Digital Preservation through Archival Collaboration: The Data Preservation Alliance for the Social Sciences

Micah Altman, Margaret O. Adams, Jonathan Crabtree, Darrell Donakowski, Marc Maynard, Amy Pienta, and Copeland H. Young

Abstract

The Data Preservation Alliance for the Social Sciences (Data-PASS) is a partnership of five major U.S. institutions with a strong focus on archiving social science research. The Library of Congress supports the partnership through its National Digital Information Infrastructure and Preservation Program (NDIIPP). The goal of Data-PASS is to acquire and preserve data from opinion polls, voting records, large-scale surveys, and other social science studies at risk of being lost to the research community. This paper discusses the agreements, processes, and infrastructure that provide a foundation for the collaboration.

About the Partnership

An international movement to archive, preserve, and share data emerged over forty years ago when digital data began to appear in volume.[1] This movement is undergoing a resurgence as the social sciences shift anew toward a reliance on vast amounts of digital data. 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, nor that newly created data will be preserved.

[1] For a history of the early development of this community, see Margaret O. Adams, "The Origins and Early Years of IASSIST," *IASSIST Quarterly* 30, no. 3 (2006): 5–15.

Why is this so? Many corporate and academic researchers assume that data they generate are their property and that they have limited obligations to share it with others or to ensure its preservation. Some individual researchers are reluctant to deposit their data in archives because they fear competition. Some lack the time or expertise to prepare the metadata required for effective sharing. And some simply do not recognize the long-term value of their data. Institutional data producers may be under legal obligation to protect proprietary information. And some data just falls through the cracks.

A huge quantity of digital social science research content lives on, for the moment, solely as files in the computers of individual researchers or of research institutions, or quite possibly as videotapes, floppy disks, or punchcards (etc.) in bookcases, libraries, and warehouses. If research sponsors, producers, and data curators do not take steps to preserve it, it will be lost forever.[2] It needs to be identified, located, assessed, acquired, processed, preserved, and shared.

Five major American social science data archives have created the Data Preservation Alliance for the Social Sciences (Data-PASS) to ensure the long-term preservation of our holdings and of materials as yet unarchived.[3] The partners are 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 Electronic Records Custodial Division of the National Archives and Records Administration (NARA), and the Henry A. Murray Research Archive, with strong technology support from the Harvard-MIT Data Center.[4] We seek to acquire and preserve data at risk of being lost to

---

[2] The members of this partnership represent the U.S. social science data archives tradition. There are other emerging approaches to preservation, including "self"-archiving, institutional archiving, and, more recently, virtual archiving. See T. A. Peters, "Digital Repositories: Individual, Discipline-based, Institutional, Consortial, or National?," *Journal of Academic Librarianship* 28, no. 6 (2001): 414–17. For a discussion of virtual archiving by one of the partners, see Micah Altman, "Transformative Effects of NDIIPP, the Case of the Henry A. Murray Archive," *Library Trends* 57, no. 3 (forthcoming).

These are important trends, and our collection policy recognizes any collection to which a longstanding institution has made a long-term preservation commitment as not "at risk." However, it is important to note that, as one recent study concludes, "faculty output is not finding its way into institutional repositories in the U.S. in large numbers, except at some of the largest, most research-intensive universities." See C. S. McDowell, "Evaluating Institutional Repository Deployment in American Academe Since Early 2005: Repositories by the Numbers, Part 2," *D-lib Magazine* 13, nos. 9–10 (2007). Furthermore, McDowell shows that, currently, most institutional repositories focus on print-related materials and do not have significant holdings of more complex (and less readily human interpretable) digital objects such as numeric data. Nor, in our own experience, do most of these repositories make full preservation commitments to preserve quantitative data resources. See also Elizabeth Yakel, Soo Young Rieh, Karen Markey, and Jihyun Kim, "Institutional Repositories and the Institutional Repository: College and University Archives and Special Collections in an Era of Change," *American Archivist* 71 (Fall/Winter 2008): 323–49.

[3] 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/presentations.html. The shared catalog is available from http://dvn.iq.harvard.edu/dvn/dv/datapass. All URLs accessed 1 August 2008.

[4] Both the Harvard-MIT Data Center and the Henry A. Murray Research Archive are now part of the Institute for Quantitative Social Science in the Faculty of Arts and Sciences at Harvard University.

the research community, from opinion polls, voting records, large-scale surveys, and other social science studies. While our organizations have a history of collaboration, this official partnership provides important benefits and has 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's National Digital Information Infrastructure and Preservation Program (NDIIPP). The NDIIPP mission is to develop a national strategy to collect, archive, and preserve digital content, especially materials created in digital format. Our partnership works 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.

