In practice, each partner has its own procedures for identifying digital content that might become part of the project. Once content is identified, the Operations Committee reviews the collection development decisions of each of the academic partners Committee, and if the content originated with a federal agency, also makes an effort to determine whether these data are already preserved at NARA, or scheduled to be transferred to NARA. Then, the acquisition and preservation of approved digital content occurs at the archive that identified the material, or less frequently, by one of the other partners.

Lessons Learned

Much of the success of our collections activity was a result of the development of a shared database of data that we

would consider for acquisition, as well as a streamlined set of agreements, procedures and guidelines. Furthermore, Data-PASS has reinforced the importance of jointly developed best practices, an open and interoperable catalog, and comparable work flows for content ingest. These gains operate throughout our project, and have provided benefits to all of the partners and allowed us to acquire and preserve far more data than would have been possible otherwise. In this section we want to emphasize several of the lessons that we have learned about the tasks needed to build collections of social science data, and the implications for the future.

The first and most important lesson we have learned was the importance of building long-term preservation and access policies into institutional processes. Data preservation is something that researchers have to initiate when they begin their research, not later, and certainly not at the end of the process or years afterward (Green & Gutmann, 2007; Higgins, 2007; Humphrey, 2006; Inter-university Consortium for Political and Social Research, 2005). A lot of the work of the Data-PASS partners took place only after data collection projects were long forgotten. By seeking to identify and preserve data long after the data were collected, we confirmed something that we knew already: it is difficult to overcome the barriers that stand in the way of effective digital preservation, especially

the researcher's lack of time, money, and knowledge. The alternative, by which metadata development, adherence to preservation standards, and planning for long term use are done early and well, requires that researchers start at the beginning of their project, and that they do so with support from their home institutions and from data archives.

The need for forethought and effective institutional support overcomes another lesson, the false sense of security that internet technology provides to researchers, universities, and private firms when they put their data on the web. Web sites are in constant flux and pages are refreshed daily, resulting in content that is replaced when it is changed, and which is eventually discarded as being out of date. Polling firms, for example, place data on a web site while the topic is fresh, but replace those data when they become dated. University-based researchers put data on a departmental web site while a project is active, but the community loses access to those data when funding runs out, a crucial student graduates, or the researcher simply loses interest. In these cases, there is rarely a longer term plan to maintain those data in up-to-date usable formats in perpetuity. ICPSR's research into NIH and NSF-funded research produced many examples of this, and each of the research firms contacted by Roper admitted having firsthand experience with such liabilities. Everyone agrees that it is necessary to take

steps earlier in the lifecycle of a dataset, ideally by securing a commitment from the principle investigator to deposit the data in an archive before the actual data collection takes place,. Now the challenge -- as we show later -- is to do this early enough to make it really happen.

Another lesson we learned was that dealing with the Private Research Organizations posed special problems if we wanted to meet our project objectives. All the PROs agreed in principle that there were benefits to preserving data, but in practice they struggled to make it a priority. In many instances, once the PROs delivered reports to their sponsors, the data disappeared from external view. After that, they asked, how could preserving the data bring them future revenue? One reason for the challenges we faced was a shortage of staff time. Staff at most PROs operate in an environment of "billable time," where every activity must be accounted for and billed to a sponsor. This made it difficult to schedule meetings with key personnel, and even more difficult to arrange the use of staff for the content search and rescue effort. In the case of NORC, the project also had to bear the cost of fees to retrieve data from a third-party warehouse.

Our relationship with the PROs required that the data archives build and maintain effective relationships with them. Sustaining these relationships is critical because only by

developing strong ties with the leadership and professional staff of the PROs is it possible to work through the variety of challenges that arise in the process of ensuring that digital content is preserved. One way that we have seen this operate is in dealing with the question of who owns the data. Most PROs are contractors: they create or manage data for some organization. They therefore have to review their contractual obligations before they can turn over the data to a third party (in this case, a data archive), which requires effort and motivation on their part. In every case the PRO needed to contact their original sponsor together with the data archive, identify the collection to be preserved, and request permission to turn it over. Only with strong and continuing relationships is this possible, and every time there are changes in the leadership or staff of the PROs, it becomes necessary to rebuild relationships. Many of these experiences are analogous to the challenges NARA faces in its records management program, given frequent changes in the leadership and staffs at federal agencies.

