

Developing Partnerships in the Social Sciences: Data Archives and Private Research Organizations

Jonathan Crabtree¹, Marc Maynard², Lois Timms-Ferrara²

¹Odum Institute, 22 Manning Hall, CB #3355, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3355, USA

²The Roper Center, 369 Fairfield Way, Unit 2164, University of Connecticut, Storrs, CT 06269-2164, USA

marc.maynard@uconn.edu

Abstract. The Data Preservation Alliance for the Social Sciences (Data-PASS) is a partnership of six major U.S. institutions with a strong focus on archiving social science research. The partnership is supported by an award from the Library of Congress through its National Digital Information Infrastructure and Preservation Program (NDIIPP). The goal of Data-PASS is to acquire and preserve data at-risk of being lost to the research community, from opinion polls, voting records, large-scale surveys, and other social science studies.

This paper will discuss issues and challenges faced by two of the Data-PASS partner archives in efforts to pursue data sharing arrangements with a variety of private research organizations (PvROs) specializing in social science research. After reviewing the background of the project, we look at the role of private research organizations in social science research within the US, followed by a general overview of issues of data sharing in the social sciences and particular challenges faced by social science archives and researchers locating and accessing data based on PvRO research. Next, we explore paths for collaboration and potential solutions. We conclude with some brief comments on the partners' future plans for pursuing these types of data resources.

Introduction

Active social science data sharing and preservation has occurred since the late-1940s and 1950s by researchers and through the formation of social science data archives around the world. The data archives have taken on the responsibility of identifying, acquiring and preserving important social science data collections for use by future generations of researchers. It has been unclear how much data could or should be preserved in this manner, but one can make an educated guess that the vast majority of potential studies have not been saved and may have been lost to researchers. Over the past three years, with support from the Library of Congress, social science data archives have been working together to track down potentially 'at-risk' data sets in an effort to expand the available materials for future generations.

The Data Preservation Partnership for the Social Sciences (Data-PASS) is a broad-based partnership led by the Inter-university Consortium for Political and Social Research (ICPSR)

at the University of Michigan, the Roper Center for Public Opinion Research at the University of Connecticut, the Howard W. Odum Institute at the University of North Carolina-Chapel Hill, the Henry A. Murray Research Archive, a members of the Institute for Quantitative Social Science at Harvard University, and the National Archives and Records Administration. The project is supported by a matched grant from the Library of Congress through its National Digital Information Infrastructure and Preservation Program (NDIIPP). Generally, the project goal is to ensure the long-term preservation of existing data holdings and of materials that have not yet archived. The partners seek to acquire and preserve data at-risk of being lost to the research community, from opinion polls, voting records, large-scale surveys, and other social science studies.

The Data-PASS partnership has provided an impetus for different models of collaboration: public-private, academic-commercial, and academic-government. One particularly promising collaborative model includes the academically-based data archives and commercial or private research organizations that have played such a central role in the advancement of social science research in the United States. Given their collection coverage and familiarity with commercial organizations, the Odum Institute and the Roper Center each began the process of establishing relationships with several private research companies in an attempt to recover 'at-risk' data sets from the past, but also to investigate ways to set a foundation for preserving future study datasets.

Founded in 1924, the H.W. Odum Institute for Research in Social Science is the nation's oldest multidisciplinary social science university institute. Indeed, it is the oldest institute or center at the nation's first public university, UNC-Chapel Hill. The mission of the Odum Institute parallels that of the university as a whole: teaching, research, and service, but the institute's focus is on the social sciences. Odum holds archived Harris polls dating back to 1958 and continues to update its Harris archives with data from Harris Interactive bimonthly telephone poll. Those data cover diverse topics, including the National Organization on Disability polls and the Equifax Privacy surveys, a topic of particular salience today. The Odum Institute also serves as a national archive of individual U.S. states' polls and has one of the most extensive collections of 1970 US census data as well as other diverse collections.

