

Data Preservation Alliance for the Social Sciences: A Model for Collaboration

Authors (listed alphabetically):

Micah Altman
Institute for Quantitative Social Science
1737 Cambridge Street
Harvard University
Cambridge, MA 02138
Micah_Altman@harvard.edu

Jonathan Crabtree
Odum Institute
Manning Hall, CB #3355
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599

Darrell Donakowski
Inter-university Consortium for Political and Social Science
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106-1248

Marc Maynard
Institute for Social Inquiry/The Roper Center
341 Mansfield Road, Unit 1164
University of Connecticut
Storrs, CT 06269-1164
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 three of the significant products that have emerged from this partnership: (1) procedures for identifying and selecting “at risk” digital materials identified by the Partnership (2) the identification of “at-risk” social science data collections from individual researchers, as well as private research organizations, (3) the design and implementation of a shared catalog describing the data holdings of all partners. We conclude with some brief comments on the partners’ future plans to develop an inter-archival syndicated storage service.

Introduction

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. Despite this notion, an international movement to archive, preserve, and share data emerged when digital data began to appear in volume. Still, we cannot say that even a majority of the digital social science research content created since the revolution in sample surveys and production of digital data has been preserved.

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 without aggressive activities 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. If we do not take steps to preserve it, it will be lost forever, and its value to our society cannot be restored. It needs to be identified, located, assessed, acquired, and preserved.

Four major American social science data archives, The Inter-university Consortium for Political and Social Research, The Roper Center for Public Opinion Research, The Howard W. Odum Institute for Research in Social Science, The Henry A. Murray Research Archive, along with the Harvard-MIT Data Center (a leader in digital library research) and the electronic records custodial division of the National Archives and Records Administration (NARA), have created the Data Preservation Alliance for the Social Science (Data-PASS) to ensure the long-term preservation of our holdings and of materials as yet un-archived. ¹ We 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. And we work together to identify, appraise, acquire, catalog, and preserve data used for social science research.

Identification and Selection

While our organizations have a history of collaboration, this official partnership has provided important benefits and taught us a great deal about the advantages of formalized collaborative relationships. Data-PASS is, in part, funded by an award from the U.S. Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP) [2]. The NDIIPP mission is to develop a national strategy to collect, archive and preserve digital content, especially materials created in digital format. Our project is working to ensure the long-term preservation of the vital heritage of digital material that allows our nation to understand itself, its social organization, and its policies and politics through social science research.

¹ The Data-PASS project website is: <http://www.icpsr.org/DATAPASS/> . All of the good practices documentation developed in this project, including the identification, appraisal and metadata practices are available from: <http://www.icpsr.org/DATAPASS/about.html> . The shared catalog is available from <http://vdc.hmdc.harvard.edu/dataverse/DATAPASS/> ,

Adopting common standards for any collaborative effort lays the groundwork for those relationships to grow and prosper. The Data-PASS partnership permits a much higher level of inter-archival cooperation, including mutually agreed-upon identification and appraisal policies. The potential volume of information which could be acquired and the need to make the most cost-effective use of limited resources have emphasized the need for selection standards.

The current focus of our project is to identify the most significant digital social science data of the past seventy-five years. We start with the premise that any social science data that is not currently in a permanent archive is considered to be at risk of being lost. If data are available at an alternative site and if there is confidence that availability will continue over time, the risk of loss is diminished. An operations committee, with representatives from each partnering organization, developed common standards that are used to identify and select data for inclusion. These criteria incorporate elements of accepted archival practice to identify the most important content to preserve and an evaluation of the risk of losing the content should acquisition not take place. The appraisal guidelines include significance of the data to the research community, significance of the source and context of data, and the uniqueness and usability of the data.

The identification and selection process is somewhat decentralized with each archive pursuing data that best represent its content area of specialization. This decentralization allows each partner to leverage their distinct capabilities in specific kinds and sources of data. However, the information gathered regarding specific data collections is brought to the committee to determine how best to proceed. Together, we try to determine if the data are from studies that were theoretically and/or methodologically groundbreaking. Other data collections of interest are from studies that are part of a seminal collection or tied to unrepeatable or rare events. We also determine if the data is highly cited in the social sciences or conducted by highly cited social scientists.