The partnership has succeeded in many areas. We identified thousands of at-risk research studies (with the help of the larger data archiving community, which contributed significant leads) and acquired hundreds of them.[5] These range from data collections created under National Science Foundation (NSF) and National Institutes of Health (NIH) grants,[6] to surveys conducted by private research organizations, to state-level polling data, to data records created by governmental research or administrative programs. This even included data from a government agency: As part of the Data-PASS partnership, the National Archives and the Roper Center collaborated to build mutually comprehensive collections of worldwide survey data collected by the U.S. Information Agency (USIA) from the 1950s through 1999. Every day more materials are identified, acquired, and processed. In the course of the partnership, we built a network of relationships among data archives, data producers, research funders, and data users. Underlying this outreach, identification, and acquisition effort was the establishment of an agreed-upon set of best practices for sustainable digital preservation of research data.

We have also established a shared electronic catalog for the tens of thousands of studies or series that comprise each partner's entire data holdings. The Data-PASS shared catalog creates, for the first time ever, a unified way to find social science digital data in major U.S. archives, completing the unification of social science data, a major goal of data archivists since the first Council of Social Science Data Archives in the 1960s.[7] It supports searching and browsing information on the

---

[5] We discuss the selection process below, in the sections describing our coordinated operations and best practices.

[6] Both NIH and NSF require most research data to be shared, although this requirement is rarely enforced. See, respectively, National Science Foundation (NSF), Grant Proposal Guide, NSF 04-23 (2004), http://www.nsf.gov/pubs/gpg/nsf04_23/, accessed 1 August 2008; National Institutes of Health (NIH), Final NIH Statement on Sharing Research Data (2003), available at http://grants.nih.gov/grants/guide/notice-files/ NOT-OD-03-032.html, accessed 1 August 2008.

[7] William A. Glaser and Ralph L. Bisco, "Plans of the Council of Social Science Data Archives," *Social Science Information* 5, no. 4 (1966): 71–96.

D I G I T A L   P R E S E R V A T I O N   t h r o u g h   A R C H I V A L
C O L L A B O R A T I O N :   T h e   D A T A   P R E S E R V A T I O N
A L L I A N C E   f o r   t h e   S O C I A L   S C I E N C E S

collections of the partners in their entirety, and it showcases data obtained directly through the partnership. The catalog provides automated interfaces so that users of other catalog systems can find data in it. The catalog also provides a single virtual collection with comprehensive content and published interfaces, which is used as a platform for additional services such as replication, discovery, and analysis. Anyone who wishes to access the data files preserved as a result of the Data-PASS partnership, as well as many of the other files described in the catalog, can do so directly through the catalog interface. The interface supports extraction of data subsets, conversion of the files to different statistical formats, and online data analysis.

The partnership succeeded in large part by engaging in three sorts of collaboration:

- loosely coupled coordination of operations,
- joint development of best practices, and
- creation of a open-source shared infrastructure.

Underlying these three *operational* areas of collaboration is a joint *strategic* agreement. In the remainder of this article we discuss this strategic agreement, these three areas of collaboration, and plans for future development.[8]

Strategic Commitments: The Data-PASS Articles of Collaboration

The concrete operational and infrastructural aspects of the collaboration are built upon a shared commitment to data preservation and institutional buy-in for collaboration. The partnership's articles of collaboration articulate these commitments to preservation and to cooperation.[9] Specifically, the articles express the partners' individual and collective commitment to ensuring that materials collected by Data-PASS remain "accessible, complete, uncorrupted, and usable over time." The articles also establish shared rights to material collected by and software developed by the partnership, and they articulate the partners' institutional commitments for three primary areas of concrete collaboration:

- Coordination of operational identification, acquisition, and cataloging activities.
- Development of best practices.
- Participation in a shared catalog and replication infrastructure.

We turn now to each of these areas of collaboration.

---

[8] For a description of Data-PASS collection development and its challenges, see Myron Gutmann, Mark Abrahamson, Margaret O. Adams, Micah Altman, Caroline Arms, Kenneth Bollen, Michael Carlson, Jonathan Crabtree, Darrell Donakowski, Gary King, Jared Lyle, Marc Maynard, Amy Pienta, Richard Rockwell, Lois Timms-Ferrara, and Copeland Young, "From Preserving the Past to Preserving the Future: The Data-PASS Project and the Challenges of Preserving Digital Social Science Data, *Library Trend*s (forthcoming), which also includes a description of how Data-PASS affected the transformation of the Murray Archive.

[9] Available from the Data-PASS website: http://www.icpsr.org/DATAPASS, accessed 1 August 2008.

Coordinated Operations

Over the first three years of the project, Data-PASS focused its collection activity on identifying and acquiring digital social science data that is "at risk," and that has had, or can be expected to have, significant influence over social science findings and public policy. The potential volume of unreclaimed social science data that could be acquired, and the need to make the most cost-effective use of limited resources, has led us to establish a coordinated approach to identification, appraisal, and processing.

The identification and selection process is somewhat decentralized, yet coordinated. Each archives independently seeks to identify data that could be acquired by the academic members of the partnership. Each partner pursues the materials that best represent its community of stakeholders and area of specialization (e.g., with respect to subject content, source, or research design). This decentralization allows each partner to leverage its distinct capabilities in specific kinds and sources of data.