The Roper Center was founded in 1947 with the mission of promoting the responsible use of public opinion data, especially through encouraging secondary analysis of individual surveys and extensive data analysis, making use of information collected at different places and times. In the early years, pursuit of the Roper Center's mandate primarily took the form of building a collection of respondent-level data sets. As survey research exploded in the decades since, the Center has focused on building long-term relationships with major media and commercial survey organizations resulting in expansion of the data library to over 16,000 surveys. The Gallup polls are the largest and the most comprehensive component of the collection. Dating back to 1936, the Gallup Poll represents a priceless longitudinal record of American opinion. Beyond a myriad of media polls, specialized survey data collections from Fortune Magazine (1938–1949), Samuel Stouffer's American Soldiers Surveys (1942–1945), United States Information Agency surveys (1952–2000), along with recent acquisitions of Henry J. Kaiser Family Foundation surveys, and Public Agenda Foundation studies, provide researchers with both a historically comprehensive, yet contemporary data resource.

The combined collections of the Odum Institute and Roper Center archives constitute one of the largest in existence. Despite the comprehensive and expansive nature of their holdings, it is unclear how many studies and how much data has not been preserved. One of the largest

potential “untapped reserve” of social science data lies in the holdings of private research organizations that have played such a prominent role in social scientific research in the past half-century. These organizations seem to be natural partners for the Data-PASS project.

Data-PASS project goals

In September, 2001 Tom Smith, senior researcher at the National Opinion Research Center at the University of Chicago, recalled a survey done just after the assassination of President John F. Kennedy that attempted to document the private emotions felt by the public at that time. His intent was to ask some of the same questions just after the 9/11 terrorist attack in the US to compare the emotional state of the nation at these two points in time. Smith documented the process of identifying the study, locating the data and documentation, reading the data—stored only on punched cards, and eventually handling the multi-punched data format and creating a usable dataset. This focused effort took hours of time over four months and involved bringing a former archivist at NORC out of retirement to assist in the rescue, driving boxes of cards representing these data across half the country to a card reader located in New York (Smith, 2001). This exercise became the catalyst for the Data-PASS project and NORC is one of the PvROs involved. NORC deserves accolades for maintaining these data and the other data that the Roper Center is working with them to secure.

The experience regarding the Kennedy Assassination data, unfortunately, is not unique. Until recently many private businesses and university-based researchers have assumed 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. There are a variety of understandable reasons for this lack of attention to preservation. Some individual researchers have been reluctant to deposit their data in archives because they wanted to avoid sharing it with potential competitors. Some lacked the time or expertise to prepare the metadata required for effective sharing. And some investigators simply did not recognize the long term value of their data. Institutional data producers may have been under contractual obligations with those who paid for data collection to protect proprietary information. And some data just fell through the cracks.

There remains a vast quantity of digital social science research content that has not been and will not be archived without aggressive pursuit by data curators. This content lives on in the computers of individual researchers or of research institutions, or quite possibly in bookcases, libraries, and warehouses and will be unavailable for scientific research unless it is found, shared and preserved.

Data Sharing and Preservation in the Social Sciences

Data sharing for scientific advancement has been a long held principle within many disciplines. Examples of successful data sharing partnerships include economics, meteorology and the Human Genome Project (Weil, 1991). Within the social sciences an ethic of data sharing has been adopted in many subject specific areas and underlies the role data archives have played to formalized data acquisition, preservation and access (Rockwell, 2001).

Sharing data with the larger research community provides a number of benefits to researchers: it makes research easier, eliminates redundant data collection, allows specialists to collect data and subjects the research results to replication/triangulation by others (Belk, 2007). Additionally, numeric data permits analysis at levels of complexity and depth not

otherwise available, can be used for research purposes unanticipated at the point of development and collection, permits extensive historical and trend analyses, and finally, can be employed to teach and train new researchers (Lievesley, 1999).

From an academic research perspective, the benefits of data sharing are clear and readily accepted to the point where the National Science Foundation, as well as, numerous journals have codified requirements that funded data collections (in the case of NSF) or data used for publication (in the case of journals), be made available through a public archive mechanism (Rockwell, 2001).

The Data-PASS project, though, is not only seeking to uncover data produced by academic researchers, but is also interested in data produced by private and commercial entities. In such an environment benefits of sharing must be clearly identified and articulated within a separate and more complex context. Lievesley (1999) states that “many of the greatest benefits (to data providers) can be classified as altruism,” including some of those mentioned above as well as contributing to the development of knowledge, ensuring data are exploited, and facilitating comparative research. But in light of an Internet environment in which institutions are negotiating cross-pressures of the democratization of information and changing intellectual property rights, these private and commercial research entities require more than just altruistic arguments to unlock their data collections or take on the expense of searching for older datasets.