As part of this process, we communicate with the producers of the data to determine their willingness to archive their data. Building and maintaining the relationships between data producers and data archives are among the most important tasks an archivist has. Those who deposit their data with archives must trust that the archive will value and preserve the information they provide and ensure that the data will remain accessible over time (Crabtree and Donakowski, 2006). Many of the data producers we encounter already know the value that each of the partnering organizations places on data preservation. Through Data-PASS, this commitment to preservation is made stronger by mutual agreements to share preservation and dissemination obligations.

Another factor influencing our interactions with individual data producers and non-profit research organizations alike is the set of data-sharing policies adopted by sponsors of research activity, such as NSF and NIH. By depositing their data in a digital archive, researchers can fulfill grant obligations that require that funded research be made available to the research community. In addition, they can avoid the administrative tasks associated with ensuring the safekeeping of the data. Depositing their data also enables researchers to demonstrate continued use of the data after the original research is completed, which can improve their prospects of securing further research money.

Federally Funded At-Risk Materials

Some federal funding agencies stipulate that data collected using their funds should be made available and shared with other researchers. The National Science Foundation, in its Grant Proposal Guide, states that it “expects PIs to share with other researchers, at no more than incremental cost and within a reasonable time, the data, samples, physical collections and other supporting materials created or gathered in the course of the work.” (NSF, 2004). The National Institutes of Health (NIH) state in its Statement on Sharing Research Data that “data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health”, and it “endorses the sharing of final research data to serve these and other important scientific goals” (NIH, 2003). In addition, any data that is produced under a federal *contract* is formally a federal record, and is subject to review for preservation by NARA. This federally funded research is a main focus of our partnership.

One of ICPSR’s roles in this partnership is to review the National Science Foundation (NSF) database. We are also reviewing the Computer Retrieval of Information on Scientific Projects (CRISP) database for federally funded data awarded between 1972 and 2003 by the National Institutes of Health. The information we retrieve is then placed in a database that includes abstracts describing the research and names of principal investigators and their institutions. Both completed and in-process research are included in the database.

We are currently in the process of contacting the principal investigators of these studies to determine if, in fact, they have made their data available for data sharing. So far, we have attempted to contact 1594 investigators (out of 5229 investigators identified in total), This sample comprises the set of PI’s who are listed as having been an investigator on only a single project. These contacts yielded 543 responses. Preliminary results suggest that few studies have been actually been archived.

LEADS PIs With Data (n=543)

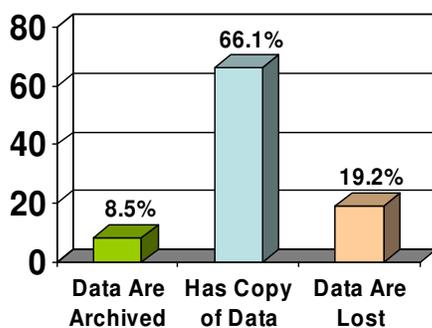


Figure 1: Preliminary Findings Regarding Data Archiving

From these initial contacts, we are finding that only 9% of principal investigators have actually archived their data with an organization such as ICPSR. Of the two-thirds who retain their own data, it is likely that only a small fraction of them make their data readily available (e.g. through a personal website). We are currently working with all of those who have not permanently archived their data to deposit it for preservation by the Data-PASS partners.

We are finding that there are a variety of reasons why data has not been archived. Some principal investigators have destroyed data stored on magnetic tapes because they mistakenly came to believe that there was no way to recover the data from those formats, some have discarded data that were in formats that they believed to be extinct. One role that data curators can play is to educate members of their research community regarding what may truly be lost and what may be recoverable. Many of those who still retain their data express a willingness to provide it on request, as long as the formats of the data and documentation are still accessible. As data curators, we must work with researchers to alleviate the barriers that prevent the preservation of their data and ensure that their data and the supporting documentation remain a long-term resource to the research community.

As individual researchers sometimes have difficulties depositing their data with archives, so too do the private non-profit research organizations that produce vast amounts of digital data. Often they are willing to share and archive those data, but need help overcoming organizational barriers. These barriers can range from the previously mentioned culture of not sharing data to an economic burden in preparing data for secondary analysis. Data archives must work with these organizations to determine which data collections are worthy of preservation and how the archives can work with the data producing organizations to alleviate perceived burdens.