A coordinated evaluation process balances decentralized search and identification activities: Each academic partner records all potential acquisitions in a shared (internal) database, and representatives from the partnership meet bimonthly (often using a teleconference) to review the newly identified studies and prioritize them for collection by the partnership. If identified content originated with a federal agency, the partners determine whether the data are already preserved at NARA or scheduled to be transferred to NARA.

Best Practices

To ensure consistency throughout this process, we developed several sets of best practices, including common criteria for content selection that guide decisions on whether data falls within the overall collection mission; appraisal guidelines to aid in prioritizing these acquisitions; and processing guidelines for making these acquisitions. All of these practices are written by the operations committee, approved by the partnership steering committee, and published on the partnership website.[10]

The content selection criteria start from the premise that any social science data not currently managed by a permanent archives is considered to be at risk of loss. Substantive criteria for selection include whether the materials support studies that are highly cited; produced by high-impact researchers; theoretically or methodologically innovative; based on a national sample; targeted a special population; part of a major policy evaluation or decision; or describe rare events.

---

[10] All of these, and other practices described below, are available on the project website.

The appraisal criteria incorporate elements of accepted archival practice in identifying the most important content to preserve and evaluating the risk of losing the content should acquisition not take place. The appraisal guidelines include assessing the significance of the data to the research community, the significance of the source and context of the data, how the materials would complement existing collections, the uniqueness of the data, the data's potential usability, and the anticipated cost of processing.

Studies identified as "high priority" are then pursued by the most appropriate partner (usually the partner that identified the study) for acquisition and processing. For NARA, this included collaborating with colleagues in NARA's Life Cycle Management Division to target disposition authorities for electronic records due for transfer from federal agencies. The details of processing at each repository differ, but always include verifying the content of the materials, preparing an inventory, performing a basic review for confidentiality, and creating required catalog metadata. For fragile materials, we developed an additional set of specific guidelines covering manner of inventory, physical handling, backup, and transportation.

Finally, we established common best practices for retention of the data. These include physical and electronic security, validation of random samples of material against Universal Numeric Fingerprints[11] (UNFs) and cryptographic hashes, guided format migration, and replication of holdings.

Shared Infrastructure: The Data-PASS Catalog

The Data-PASS shared catalog provides essential infrastructure for the partnership's cataloging, dissemination, and preservation activities. (See Figure 1.) It is publicly available and linked from the partnership website.[12] The shared catalog supports three general categories of services.

---

[11] Universal Numeric Fingerprints, or UNFs, are semantic fingerprints that uniquely identify and validate datasets. The general algorithm for this was first published along with a sample implementation in Micah Altman, Michael P. McDonald, and Jeff Gill, *Numerical Issues in Statistical Computing for the Social Scientist* (New York: John Wiley and Sons, 2004). Different versions of the algorithm have been developed since its initial publication. Version 3 was the first to be implemented in publicly available software: The UNF package for R was made available through the Comprehensive R Archive Network (CRAN, a code archive developed as part of the R Project). This version of the software also introduced the name "Universal Numeric Fingerprint." Version 5 is the current version addressing security vulnerabilities and adding time, date, and other types, see Micah Altman, "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 (Dordrecht: Springer-Verlag, 2008). As discussed below, the UNF was later incorporated in the proposed Altman-King citation standard for data: Micah Altman and Gary King, "A Proposed Standard for the Scholarly Citation of Quantitative Data," *D-Lib Magazine* 13, nos. 3–4 (2007).