For private research organizations, recovering these studies and data collections for their own use provides a vehicle to better research and, in some cases, better business. One such benefit is the resulting electronic access to their commercially significant datasets that will provide PvROs the ability to track what is being downloaded and used by prospective clients. They can use this information to guide future research agendas and to better direct resources. Uncovering potentially new and distinct research areas based on detailed analysis of previous research efforts is also of undoubted (if difficult to measure) value to the field of social science research. Additional benefits to the PvROs might include reduced storage costs, access to fully migrated data resources and digitization of internal library metadata records (many of which are themselves paper-based, fractured and at risk).

While the potential benefits to both the research community and data providers are numerous, data sharing still presents challenges for full acceptance across disciplines. Further understanding of private research organizations provides the context for identifying and developing solutions to those challenges.

Private Research Organizations

The late-1940s and 1950s witnessed the rise of private organizations and firms that deal almost exclusively in the production and analysis of information, knowledge, and public policy. These organizations are potentially a major source of social science research on important theoretical and public policy issues. They do much of their work under contract with public and private agencies and these agreements may not have included requirements that data collected and analyzed also be archived.

The Data-PASS partnership has provided a platform to uncover and explore this vast research trove developed by PvROs over the past half century. Organizations such as Research Triangle Institute (RTI International), the National Opinion Research Center (NORC), Westat, and ABT Associates have played primary roles in the advancement of scientific

research in the social sciences. They are involved in a significant portion of governmental and scholarly research in substantive areas of social, health, and cultural research. Other organizations such as the Henry J. Kaiser Family Foundation, Public Agenda Foundation and AARP also now produce high quality social science research on a myriad of topics from education, health care and aging, among others.

What we have termed private research organizations are not a homogeneous group of institutions. PvROs fall under a number of operating models depending upon their institutional mission, connection to a university, presence of a broad research agenda, technical capacity, and substantive expertise. As with any other type of organization, private research organizations vary greatly in their purpose, size, internal structures, operational styles and settings. For the purpose of this project, it seems that private research organizations can be grouped broadly on two dimensions regarding purpose and commercial status.

The purpose of a private research organization can be defined as either research driven or service driven. A research driven organization maintains an active research agenda limited only by the mission and management of the organization. They conduct research within specific subject areas either for their own interests (publicity, policy development, etc.) or in collaboration with funding agencies with similar interests. Service driven organizations (or contract research organizations) provide external clients with the substantive or methodological expertise, as well as the infrastructure to conduct research. These organizations typically do not have a specifically focused research agenda.

Private research organizations may be purely commercial for-profit firms, non-profit entities, reside in an academic setting, or be some type of hybrid. The legal forms these organizations take can (and often do) influence what potential interest they have in considering data sharing and preservation.

Long-term relations between data archives and PvROs have been uneven and sporadic at best. These relationships have been impacted by industry trends, research advancements and archival mandates. In some cases, past experiences have laid a foundation on which to base further development of data sharing and preservation efforts.

Since their inception The Odum Institute and Roper Center have been working with data producers to acquire and preserve important social science information. A key factor guiding those acquisition activities has been growing recognition that the relationships developed with researchers, i.e., data producers, are critically important. Just as research technologies have evolved over time, so too have the organizations guiding the task. This has required archiving institutions to consistently build new and renewed relationships (Crabtree, 2006). Both Odum and Roper have sought to develop relationships with local private research organizations as a part of this on going effort. In some cases this culminates in the co-sponsoring of research projects or the co-funding of research positions. The Data-PASS experience has allowed us the opportunity to build on these existing relationships. The existing relationships were instrumental in getting the appropriate people within the organizations to the table and resulted in the discovery of many important studies currently not archived.

During the discovery phase of our project we began to see some of the many challenges we were to face when working with the private research organizations. These challenges were

not insurmountable, but require more planning and substantially more time than originally anticipated.

Challenges

Within the current culture of PvROs, the Data-PASS partners' experiences noted that the barriers that most affected the rescue effort fell into one or more of four categories: attitude, economics, technology, and the broad category of logistics. (Crow, 2002).