Expanding Partnerships: Data Archives and 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 public policy issues. They do much of their work under contract with public and private agencies, and these agreements may not have requirements that data collected and analyzed also be archived.

The Data-PASS partnership has provided a platform to explore, uncover, acquire and preserve this vast research trove developed by private research organizations (PROs) 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. While successful and long-term relations between data archives and these types of organizations have been uneven and sporadic at best, their role in the world of social science research make them a natural partner in the work of NDIIPP and specifically the Data-PASS partnership. Discussions have begun with several PROs in efforts to develop strategies to identify and recover older materials, as well as set a foundation for future arrangements for digital preservation.

The potential benefits to any number of stakeholders in this type of arrangement are plentiful. Preservation of surveys conducted by PROs that are currently at risk of being lost or misplaced is one of the primary goals of this collaboration. For researchers, these studies represent an untapped resource to supplement already available materials. For the PRO, recovering the value of these studies for their use provides a vehicle to better research and better business. One such benefit is the resulting electronic access to their commercially significant datasets that will provide PROs the ability to track what is being downloaded and used by potential prospective customers. 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 PROs could 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).

It is unclear whether the benefits are convincing enough, in and of themselves, to PROs within the context of their core business operations. The research and economic climate is not what it was when the first PRO data collections were created, acquired and processed. PROs must ask these questions in order to squeeze value out of preservation:

- If datasets are assets, what is their value to our PRO? Can they be used to leverage existing research or identify new areas of interest?
- Do datasets have value to other organizations that might be willing to pay for them?
- What is the best way to identify datasets with commercial value or historical significance?
- What legal, technical and financial issues are involved?
- Can we make sure that our datasets are “born digital” as an effort to make preservation affordable?
- How can we build a business case for preservation at our PRO?
- Would archiving data with appropriate documentation be an asset in funding proposals, (since the funding agency would then have documentation and access to their data into the future)?

In addition issues such as privacy and confidentiality with respect to both the respondent and the funding or sponsoring agency must be addressed. Contractual obligations on behalf of the PRO must be reviewed. Ultimately permissions must be obtained from the funding agency for release, preservation, redistribution and rules governing access. PROs are businesses that must respond to the economic situation in which they operate, therefore any benefit must also provide a reasonable economic incentive to be successful. 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 PRO to integrate data archiving activities into their workflows.

However, the PROs 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 PRO. Funding agencies may be able to set forth pre-conditions as part of contractual language for archiving final versions of the data.

Much work still remains, but Data-PASS efforts have identified a number of conceptual issues that must be addressed before firm PRO partnerships can be initiated. Business and economic concerns certainly outrank other challenges faced by the archives. Assessing completed projects and translating their value into real numbers is hard for PROs to do. Additionally, modifying workflows to more efficiently and effectively gather materials for the purpose of archival activities, while desirable in theory, is very difficult to implement in practice. It is our

hope that with a coordinated effort we will be able to help provide a valid business model and approach to some of these issues faced by PROs, funding agencies and data archives alike.

Expanding Partnerships: NARA – Roper Center Collaboration

Data-PASS partnership has provided the impetus for different models of collaboration: public-private, academic-commercial, and academic-government. A notable example of the latter is a collaboration between the Roper Center for Public Opinion Research and the National Archives and Records Administration (NARA) to recover, preserve, document and make accessible public opinion survey data conducted on behalf of the United States Information Agency (USIA) Office of Research from 1952 through 1999.

The USIA data collection is estimated at over 2,000 surveys conducted in dozens of countries that contributed to the formulation of US foreign and defense policy. Some of the surveys represent the only opinion surveys available from specific countries. A typical study includes electronic data files containing coded responses of individuals to survey questions, questionnaires that provide a type of map to the data file contents, methodological and summary reports and assorted correspondence.

Substantial, but partial, collections of these survey files have been stored 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 housed at NARA, primarily consisting of select survey data from the early 1970s to 1999. Additionally, NARA has described 1,555 research reports from 1960 to 1982 in its archival research catalog and 675 series of related USIA records that are not necessarily survey related. The Roper Center maintains a subset of USIA survey materials, primarily focused on electronic datasets from 1952 to the early-1970s and 1990-1993.