[12] The shared catalog can also be accessed directly at http://dvn.iq.harvard.edu/dvn/dv/datapass, accessed 1 February 2009.
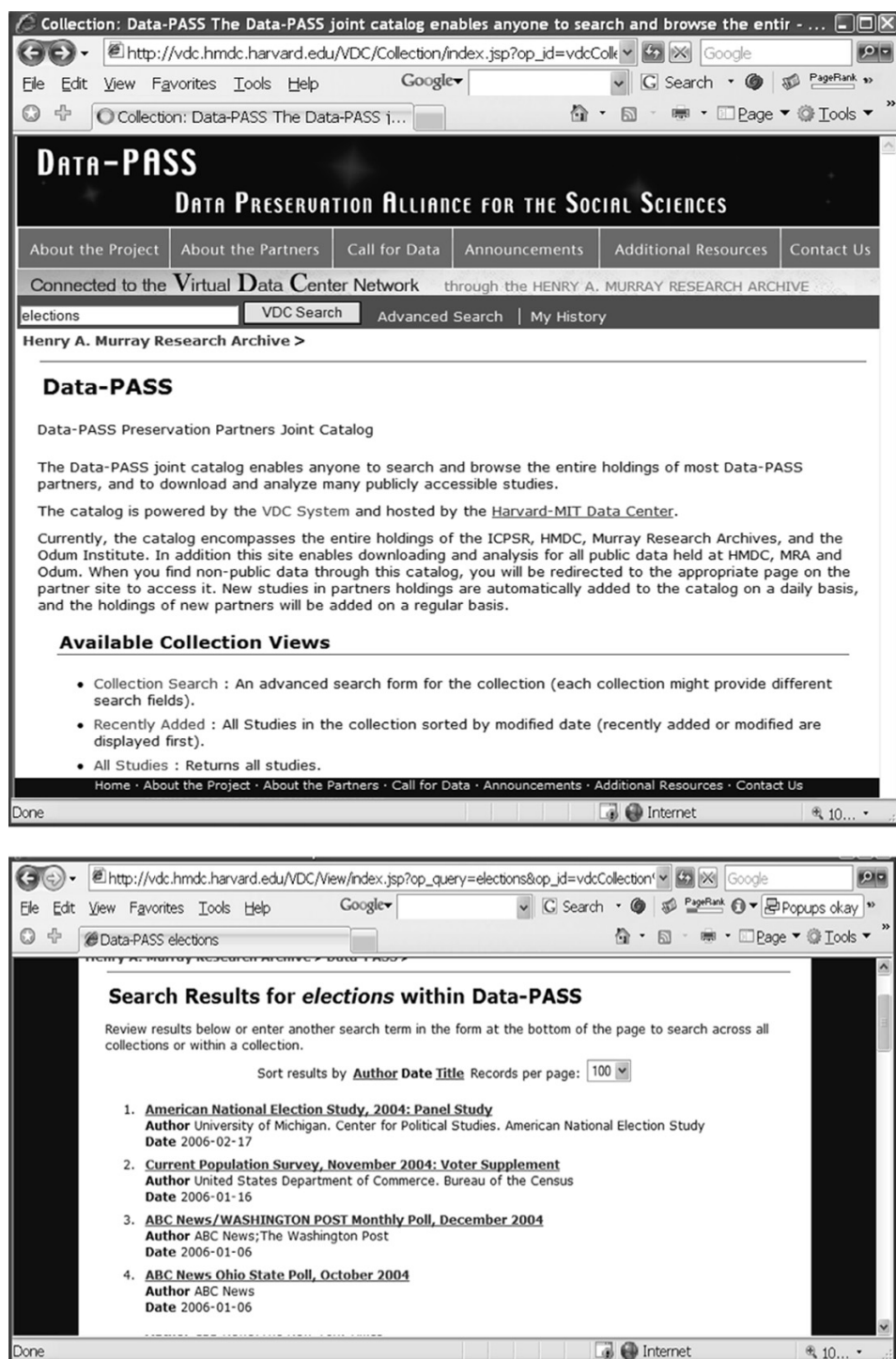
F I G U R E   1 .   **Screenshots of the shared catalog interface. (continued on next page)**
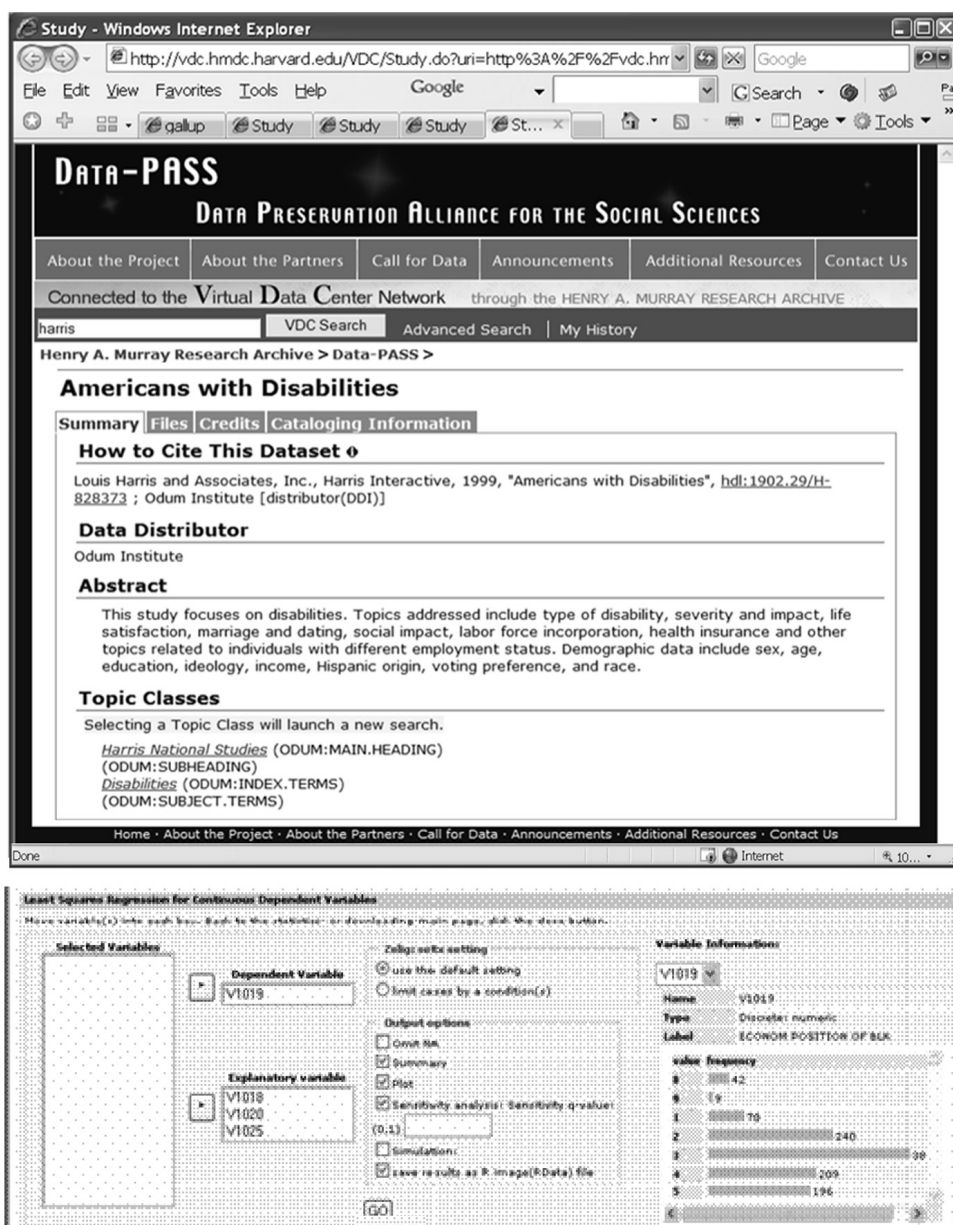
FIGURE 1. (continued).