Attitude Towards Archiving Data

In general, initial discussions with PvROs were encouraging and constructive. There was mutual agreement about the need to preserve the data in perpetuity, and there was the sense of camaraderie in the endeavor. For those private research organizations with substantive interest in the outcomes of the research, the argument was obvious as they were part of the social science community that would benefit from the data becoming accessible. Some studies for which the PvRO served in the capacity of executing the research rather than as a full study partner, the rescue effort had less appeal. Studies of both types suffered the fate of being misplaced or placed in locations unknown to current staff. While prospects of the project elicited positive responses, it was usually the case that optimism on the part of the PvRO was not enough to circumvent other obstacles.

Economics

The benefits of archiving are perhaps not convincing enough, in and of themselves, to PvROs within the context of their core business operations. The research and economic climate is not what it was when the first PvRO data collections were created, acquired and archived a half century ago. Today, PvROs are faced with a myriad of questions regarding the economic impact of documenting and preservation including: (1) If datasets are assets, what is their value to our organization or others? Can they be used to leverage existing research or identify new areas of interest? (2) How can we build a business case for preservation at our firm? Can we make sure that our datasets are "born digital" as an effort to make preservation affordable? (3) Additionally, would archiving data with appropriate documentation be an asset in funding proposals and more likely to support long term relationships with funding agencies?

The archives had not anticipated the extent to which these economic complications came into play. The argument that these selected studies were assessed by a team of archivists and deemed to have value to future social science researchers was met in the context of the current business model: if they have value, then the PvRO should be take steps to realize any financial benefits these data could yield. Furthermore, the economic climate incorporating the concept of 'billable time' had to be considered as any resources allocated to the rescue and duplication effort had to be compensated. Access and communication with key personnel within the PvRO proved difficult, primarily due to the lack of financial resources to cover their time. Fees for the private warehouse facilities to access stored PvRO materials also had to be budgeted and strategies configured to minimize such labor. On the other side of this economics equation was the argument that once the materials were retrieved from the warehouse and properly archived, the storage expense would dissipate for the PvRO. Physical distance between the archives and the PvRO inhibited action and the sharing of archive staff resources.

Technology

Within the scope of this project, technological challenges, both in terms of technical advancement and obsolescence, were evident. The advent of Internet technology has provided a false sense of security that companies have archived their data by making it available on their web site. Web sites are in constant flux and pages are refreshed daily, resulting in replaced and discarded content. Polling firms will often place data on a web site while the topic is fresh, but will replace those data when it becomes dated. When organizations do this, there is rarely a longer term plan to maintain those data in up-to-date usable formats in perpetuity.

In a case of historical data that was to be rescued by the Data-PASS team, the data collected by the PvRO existed only on the original punched cards prompting additional technological concerns. First, the quality of the punch cards was critical to the rescue effort; many of studies were boxed and stored for decades. Second, a card reader was required. Finally, while the project had secured a card reader for one of the archives, negotiating the removal of the only copy of the data in existence from the PvRO's facilities and identifying mutually agreed upon modes for transporting the materials proved another challenge.

Logistics

Logistical obstacles persisted at every point. Given the broad nature of the targeted studies and the sensitive issues of confidentiality, often more than one unit within the PvRO was involved and that meant conferring with different personnel and navigating unfamiliar organizational politics. Staff changes within the organizations occurred between initial discussions and follow up conversations, hindering commitments made with key staff at the outset of the effort and making data and metadata more difficult to definitively identify and locate. Issues such as privacy and confidentiality with respect to both the respondent and the funding or sponsoring agency had to be addressed. Contractual obligations on behalf of the PvRO had to be reviewed and ultimately, permissions obtained from the funding agency for release, preservation, redistribution and rules governing access. Finally, assuming these concerns can be addressed in a satisfactory manner, the acquisition team must still locate the storage media and appropriate supporting documentation or persuade the PvRO to integrate data archiving activities into their workflows. In the case of historical data collected in the 1950s and 1960s where the principal investigators were no longer available, the difficulties lay with cryptic finding aids and tracking down boxes of punched cards and documentation that had been stored at multiple facilities.