In concluding this summary of lessons learned, it is valuable to emphasize how different this project is from others in the realm of digital preservation, especially in terms of identifying and acquiring content. Many -- if not most -- digital preservation projects begin with a discrete and known

body of content that needs to be preserved. It might be the digital holdings of a library or archive, or it might be a kind of web site that needs to be harvested before it disappears. Our project had as its goal the identification of disparate content in many locations, owned by many individuals and organizations, not all of whom could be identified or convinced to turn over their content, and not all of whom could produce the content that they thought they wanted to turn over for preservation. What that meant for the Data-PASS project was that we needed to work together, and that in the future we will need to find ways to work with the community through an early start, strong institutional support, and relationships with individual researchers, research organizations, and research sponsors to ensure that the important digital content that documents American society is preserved. The impact of collaboration among the nation's social science data archives in meeting this challenge cannot be overstated.

At the same time, we have learned that an early start, relationships, collaboration, and persistence will probably not allow us to preserve all the valuable digital social science content, no matter how hard we try. Many researchers and research organizations lack the commitment, incentives, and resources needed to cooperate with the data archives community to ensure preservation, and no matter how hard the preservation

community works, we have learned that this is unlikely to change. Ensuring preservation of important digital social science content is going to require changes in policy at the institutional and governmental level, so that research organizations, universities, and government agencies go beyond the tepid -- and weakly enforced -- current policies that require preservation and sharing of research data, to something that will have real impact.

Social Science Data Preservation in the Future

Building upon these lessons learned, Data-PASS has begun to work toward archiving the products of recent and ongoing data projects. In recognition of these burgeoning "living collections", the Data-PASS partnership's Operations Committee has standardized and published appraisal and acquisition guidelines for active content (Data Preservation Alliance for the Social Sciences, 2007). Identifying and appraising newer data collections pose interesting questions of divining the future: How to appraise data resulting from dissertation awards, where the researchers are not yet professionally established and publications stemming from the data are still forthcoming? How to identify promising but yet-to-be-collected data that are known to a just a very small cadre of researchers? Harvard IQSS, for example, is approaching "living collection" appraisal

by casting a targeted net of published research data through collaborating with scholarly journals and departments to automatically archive the data associated with research articles and dissertations. ICPSR, for its part, is expanding its database of NSF and NIH awards to accommodate ongoing identification and selection of awards announced by federal sponsoring agencies. And, NARA is assisting federal agencies to identify and schedule by the end of FY 2009 (i.e., to propose disposition) all their electronic records, including data files, as mandated by the E-Government Act of 2002.

Among the most important ways that the Data-PASS partners are looking to the future is by taking steps earlier in the data lifecycle to encourage the deposit of data in a data archives, and then following up on those steps as research projects evolve. We briefly describe the steps taken by each archive in the pages that follow, but we introduce this issue by making these general points. First, it is essential that data producers and research projects that are collecting data be aware of their data preservation and sharing opportunities and requirements, and that the data archives work with them throughout the research life cycle. Second, the data archives need to understand that the research data world is changing rapidly, and that old models that involved archiving data only after a project was completed are no longer sufficient. For one thing,

many projects are longitudinal or otherwise long-lived, and these living collections should have the opportunity to share their data in a fashion that serves their needs. The long-lived nature of many projects also means that the creation of effective metadata is not something that is only done at the end of a long process. Rather, it needs to be done early in the life cycle, made available earlier, and potentially updated just as data in living collections are updated from time to time. Finally, the archives and the user community need to recognize the legitimate right of data producers to have exclusive access to their data for some period of time, and create mechanisms that ensure early preservation and metadata creation while allowing delayed release, perhaps through a series of well-articulated embargoes.