First, the catalog facilitates *discovery,* since it provides a single access point from which patrons can search or browse all of the holdings collected specifically under the Data-PASS partnership, or descriptions of the entire holdings of all of the partners. Both simple and fielded search of descriptive study and variable-level metadata is supported, along with browsing by subject, date, and archival source.

Second, the catalog provides layered *data extraction and analysis* services for a selection of 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 online data analysis. Some content, including all of that preserved by the National Archives and restricted content from the other partners, is discoverable through the catalog but accessible only from the home archives. For such studies, the catalog provides a direct link to the study or series description within the native catalog of the partner responsible for it.

Third, the catalog facilitates *management of the collection* by providing a single standard interface for harvesting, via the widely used OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) method.[13] This interface was also used to support a prototype preservation mirror of the Data-PASS collected content, hosted at the Harvard-MIT Data Center.[14] Since the shared catalog combines information from several different sources, we emphasized provenance in its design. The descriptive information for each study includes information about every stage of authorship and curation, including the author, producer, and original distributor of the record. The descriptive information for each includes a link back to the study at the home archives, citations supplied by the archives,[15] and a citation using the Altman-King data citation standard.[16] The latter includes, where available, a UNF that can be used to validate the data, even after reformatting.

The shared catalog was developed using the Virtual Data Center (VDC) software[17] and migrated to its successor, the Dataverse Network (DVN)[18] at the end of 2007. The DVN is also used to manage the content of the Murray and Odum archives and to harvest the metadata from all archives into a central index, while other archives use their own systems[19] to manage their own content, sharing it with the main catalog via metadata harvesting.

[13] Carl Lagoze, Herbert Van de Sompel, M. Nelson, and S. Warner, "The Open Archives Initiative Protocol for Metadata Harvesting—Version 2.0," 2002, available at http://www.openarchives.org/OAI/ openarchivesprotocol.html, accessed 1 August 2008.

[14] At the end of the initial award period, a copy of this preservation mirror was delivered to the Library of Congress for preservation there.

[15] In this article, following SAA nomenclature, we use "archives" to denote a *single* archival institution.

[16] Altman and King, "A Proposed Standard."

[17] See Micah Altman, Leonid Andreev, Mark Diggory, Michael Krot, Gary King, Daniel Kiskis, Akio Sone, and Sidney Verba, "A Digital Library for the Dissemination and Replication of Quantitative Social Science Research," *Social Science Computer Review* 19, no. 4 (2001): 458–71.

[18] Gary King, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," *Sociological Methods and Research* 32, no. 2 (2007): 173–99.

[19] Roper, NARA, and ICPSR use custom systems, developed in-house, to manage their holdings. NARA envisions moving all of its archival management processes and storage to the ERA system, in development. See Kenneth Thibodeau, "Building the Archives of the Future: Advances in Preserving Electronic Records at the National Archives and Records Administration," *D-lib Magazine* 7, no. 2 (2001).