The private research organizations are not the only vehicle to recover these at-risk data sets. Funding agencies and research partners also provide another point of entry. Funding agencies tend to be the ultimate owners of the research data and therefore must be contacted to provide release clearance and other permissions. They may provide the best entry point for pursuit of a particular study due to their client status of the PvRO.

Paths for Collaboration

Attitude Towards Archiving Data

The PvROs expressed significant interest in the effort and understood the value of the studies, so developing cogent strategies to move the collaboration forward merited the effort. It is well known that archiving data is good practice as the archives can be depended on to (a) be sure the data exist in perpetuity, backing up holdings regularly; (b) maintain the data and documentation in current formats; (c) index and create finding aids to let researchers know they are available; and (d) provide access to the research community. However, arguments

for the PvROs to archive their data had to be made soundly and with specificity. The Data-PASS Operations Committee developed a list of criteria for data selection that included specific recognition for content that exhibited these characteristics:

- Research that was theoretically and/or methodologically ground breaking
- National samples, important regional samples, or samples of hard to reach populations
- Data collected as part of a policy evaluation
- Data cited as a part of a seminal collection or tied to unrepeated or rare events

Defining how the data collection fits into the larger scheme of social science research data helped identify its place within the archives. Stand alone datasets are important, but they become critical when set within the context of the broader collection of related science that is properly archived for use by the scholarly community. In the case study cited earlier involving Tom Smith of NORC, the stand alone data no doubt had value, but its value increased exponentially once the questions were replicated for comparative purposes. This argument also offsets the notion that if these data have value, then the PvRO ought to save the data to reap any financial gain. The true value resides in the ability to analyze all of the related collections.

Economics

It became clear that altruism alone would not carry the projects; pragmatic approaches to the economic realities of the PvRO had to be taken into account, notwithstanding the limited resources of the archives. It became understood that PvRO resources dedicated to the endeavor could be scarce, particularly if the data were not current or difficult to locate. Through Data-PASS, financial arrangements were made to accommodate the costs for duplication, and eventually funds were made available to offset salary costs for staff involved in the recovery. Another approach was for the archives to offer staff assistance to the PvRO or off-site storage facilities to assist in data recovery. This option was well-received, but ultimately, for unrelated reasons, was not utilized. In the case of recovering historical materials, it was necessary to take into account access fees for private storage facilities and to document cost savings for the PvRO once archiving the materials would eliminate external storage charges. The archives learned to proactively take steps to streamline any bureaucratic processes that could delay transactions for reimbursement.

Technology

Social science archives take seriously their obligation to preserve and maintain the data that is entrusted to their care. This task is not the mandate of the data collectors or the financiers of survey work. Once the curation process was articulated, one PvRO quite sensibly requested that a copy of the data be returned in their enhanced state so they might be available via the PvRO's own web site in addition to the archives'. The archives, in many cases, can provide certain value-added services that the PvROs are not willing or able to undertake themselves. It became more apparent that leveraging the strengths and expertise of the archives for the purpose of upgrading and migrating datasets, developing richer metadata, and overlay existing tools for data discovery and exploration were potential discussion points toward unlocking PvRO data collections.

Logistics

Work flow problems were difficult to overcome, but once the ebb and flow of the PvRO workplace was distinguishable, the archives were sometimes better able to offer assistance and divisions of labor to move the project forward. Communication was a struggle as the time

of PvRO staff was stretched on billable projects. Scheduling briefing time in person or via telephone proved difficult and ultimately electronic communication proved most successful. While face-to-face contact was rare, it did serve to invigorate the process and jump start various tasks along the way.

The archives prepared written statements about the intentions of the project to be shared with the funding agencies in order to receive approval for the transfer of data. The matter of who would make the contact required some tact and was ultimately a matter for the PvRO to decide. Since the PvRO would have to engage resources to provide the contact information to the archives anyway, in most cases the private research organization opted to make the inquiry themselves. The archives had to be clear to the PvROs and the sponsoring organizations about why the data were important and how the archives could add value to them. Preservation was part of the equation, as was the reformatting of the historical data, but the benefits of easy access to these data for the PvRO itself proved an important argument.