By digitizing documents directly required for research data use and pointing users back to NARA for any additional resources, the project is able to leverage the relative strengths of each partner while creating a richer collection of USIA materials. NARA provides the structure for working with the State Department in the context of its mandate for preservation of federal electronic records; standards for appraising, cataloging and preserving electronic records; and permanent storage and file-level access for all materials related to the collection. NARA maintains the available additional USIA records in the form of reports, correspondence, and related federal government records. The Roper Center provides potential flexibility in communications and approach, supplementing the federal government agency-to-agency protocols; experience working with a variety of organizations to acquire data resources; active migration and management of data; more streamlined access to data-based materials; and access to related public opinion survey data from the private and non-federal public sector.

In the end it is hoped that such a collaboration will have benefits along a number of dimensions including improved user access, improved collection building and synergies resulting in further opportunities. Enhanced and richer access to multiple facets of the collection is the main advantage of the collaboration. Increased volume of metadata and availability of contemporary formats provide researchers with a much deeper resource for further exploration. As this core collection of survey data is solidified, richer and more complete extensions can be envisioned tying in supplementary contextual research materials. Finally, opportunities for various synergies such as coordinated acquisitions, processing and continually upgraded migration are anticipated.

As previously mentioned a key aspect in the relationships between data producers and data archives is trust; trust in the capabilities and integrity of the archivists and trust that the data will be securely stored and reliably preserved. That trust is also an integral part of the relationship between the archives and those who use the data. Those who use the data must be able to trust that the data are not only preserved, but will remain accessible over time.

The Data-PASS Shared Catalog

The Data-PASS shared catalog (see Figure 2) provides support for the partnerships cataloging, dissemination, and preservation activities. It is publicly available at:

<http://vdc.hmdc.harvard.edu/dataverse/DATAPASS/> . It provides three types of services:

First it facilitates discovery by providing a single access point from which patrons can search or browse all of the holdings collected specifically under the Data-PASS project, or the entire holdings of all of the partners. Both simple and fielded search is supported, along with browsing by subject, date, and source. (Search on variable-level descriptive information will be supported in the next release.)

Second, the shared catalog provides layered data extraction and analysis services on publicly-distributed data. Users who wish to access this public content can do so directly through the catalog interface, which supports extraction of data subsets, conversion to different statistical formats, and on-line data analysis. (Restricted content is discoverable through the catalog but accessible only directly from the home archive..)

Third, the shared catalog facilitates management of the collection by providing a single standard interface for harvesting via OAI-PMH. This interface is also used to support a single preservation mirror of the Data-PASS collected content, hosted at the Harvard-MIT Data Center.

Since the shared catalog combines information from several different sources, we designed it with an emphasis on provenance. The descriptive information for each study includes the chain of custody: the author, producer, and original distributor of the record. The descriptive information for each includes a link back to the study at the home archive, citations supplied by the archive, and a citation using the Altman-King [2007] data citation standard. This latter includes, where available, a Universal Numeric Fingerprint (UNF) which can be used to validate the data, even after reformatting. [see Altman, Gill, McDonald 2003]

Search Results for elections within Data-PASS

Review results below or enter another search term in the form at the bottom of the page to search across all collections or within a collection.

Sort results by [Author](#) [Date](#) [Title](#) Records per page: 100

- American National Election Study, 2004: Panel Study**
 Author University of Michigan. Center for Political Studies. American National Election Study
 Date 2006-02-17
- Current Population Survey, November 2004: Voter Supplement**
 Author United States Department of Commerce. Bureau of the Census
 Date 2006-01-16
- ABC News/WASHINGTON POST Monthly Poll, December 2004**
 Author ABC News;The Washington Post
 Date 2006-01-06
- ABC News Ohio State Poll, October 2004**
 Author ABC News
 Date 2006-01-06

Americans with Disabilities

Summary | Files | Credits | Cataloging Information

How to Cite This Dataset

Louis Harris and Associates, Inc., Harris Interactive, 1999, "Americans with Disabilities", [hdl:1902.29/H-828373](https://hdl.handle.net/1902.29/H-828373) ; Odum Institute [distributor(DDI)]

Data Distributor

Odum Institute

Abstract

This study focuses on disabilities. Topics addressed include type of disability, severity and impact, life satisfaction, marriage and dating, social impact, labor force incorporation, health insurance and other topics related to individuals with different employment status. Demographic data include sex, age, education, ideology, income, Hispanic origin, voting preference, and race.

Topic Classes

Selecting a Topic Class will launch a new search.