This metadata supports navigation and presentation of the catalog. The DVN also provides the "layered" online data formatting, extracting, and analysis mentioned above, by dynamically retrieving data from each of the archives, processing it, and delivering it to end-users. Advanced statistical analysis is provided through the R Statistical language using interfaces developed to extend the Zelig library.[20]

A conceptual model of the catalog and related services is shown in Figure 2. 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 online analysis. These "layered" services are provided for the user dynamically by the catalog, without having to reprocess data or install new software at the source archives. We used the OAI-PMH protocol as an exchange mechanism and identified a subset of the Data Documentation Initiative's DDI-lite specification to format the metadata being exchanged.[21] The full metadata requirements are documented in detail on the Data-PASS website.

We intentionally made the metadata *requirements* minimal. Each of the archives is required to provide only a title, a permanent and unique identifier, and an abstract for the study, along with a link to a corresponding catalog page hosted by that institution. However, most archives supply additional metadata, since this enables the catalog to provide increased levels of service; for example:

- Adding additional provenance information enhances branding; for example, if logos are included in the metadata, they will be displayed with each catalog record.
- Adding keywords facilitates searching and browsing, and makes it more likely that a user will find that resource.
- Adding file names and links enables integrated download services and facilitates replication of the data for preservation.
- Adding UNFs or a standard cryptographic hash such as MD5 enhances integrity.
- Adding variable-level information enables the catalog to support data analysis, extraction, and variable-level search.

[20] Kosuke Imai, Gary King, and Olivia Lau, *Zelig: Everyone's Statistical Software.* R package version 2.7-4 (2006).

[21] OAI-PMH is an open-protocol used to expose the contents of digital library catalogs. See Lagoze et al., "The Open Archives Initiative Protocol." The Data Documentation Initiative (DDI) schema is a metadata specification designed specifically for documenting social science research data. See Grant Blank and Karsten Boye Rasmussen, "The Data Documentation Initiative. The Value and Significance of a Worldwide Standard," *Social Science Computer Review* 22, no. 3 (2004): 307–18.

F I G U R E   2 .   Model of the Data-PASS shared catalog.
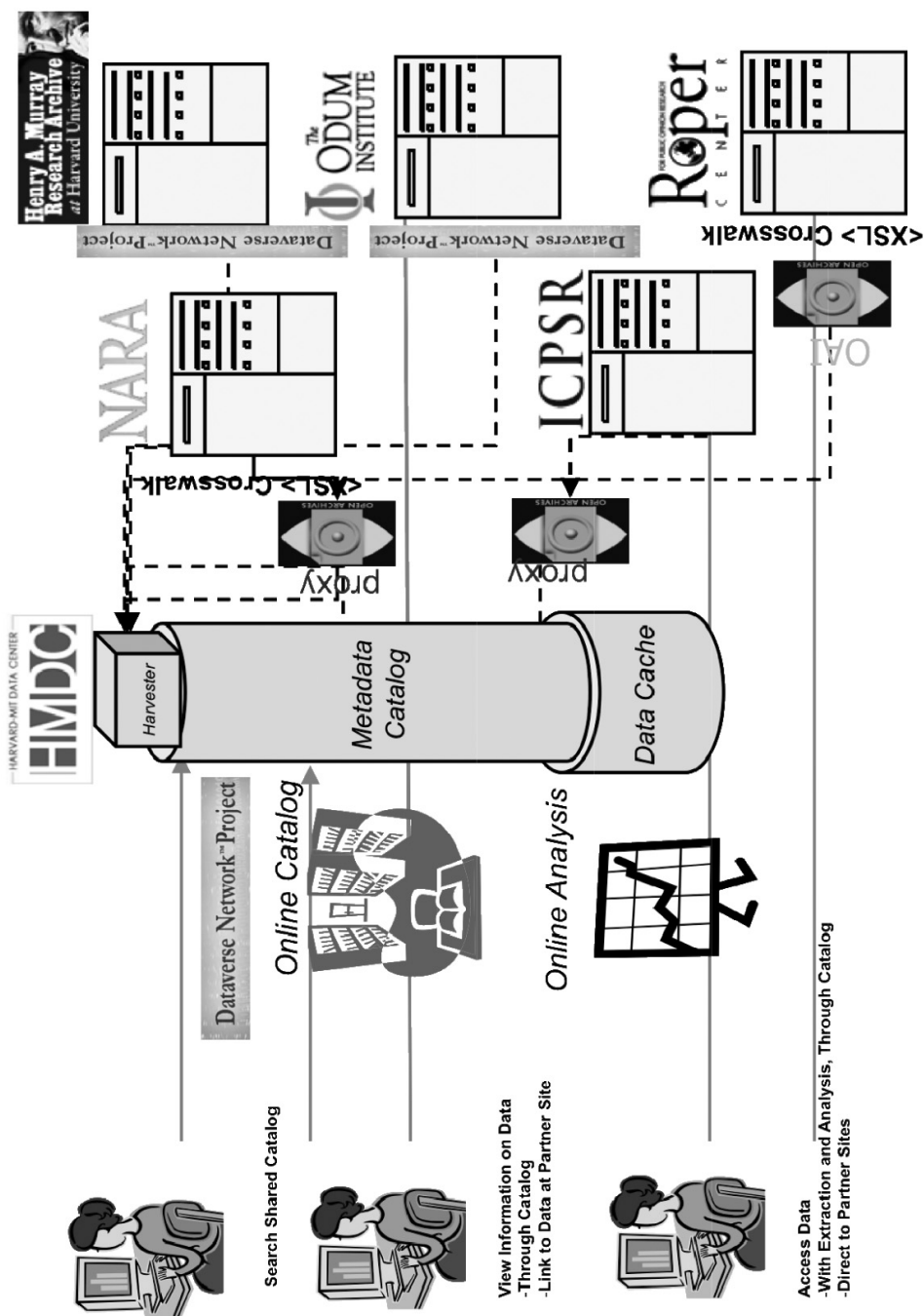
- Adding archives-specific usage terms facilitates conditional access to study files through the catalog. These usage terms are then incorporated into an online click-through agreement that the catalog presents to patrons (and to which they must agree to gain access to the restricted files).