Harvard IQSS

The Henry A. Murray Research Archive (MRA) collects and preserves all types of data of interest to the social science research community, including numerical data, qualitative text, video, audio, and other emerging forms of social science data.⁸ As long as the data were collected or analyzed as part of a research design aimed at answering social science questions, it is a candidate for selection into the MRA. In order to accommodate this flow of data, the MRA has automated every stage

of its archival workflow. Ingestion of materials, cataloging, and processing are all accomplished through the Dataverse Network (DVN) system.⁹ The DVN supports the creation of virtual archives, called *dataverses*, for scholars and organizations (King, 2007). Individual dataverses are self-contained data archives virtually hosted on the DVN.

The owner of each dataverse controls its branding, graphic design, content, and dissemination rules. The DVN system automatically creates permanent citations for the data; converts data in common statistical software formats to preservation-ready formats; extracts and captures structural, descriptive and preservation metadata; enables discovery through browsing, searching, and harvesting; and automates usage agreements. The MRA endowment ensures permanent bit-level preservation of data stored in the IQSS DVN Network, and the MRA staff provides supplementary professional cataloging and documentation of selected studies. Together this combines many of the advantages of professional archiving with "self archiving".

The MRA's forward-looking collection development strategy is built upon finding ways to use the DVN and its virtual archiving functionality to integrate archiving into the institutional processes of organizations that produce data and intellectual works based on data. In many cases, the end result is a living collection of data that is managed by the publisher,

with preservation guaranteed by the MRA. For example, during the first phase of the Data-PASS project, the MRA worked with two journals, *International Studies Quarterly* and the *Annals of Applied Statistics*, to create virtual archives that the journals now use on an ongoing basis to store replication data for the articles they publish.

IQSS is now extending the DVN system to integrate seamlessly with popular journal management software. This lowers technical barriers of archiving replication data, as well as makes it easier for journals to monitor their own data replication and citation policies. And this provides a positive incentive to authors, since depositing data in the system enables separate standardized citation of their data. The IQSS staff expect that this will also change incentives - as more journals enforce data deposit, authors will have increased incentives to plan for data archiving in their research projects. More generally, technological infrastructure that supports data producing institutions to manage their data will lower the costs and increase the benefits of data preservation for those organizations and their stakeholders.

ICPSR

ICPSR has a broad overall collection development strategy that involves many kinds of data, and is supported by a variety

of funding sources. These include the kinds of investigator-initiated projects described as part of this article, but also U.S. and international data that are collected by government agencies, polling firms, researchers, and others. For the Data-PASS project and its sequels, ICPSR will continue its focus on research data with support of NIH, NSF, and other important funding organizations. It will target data collections funded by new and recent awards made by NSF, NIH, and other funding organizations.¹⁰ This prospective approach to building the ICPSR collection will be accomplished by automating as much of the acquisition workflow as possible, thereby allowing ICPSR to cast a wide net for the identification of research data. This entails ongoing review, selection, and appraisal of NIH and NSF funded research data, providing ICPSR with a systematic source of information about researchers who are initiating new data collections of interest to the social sciences. Thus, ICPSR can provide early and ongoing consultation about the requirements for long-term data preservation and access to researchers with valuable data collections

ICPSR has specified the steps to take in this, which focus on knowing the universe of research projects that are creating data at any given time, and building the relationships with researchers and institutions that are needed to ensure preservation of data. The first step is a weekly harvest of new

awards by research funding agencies, starting with NIH and NSF, and then adding other funding agencies over time. From those lists of new awards, ICPSR assigns staff to review project abstracts and select those projects likely to produce social science data of lasting value. The third step is an immediate contact with the Principal Investigator, expressing interest, opening communications, and asking about how ICPSR might stay in touch in the future. The subsequent steps are annual or more frequent contact until the last year of project funding, and then more frequent contact as needed until the researcher or organization agrees to or declines to deposit data. If data are then deposited, they enter ICPSR's normal processing, preservation, and dissemination process. Over the next year ICPSR will expand this process and evaluate its success.