- [Harris National Studies](#) (ODUM:MAIN.HEADING)
- [Disabilities](#) (ODUM:INDEX.TERMS)
- [Disabilities](#) (ODUM:SUBJECT.TERMS)

Figure 2: Screenshots of the Shared Catalog Interface

The shared catalog is powered by the Virtual Data Center (VDC) 1.07 software [Altman, et. al 2001] (It will be migrated to its successor, *The Dataverse Network*, when that is publicly released.) The VDC is used to manage the content of the Murray and Odum archives, harvest the metadata from all archives into a central index. This metadata supports navigation and present of the catalog. The data analysis also provides “layered” on-line data formatting, extracting, and analysis, by dynamically retrieving data from each archive, processing it, and delivering it to end-users. (Advanced statistical analysis is provided through the R Statistical language [R Core Development team 2006] using interfaces developed to extend the Zelig [Imai, King, and Lau 2006] library.) A conceptual model of the catalog and related services is shown in Figure 3.

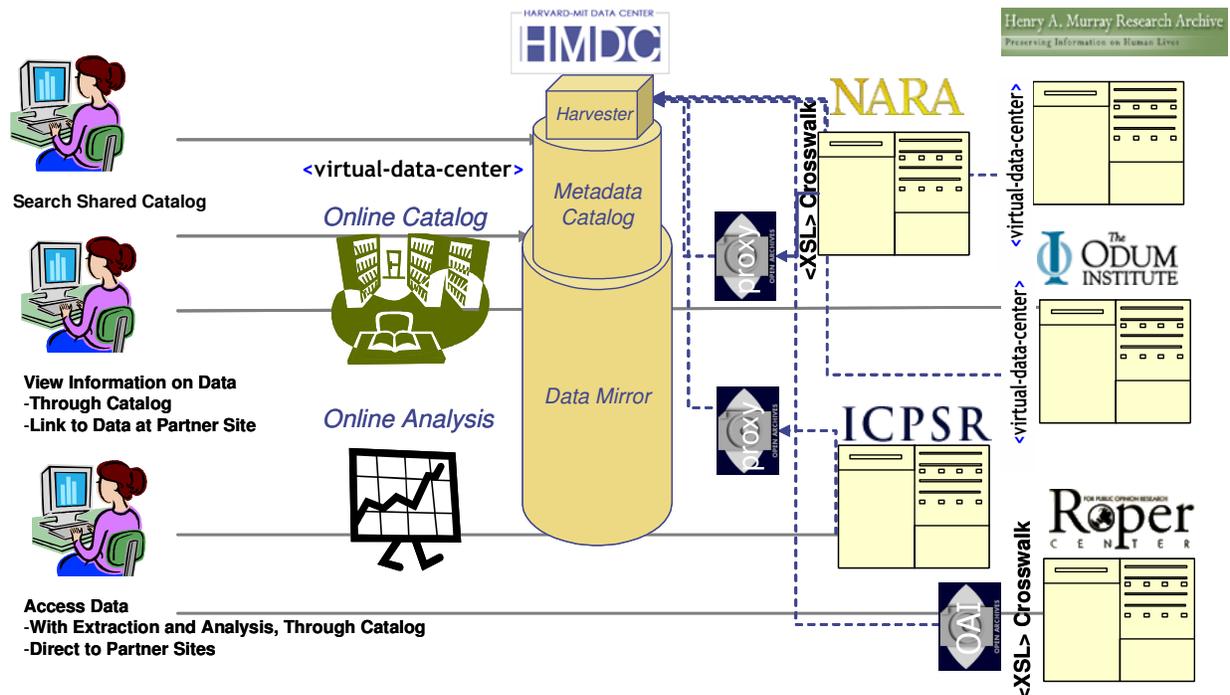


Figure 3: Model of the Data-PASS Shared Catalog

Metadata is naturally the linchpin of a common catalog. And the Data-PASS catalog builds upon shared practices for metadata content, organization, and exchange. Metadata supports many services, including: resource discovery, resource identification and citation, resource location, resource administration, data integrity, provenance, access control, and layered services such as variable level search, reformatting, and on-line analysis. We used the OAI-PMH protocol [Lagoze, et. al 2002] as an exchange mechanism, and identified a subset of the DDI-lite specification [see Blank & Rasmussen 2004] to format the metadata being exchanged. The full metadata standards are documented in detail on the Data-PASS web site.