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 of the archives and the DDI-lite schema used for exchange. These crosswalks were typically implemented in actionable form using XSLT (eXtensible Stylesheet Language Tranformations) for metadata sources already in XML (eXtensible Markup Language) form, and in Perl script for metadata in other forms. Another significant step was to create proxy OAI servers that exposed content through OAI for the archives that provided metadata only through other interfaces (such as FTP, HTTP, or other ad-hoc interfaces). The combination of minimal requirements, actionable metadata crosswalks, and proxy OAI services creates a uniform interface for each of the archives, which enables the core of the shared catalog implementation to treat all member archives as standardized sources.[22]

Future Research in Syndicated Storage

There are many threats to archived digital information. *Physical* threats result from chance, natural events, or age, and include failures in media, hardware, storage facilities, and so forth. *Technological* threats include format obsolescence and destructive software errors. *Human* threats include curatorial error, and insider and outsider attacks. *Institutional* threats include mission change, change of legal regime, or economic failure.

Many of these threats are ameliorated through replication of the materials to be preserved, combined with regular auditing.[23] When a set of institutions replicates their holdings and each holds these "in trust" for the others, the risk of preservation failure is greatly diminished. This is especially true when the institutions are diversified with respect to the legal regimes and economic models under which they operate and the technical preservation strategies that they employ. Current best practice is moving toward a systematic approach to data replication, which includes maintaining consistent unique identifiers for each resource; explicit metadata describing the resources, provenance, version, and associated rights; and a managed set of replication services. Best practice is moving toward more systematic and explicit replication policies that include multiply replicating

---

[22] These proxy OAI servers were initially implemented as custom wrappers and are now implemented as "harvested dataverses" in the IQSS dataverse network, using features built into the Dataverse Network.

[23] David S. Rosenthal, Thomas Robertson, Tom Lipkin, Vicky Reich, and Seth Morabito, "Requirements for Digital Preservation: A Bottom-Up Approach," *D-Lib Magazine* 11, no. 11 (2005).

entire collections off-site, explicit versioning, and a process of regularly refreshing and verifying replicated content. In a separately funded follow-on phase of the Library of Congress award scheduled to be completed during 2009, the academic partners will prototype a policy-driven replication service for the partnership.

Data-PASS partners, as well as others who archive social science data, are searching for a "syndicated storage" layer that would assist them in such preservation-oriented replication. This system will serve two institutional goals. First, it will help each institution ensure against media, software, hardware, and physical failure, since geographically distributed partners will keep separate "back-up" copies. Second, it will help the partnership ensure against institutional failure, since, if a partner should suffer institutional failure, the partnership as a whole will still retain copies of the holdings of failed partners and will be able to redistribute them.

For syndicated storage technology to be effective, it must support the archival life-cycle. Syndicated storage solutions must also be designed to support and integrate smoothly with intra-archival and interarchival policies. Our primary goal for the behavior of the syndicated storage system is that archival policies govern it directly—this should include systematizing the commitment of resources each of the archives has made to preserving the contents of the other partners, the auditing commitments each has made to its depositors, and the legal policies supporting access to the data by other partners in the case of institutional failure.

Another institutional issue we plan to address is the asymmetrical nature of storage needs among current and potential partners. How do we construct systems that serve both the technology needs and the business needs of a collective when some members may require an order of magnitude more storage than others? For example, ICPSR's distribution data collection is about 300 gigabytes compressed and about 1.3 terabytes uncompressed, and the Murray Archive's collection of digital audio and video is approximately 60 terabytes with compression. In comparison, a small institution may have a total collection of 10 to 50 gigabytes of data. We cannot easily ask the small institution to mirror all Data-PASS partners or even a single large archives. Instead, we are prototyping a replication system designed to function with such asymmetries.