Odum

The Odum Institute has plans on several fronts. First Odum will continue to be the archival home of the National Network of State Polls (NNSP), building on a two-part strategy. Adding to the effective partnership with IQSS and the Dataverse network, Odum is developing software solutions that will reduce the cost or effort to make a submission to the archive. The ability to provide "virtual archives" for organizations will prove to be a key function in the process of making the ingest process more

automated and less intimidating for the researchers. Allowing data producers to use automated procedures to upload studies coupled with the review of discipline based professional archivist lowers the barrier to high quality data submissions. This feature of Dataverse coupled with intervention earlier in the social science research data life cycle should help turn the tide and help increase data submission and archival rates (Green & Gutmann, 2007). Odum also will continue to follow up on their leads of organizations that have state poll data suitable for submission.

Another part of Odum's plan is to continue the connection with Harris International. The Data-PASS project helped solidify that relationship, but Odum also learned that frequent contact is necessary to maintain the flow of data. They will maintain that regular communication. A third planned Odum activity will establish relationships with private foundations that sponsor social science data collection efforts to see if Odum can archive data collected under their sponsorship. The experience with RTI and NORC taught the Odum staff that it is extremely difficult to obtain data from the organizations contracted to collect the data. The alternative is to build relationships with a select group of organizations that sponsor research, and obtain data deposits that way.

Roper

The Roper Center will continue to focus its data collection efforts on public opinion surveys from commercial survey firms and media outlets in the United States and internationally.

Within this broad scope, priorities will include efforts to solidify relationships with organizations that have historically archived their studies at the Center. In spite of years of data acquisition and relationship building experience, the Roper Center continues to feel the residual effects of the unique challenges facing the media-based polling organizations.

Personnel changes and economic pressures similar to those faced by private research organizations are exacerbated by the 24-hour news cycle. These organizations are moving targets as new polling partnerships and projects are formed on what seems like a weekly basis. Much effort will be placed on identifying critical components to strengthening these relationships, especially incentives for data archiving in this environment.

Beyond these core commercial and media-based polling firms, the Center will seek to develop relationships with other private research organizations who have conducted many polls already found in the Roper Center's TRACES database. These studies typically are done in particular subject areas by organizations such as the AARP, Kaiser Family Foundation, Carnegie Foundation for the Advancement of Teaching and the American Cancer Society.

The focus of these acquisition efforts will not be on past polls, but on establishing relationships focused on surveys of the future. Studies by these types of organizations tend to fall between the federally funded initiatives and commercially driven studies, and are more likely to fall through the cracks of even comprehensive acquisition efforts of the Data-PASS partners.

NARA

NARA's development of its Electronic Records Archives (ERA), which builds on 40 years of experience preserving and providing access to digital federal data records, positions NARA to assure the long-term viability of the burgeoning volume and variety of valuable federal digital records. ERA is NARA's strategic initiative not only to preserve and provide long-term access to uniquely valuable electronic records of the U.S. Government, but also to transition government-wide management of the lifecycle of all records into the realm of e-government (Thibodeau, 2001). As this article goes to press, the ERA Initial Operating Capability (IOC) has just been announced. With IOC, ERA will support the basic process of determining how long Federal agencies need to keep records and whether the records should be preserved in the National Archives afterwards. This stage also supports NARA in beginning the ingest of

approximately three and a half million already-accessioned electronic records files into ERA. Like so much of digital preservation work, this is a story that is in the midst of rapid and substantial change that needs to be monitored frequently. For an update on ERA as it evolves, see its web site at <http://www.archives.gov/era>.

Concluding Remarks

The importance of long-term access to and preservation of data for research and educational use is now widely recognized. In addition, the Federal Records Act covers data records created by federal agencies or their contractors, and requires a plan for their long-term disposition. Good practice is clear - data producers should plan for archiving of data early, so that data are available for future research and policy analysis.¹¹

The successes of the Data-PASS project reflect the importance of building a partnership that drew together experienced digital archives to identify, acquire, curate, and preserve a broad range of digital content. The partnership enabled us to agree on standards, work together on technology, and share the responsibility for identifying, acquiring, and preserving the content in our field of activity. The tangible result is a significant amount of digital content preserved, which constitutes one of the core goals of the NDIIPP program.

Perhaps more importantly, the partnership showed a way toward the future of digital preservation, which has been an even more fundamental goal of NDIIPP. Data-PASS demonstrated how to preserve an ever-larger share of digital social science data, and to do so in a structure that is sustainable for the very long term.