The metadata *requirements* were intentionally made minimal. Each archive is required only to provide a title, identifier, data, and abstract for the study, along with a link to a corresponding catalog page hosted by that archive. However, most archives supply optional metadata, since we have identified fields that will enable more services to be provided by the catalog:

- Additional provenance information can be included, including logos to be displayed with each catalog record.
- Additional subject keywords can be provided to facilitate search and topic browsing
- File names, and links can be provided (along with MD5 or UNF's for validation), in order to enable download services.
- Variable level information can be provided, to support data analysis, extraction and (in the future) variable-level search.
- Usage terms can be supplied, if an archive wishes to allow public access to study files through the catalog, conditional on specific terms of use on access to the study. (These usage terms are then incorporated in an on-line click-through agreement to which patrons must agree in order to gain access to the restricted files.)

Since each organization followed its own practices internally, a significant part of establishing a shared catalog was to develop automated crosswalks between the metadata schema used internally by each archive and the DDI-lite schema. (These were typically implemented using XSLT.) Another significant step was to create proxy OAI servers that exposed the archive content through OAI for the archives that provided metadata only through other interfaces (such as FTP or HTTP, or other ad-hoc interfaces). The combination of metadata crosswalks and proxy OAI services creates a uniform interface for each archive, which enables the core of the shared catalog implementation to treat all member archives uniformly.

Future Research in Syndicated Storage

Data-PASS partners, as well as others, who archive social science data, have a need for syndicated storage that would assist them in their preservation activities, and we have begun to explore this area..This is the problem that we face: each archive has a unique collection. While there has in the past been some duplication of archival collections, current best practices are moving away from multiple unmanaged local copies. This is because of the need to ensure that research can be replicated using the exact data source that the original author used, and because the availability of on-demand web-based data distribution has reduced -- if not eliminated -- the need for local copies of widely-available resources. While a single point of responsibility for collections increases research replicability and reliability by reducing the possibility of versioning problems, they put the data at greater risk by reducing the number of copies and by putting those copies under a single institution's control. Moreover, an ever-increasing concern about preserving confidentiality makes strategies for storage, retrieval, and preservation all the more sensitive.

Current best practice is moving towards a more fully documented approach to data duplication, which includes maintaining consistent unique identifiers for each resource, and explicit metadata describing the resources, provenance, version, and associated rights. Best practice is moving towards more systematic and explicit duplication policies, including multiple mirroring of entire collections (rather than ad-hoc selections of individual items), and a process of regularly updating a mirror in order to preserve both the original and newer versions of a selected collection. A central issue that Data-PASS need to be solve is the asymmetrical nature of storage needs among the partners or potential partners. How do we construct systems that serve both the technology needs and the business needs for a given syndicate when some members may require an order of magnitude more storage than others? For example, ICPSR's distribution collection is about 300 gigabytes compressed and about 1.3 terabytes uncompressed, the Murray Archive's collection of audio and video is approximately 60 terabytes with compression. That might compare with a small archive that has a total collection of 10 to 50

gigabytes of data. We cannot easily ask the small collection to mirror all Data-PASS partners or even a single large archive. This will require negotiation and development of particular technological and institutional tools.

For syndicated storage technology to be effective, it must support the archival lifecycle. Syndicated storage must be compatible with the workflows for format migration, which is an essential and regular activity in social science data preservation. Syndicated storage solutions must also be designed to integrate with archival and inter-archival policies: The coverage, freshness, and correctness of redundant copies in the syndicated storage system will ideally be driven automatically by formal statements describing the desired archival relationships, and be fully auditable by all members of the partnership

Recent technologies such as LOCKSS such as LOCKSS [Reich & Rosenthal 2000], SRB [Moore, et. al ,2000] (and its imminent successor, IRODS [Rajasekar 2006]), and the emerging Distributed Data Manager now in development for incorporation in the Globus Toolkit [Foster 2006] may individually or in combination suitable as a base platform to build a service for the distributed preservation of social science data and documentation. Can these systems be adapted for managing asymmetrical collections? How tolerant are these system to human errors in archival management? To what extent do these systems provide for externally auditing for policy compliance? What can and should be incorporated into a schemas be developed that would accurately describe the policies governing inter-archival replication, and that can automatically coordinate the social science syndicated storage fabric?