We plan to develop a formal schema that will precisely describe these policy commitments. These formalized commitments are likely to include storage resource commitments from and to each partner; details of the replication policy to be applied, such as freshness, number of copies, and versioning; and audit and verification requirements.

We will then develop a set of software tools that translates this set of commitments into a set of replication and verification actions, to be performed on top of an existing distributed storage platform. Recently developed distributed technologies such as LOCKSS, SRB and its developing successor, IRODS, and other emerging systems are examples of distributed storage technologies that provide a suitable base platform on which to build a service for the

distributed preservation of social science data and documentation.[24] One of these systems, along with the software being developed by the academic partners, would constitute an internal storage layer. The archives' systems, rather than end-users, would access it to manage replication of digital objects.

Can these systems be adapted for managing asymmetrical collections? How tolerant are these systems to human errors in archival management? To what extent do these systems provide for external auditing for policy compliance? What can and should be incorporated into a schema to describe accurately the policies governing interarchival replication and to coordinate automatically the social science syndicated storage fabric?

These questions and others will need to be answered. Clear at this point is that different technologies offer syndicated storage capabilities, but take divergent practical and theoretical approaches to replication and management. They include 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, there are differences in the sorts of threat models envisioned, whether it is necessary to protect against unintentional human error, and whether unilateral decisions by the institution holding the "master" copy are permissible. We have begun to prototype the use of these systems in the context of social science data archiving. When the system is complete and has been operating long enough to observe its stable behavior, we will be able to report on our findings in more detail.

Future Collaboration

Truly useful collaborations are often difficult to start, but easy to maintain. The first phase of the partnership was started with a substantial infusion of grant funding to the academic partners (NARA's effort was entirely contributed), which helped to provide the staff time and focus to initiate collaborative activities. This phase and its funding are over, although the Library of Congress generously funded a much smaller follow-on grant, primarily to develop the replicated storage solution described above.

---

[24] LOCKSS (Lots of Copies Keeps Stuff Safe) is a project to produce software for peer-to-peer archival replication of online digital resources that are primarily print related or narrative in nature. See Victoria Reich and David S. Rosenthal, "LOCKSS (Lots Of Copies Keep Stuff Safe)," *Preservation 2000, The New Review of Academic Librarianship* 6 (2000): 155–61. SRB (storage resource broker), IRODS (i Rule Oriented Data Systems), and DDM (Distributed Data Manager) are systems designed for use in high-performance computing systems as replicated storage. See, respectively, Reagan Moore, Chaitan Baru, Arcot Rajasekar, Bertram Ludaescher, Richard Marciano, Michael Wan, Wayne Schroeder, and Amarnath Gupta, "Collection-Based Persistent Digital Archives," *D-Lib Magazine* 6, nos. 3 and 4 (2000). For a review of storage grid technologies, see Srikumar Venugopal, Rajkumar Buyya, and Kotagiri Ramamoanarao, "A Taxonomy of Data Grid for Distributed Data Sharing, Management, and Processing," *ACM Computing Surveys* 38, no. 1 (2006): 1–53.

Funding notwithstanding, we have developed channels of communication, harmonized practices, and created joint infrastructure that make ongoing activities easier to continue. Most of our communication is virtual: through email, telephone, and wikis. We have dramatically decreased the marginal costs of sharing collections acquired in the future by harmonizing our metadata and deposit agreements. We have also decreased the marginal cost of further shared infrastructure creation by developing an open-source base and building on other open-source tools—anyone can contribute. These actions make the costs of continuing collaborative activities relatively low.

Moreover, the partnership has demonstrated to us the benefits of collaborative activity. The things that the partnership provides, such as a shared catalog, recognized best practices, and expert review of acquisitions, serve the core mission of each individual institution. Through collaboration, we are able to pool expertise and (virtual) collections, facilitating broader and deeper solutions to archival problems. So too, in future, each participating institution will benefit from replication of holdings for disaster recovery. This creates strong incentives to continue to collaborate, particular grants notwithstanding. We plan to continue and expand the partnership.

Conclusions

We believe the future of digital curation will depend on collaborative efforts such as Data-PASS. The Data-PASS partnership demonstrates how such collaboration can be effective in identifying, acquiring, and preserving digital material. Through three years of partnership, we have acquired hundreds of at-risk research studies, identified thousands more, and fulfilled a long-term vision in the social sciences of providing a unified catalog of U.S. social science data.

The collaboration has taken part largely in three areas. First, we regularly coordinated our loosely coupled operations, particularly in the areas of identification and selection of materials for acquisition. Second, we established joint best practices, especially in the area of appraisal and metadata exchange. Third, we created a shared catalog infrastructure. Strategic articles of collaboration that articulate commitments to coordinating practices, to participating in a shared catalog and replication infrastructure, and to sharing rights in the nonfederal government materials acquired and technologies developed by the partnership activities made all of these areas of collaboration possible.

The 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 partnership or to contribute to it, please visit our website: http://www.data-pass.org.