The project's success has also revealed areas where the scientific, policy and digital preservation communities need to do more. One area where we met challenges was in dealing with Private Research Organizations, which seem far less committed to publicly preserving and sharing the data they have collected than public and private opinion polling organizations, university-based researchers, and government agencies. More generally, we have reinforced our belief in the need for policy that ensures that research data supported by public funds are preserved for future analysis. Without stronger policies by NIH and NSF to begin with, and then by other agencies that support the collection of research data, we cannot be certain that this important class of digital content will be preserved.

The future plans of the Data-PASS partners and the partnership as a whole do ensure that these public and university-based archives will continue the aggressive effort that they have demonstrated thus far. We have already enriched the future of digital content for the social sciences, and with

the plans we have made for the future we ensure that ever more of the content that is important will be preserved.

References

- Altman, M. (in press-a). A fingerprint method for verification of scientific data. In *Advances in Systems, Computing Sciences and Software Engineering* (Proceedings of the International Conference on Systems, Computing Sciences and Software Engineering 2007). Springer-Verlag.
- Altman, M. (in press-b). Transformative effects of NDIIPP, the case of the Henry A. Murray Archive. *Library Trends*.
- Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (in press). Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. *The American Archivist*.
- Altman, M., Andreev, L., Diggory, M., Krot, M., King, G., Kiskis, D., Sone, A., & Verba, S. (2001). A digital library for the dissemination and replication of quantitative social science research. *Social Science Computer Review*, 19(4), 458-470.

- Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative Data. *D-Lib Magazine*, 13(3/4). Retrieved June 29, 2008, from <http://www.dlib.org/dlib/march07/altman/03altman.html>
- Anderson, M.J. (1988). *The American Census*. New Haven, CT: Yale University Press.
- Converse, J.M. (1987). *Survey research in the United States: roots and emergence 1890-1960*. Berkeley, CA: University of California Press.
- Data Preservation Alliance for the Social Sciences. (2007). *Appraisal and acquisition of actively curated collections*. [PDF document]. Retrieved June 29, 2008, from <https://www.icpsr.umich.edu/DATAPASS/pdf/LivingCollections.pdf>
- Freese, J. (2007). Replication standards for quantitative social science. *Sociological Methods and Research*, 36(2), 153-172.
- Green, A., & Gutmann, M. (2007). Building partnerships among social science researchers, intuitional-based repositories and domain specific data archives. *OCLC Systems and Services: International Digital Library Perspectives*, 23(1), 35-53.

- Heim, K.M. (ed.) (1982). *Library Trends*, 30(3).
- Higgins, S. (2007). Draft DCC curation lifecycle model.
International Journal of Digital Curation, 2(2), 82-87.
- Humphrey, C. (2006). *e-Science and the life cycle of research*
[Word document]. Retrieved June 29, 2008, from
<http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>
- Igo, S.E. (2006). *The averaged American: Surveys, citizens and the making of a mass public*. Cambridge, MA: Harvard University Press.
- International Council for Science. (2004). ICSU report of the CSPR assessment panel on scientific data and information. ISBN 0-930357-60-4
- Inter-university Consortium for Political and Social Research (2005). *Guide to social science data preparation and archiving, 3rd edition*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.
- King, G. (1995). Replication, replication. *PS: Political Science and Politics*, 28(3), 444-452.

- King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, 36(2), 173-199.
- King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49-69.
- Lanza-Kaduce, L., Dunham, R., Akers, R.L., & Cromwell, P. (1998). Policing in the wake of Hurricane Andrew: Comparing citizens' and police priorities. *Disaster Prevention and Management*, 7(5), 413-419.
- Mortimer, J.T., Finch, M.D., Shanahan, M.J., & Ryu, S. (1992). Work experience, mental health, and behavioral adjustment in adolescence. *Journal of Research on Adolescence*, 2(1), 25-27.
- National Institutes of Health Office of Extramural Research. (2003). *NIH data sharing policy and implementation guidelines*. Retrieved June 29, 2008, from http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