These questions and others will need to be answered. What is clear at this point is that different technologies offer syndicated storage capabilities, but take diverging practical and theoretical approaches to replication and management, including differences in source licensing, cost of ownership, integration with digital library and computing grid protocols, scalability in size and number of replicas. Most important, these different storage technologies are designed under different philosophies regarding robustness, for example: what sorts of threat models are envisioned, whether it is necessary to protect against unintentional human error, and whether unilateral decisions by the archive holding the “master” copy.

We have begun to prototype the use of these systems in the context of social science data archiving, and to understand where the gaps between technology, policies and workflow are, and how to bridge these. In the coming year we plan to report on our findings in more detail.

Conclusions

We are in an age of unlimited digital resources. Data curators need skills and experiences to identify what can and should be preserved and what can not. Our goal is to ensure that the materials we include in our holdings remain accessible, complete and usable over time. The Data-PASS partnership continues to evolve, and to work closely with the social science research community in its search for classic data in need of archiving, potential partners, and new technologies in support of preservation.

The Data-PASS partnership continues to evolve, and to work closely with the social science research community in its search for classic data in need of archiving, potential partners, and new technologies in support of preservation. To learn more about the project or to recommend data for preservation, please visit our web site: <http://www.icpsr.org/DATAPASS/>

The future of digital curation will depend on collaborative efforts such as Data-PASS. Within disciplines, collaborative efforts should occur among researchers, and between researchers and archivists, curators and other information specialists to ensure that the data collections are available and usable. Collaborative efforts among archives, and between archives and individual researchers or private research organizations, can provide opportunities to learn from each others experiences and provide fresh perspectives and ways to deal with challenges that we all face.

Acknowledgments

This project was supported by an award (PA#NDP03-1) from the Library of Congress through its National Digital Information Infrastructure and Preservation Program (NDIIPP).

References

- 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):458-71.
- Altman, M., McDonald, M., & Gill, J. (2003). *Numerical Issues in Statistical Computing for the Social Scientist*, John Wiley & Sons: New York.
- Altman, M., & King, G. (2007). "A Proposed Standard for the Scholarly Citation of Quantitative Data", *D-Lib* 13(3/4).
- Crabtree, J. and Donakowski, D. (2006). *Building Relationships: A Foundation for Digital Archives*. Paper presented at JCDL 2006 Workshop: Digital Curation & Trusted Repositories: Seeking Success. Retrieved March 20, 2007 from <http://sil.unc.edu/events/2006jcdl/digitalcuration/Crabtree-JCDLWorkshop2006.pdf>.
- Data-PASS. Web site. <http://www.icpsr.umich.edu/DATAPASS/> (03/14/07)
- Foster, I. (2006). "Globus Toolkit Version 4: Software for Service-Oriented Systems". *IFIP International Conference on Network and Parallel Computing*, Springer-Verlag LNCS 3779, pp 2-13.
- Imai, K., King, G., & Lau, O. (2006). Zelig: Everyone's Statistical Software. R package version 2.7-4.
- Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting - Version 2.0. <<http://www.openarchives.org/OAI/openarchivesprotocol.html>>
- Moore, R. W., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., & Gupta, A. (2000). "Collection-Based Persistent Digital Archives", *D-Lib Magazine* 6 (3,4).
- National Science Foundation (NSF). (2004). Grant Proposal Guide, NSF 04-23. Retrieved March 20, 2007 from http://www.nsf.gov/pubs/gpg/nsf04_23/.

National Institutes of Health (NIH). (2003). Final NIH Statement on Sharing Research Data. Retrieved March 20, 2007 from <http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>

NDIIPP. Web site. <http://www.digitalpreservation.gov/about/index.html> (03/14/07)

Rajasekar, A., Wan, M., Moore, R. W., & Schroeder, W. (2006). "A Prototype Rule-based Distributed Data Management System", *HPDC workshop on Next Generation Distributed Data Management*.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Reich, V. & Rosenthal, D.S. (2000). "LOCKSS (Lots Of Copies Keep Stuff Safe)", *Preservation 2000, The New Review of Academic Librarianship 6: 155- 161*.

License

You may use this work under the:

Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License
(see <http://creativecommons.org/licenses/by-nc-nd/3.0/us/>)