Conclusions and Future Work

The Data-PASS project has provided the catalyst to begin the process of archiving important studies produced by the diverse group of PvRO's around the country. While our effort will temporarily culminate with the rescue of many "at risk" datasets, more work needs to be done in order to invoke a lasting change. Preservation workflows are at the core of the data archives yet within the complex organization of the PvRO they get lost in the economics of opportunity costs. The partners will continue to work with the PvROs to determine where and how these tasks can be integrated to their organizational structure.

In order to make a substantive change in the process we will need to bring all the vested parties to the table. The funding agencies are the ultimate owners of the data and are vital to this endeavor. Unfortunately the economic environment around data producing projects tends to force researchers to trim the funding for archival efforts to meet budget requirements. Some funding agencies have policies that "require" the data produced be archived. While this is a good first step it leaves out the critical component of relationship building. Archives have the incentive and seek to build partnerships that potentially lower this cost. The funding agencies, PvROs, and the archives could provide a team of experts with the knowledge, incentive, and means to solve many of the barriers to archiving these valuable works.

Productive partnerships leverage each party's strengths and the solutions to some of the caveats encountered have to creatively address each organization's needs. In future work, the Data-PASS members seek to bring these parties together. Our intent is to sponsor workshops that bring members from funding agencies and PvROs to meet and talk with archives about the issues around data preservation. Each group has existing ties either through funding or collaboration but we feel a focused, face-to-face effort would provide the most traction to keep things moving. Our focus will be to enter into open discussion intent on exposing the barriers to archival realized in our past efforts as well as uncovering other potential problems. Our goal will be to bring these issues into focus and provide solutions developed within these newly defined relationships.

Patience and perseverance gives way to persistent resolve and the archives have now made headway into some of the collections of some of the PvROs. There is much more to be done. Further discussions among the groups will demand more imagination and ingenuity, but all parties seem to agree that the strengths of the archives are different from those of the private

research organizations, and yet, equally important. Smith and Forstrom said it best after rescuing the 1963 Kennedy Assassination poll:

“Survey data must be sent to survey archives...where the documentation and data will be preserved, backed-up, periodically updated as technologies change, indexed, and made routinely and easily accessible to researchers. Failure to archive studies is poor science and a disservice to other contemporary researchers and those in the future.”

Tom Smith and Michael Forstrom, *IASSIST Quarterly* Winter 2001

Acknowledgments

This work was supported in part by an award (PA#NDP03-1) from the Library of Congress through its National Digital Information Infrastructure and Preservation Program (NDIIPP). The authors would like to thank the Odum Institute, the Roper Center and our Data-PASS partners, especially those individuals who have worked so diligently to advance the PvRO discussions. In addition we also extend our gratitude to the private research organizations, especially National Opinion Research Center (NORC) and Research Triangle Institute (RTI International), for their assistance.

References

- Belk, Russell, (May 2007), “Why Not Share Rather Than Own?” *The Annals of the American Academy of Political and Social Science*, Vol. 611, pp. 126-140.
- Crabtree, Jonathan and Darrell Donakowski, (2006), “Building Relationships: A Foundation for Digital Archives.” Paper prepared for the Joint Conference on Digital Libraries (JCDL), held in Chapel Hill, NC 2006.
- Crow, Raym, (August 2002) “The Case for Institutional Repositories: A SPARC Position Paper.” *ARL Bimonthly Report* 223.
- Green, Ann G. and Myron P. Gutmann, (2007), “Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives,” *OCLC Systems & Services*, Vol. 23, Issue 1, pp. 35-53.
- Lievesley, Denise (1999), “Sharing and Preserving Data for the Social Sciences.” Paper prepared for the Workshop on Infrastructure Needs for Social Sciences, held in Ottawa, Canada 1999.
- Rockwell, Richard, “Data Archives: International,” *International Encyclopedia of the Social and Behavioral Sciences*, Paris: Elsevier, 2001.
- Smith, Tom and Michael Forstrom, (2001), “In Praise of Data Archives: Finding and Recovering the 1963 Kennedy Assassination Study,” *IASSIST Quarterly*, Winter 2001, Vol. 25, No. 4, pp. 12-14.
- Weil, V. and Hollander, R., “Normative Issues in Data Sharing.” *Sharing Social Science Data: Advantages and Challenges*. Joan Sieber, (Ed.). (1991). Newbury Park, CA: Sage Publications.