- National Research Council. (1985). *Sharing research data*. S. Fienberg, L. Martin, & M. Straf (Eds.). Washington, DC: National Academies Press.
- National Research Council. (2001). Principles and practices for a federal statistical agency (2nd edition). Washington, DC: National Academies Press.
- National Research Council. (2003). *Access to research data in the 21st Century: An ongoing dialogue among interested parties*. Science, Technology, and Law Panel, Policy and Global Affairs. Washington, DC: National Academies Press.
- National Research Council. (2004). *Protecting participants and facilitating social and behavioral science research*. Washington, DC: National Academies Press.
- National Research Council. (2005). *Expanding access to research data: Reconciling risks and opportunities*. Washington: National Academies Press.
- National Science Board. (2005). Long-lived digital data collections: Enabling research and education in the 21st Century, NSF. (NSB-05-40).

National Science Foundation. (2001). *Grant general conditions (GC-1)*. [PDF document]. Retrieved June 29, 2008, from <http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf>

Pienta, A. (Chair), Adams, M., Altman, M., Crabtree, J., Donakowski, D., & Maynard, M. (Panelists). (2007, August). *Data Preservation Alliance for the Social Sciences: A model for collaboration*. Session at the Society of American Archivists Annual Meeting, Chicago, IL.

Pienta, A.M., Gutmann, M., Hoelter, L., Lyle, J., & Donakowski, D. (2008, August). *The LEADS database at ICPSR: Identifying important "at risk" social science data*. Paper to be presented at the Quantitative Methodology Open Refereed Roundtable session at the 2008 American Sociological Association Annual Meeting, Boston, MA.

Smith, T., & Forstrom, M. (2001). In praise of data archives: Finding and recovering the 1963 Kennedy Assassination Study. *IASSIST Quarterly*, 25(4), 12-14.

Thibodeau, K. (2001). Building the archives of the future: Advances in preserving electronic records at the National Archives and Records Administration. *D-Lib Magazine*, 7(2). Retrieved June 30, 2008, from

[http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.h
tml](http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html)

Endnotes

¹ This project is supported by the Library of Congress's National Digital Information Infrastructure and Preservation Program. We are grateful to Lisa Quist, Piper Simmons, and Tannaz Sabet-Fakhri at ICPSR; Cynthia Teixeira, Marilyn Milliken, Terry Emmons, Hang Nguyen and Huong Hoang at the Roper Center; Linda J. Henry and Lynn A. Goodsell at NARA.... for their contributions. We are also grateful to our institutions... Harvard University, the University of Connecticut, and the University of Michigan for their part in the Data-PASS project's cost sharing. The following research projects supported some portion of the work described in this article: At ICPSR, NIH Grants: "Human Subject Protection and Disclosure Risk Reduction" P01 HD045753 (PIs: Myron Gutmann, JoAnne McFarland O'Rourke and James McNally), "Data Sharing for Demographic Research" U24 HD048404 (PI: Felicia LeClere), and "Factors in Aging: Development Research Resources" P30 AG004590 (PI: James McNally). ² For details of this technology and the shared practices developed with it see: Altman et al. (in press).

³ Roughly one-fourth of all articles published in political science and one-half of all quantitative articles use survey data. See footnote 1 in King, Honaker, Joseph, & Scheve (2001).

⁴ For NSF, see article 36 in National Science Foundation (2001); for NIH, see National Institutes of Health Office of Extramural Research (2003).

⁵ Also see Altman (in press-b) for details and reflections on this transformation.

⁶ <http://www.dataxiom.com/products/statmost/index.html>

⁷ Many of these documents are available on the project website, at <http://www.icpsr.umich.edu/datapass>. All others may be requested from the author.

⁸ Also see Altman (in press-b) on how digital infrastructure has catalyzed change in archival workflows and collections.

⁹ <http://thedata.org/>

¹⁰ One good model for this sort of lifecycle management of academic research data operates at the Economic and Social Data Service and the UK Data Archive. See:

<http://www.esds.ac.uk/aandp/create/createintro.asp>

¹¹ For clear statements of the importance of these practices see National Research Council (2001); National Science Board (2005); and International Council for Science (2004).