

**Transformative Effects of NDIIPP, the case of the  
Henry A. Murray Archive**

Micah Altman<sup>i</sup>

Harvard University

Institute for Quantitative Social Science

Cambridge, MA 02138

617-496-3847

Micah\_Altman@harvard.edu

Paper prepared for Library Trends Special Issue about the NDIIPP  
Program

July, 2008

**Abstract:**

This article comprises reflections on the changes to the Henry A. Murray Research Archive, catalyzed by involvement with the NDIIPP partnership, and the accompanying introduction of next-generation digital library software.

Founded in 1976 at Radcliffe, the Henry A. Murray Research Archive is the endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science, in Harvard University. The Murray preserves in perpetuity all types of data of interest to the research community, including numerical, video, audio, interview notes, and other types. The Center is unique among data archives in the United States in the extent of its holdings in quantitative, qualitative, and mixed quantitative-qualitative research.

The Murray took part in an NDIIP-funded collaboration with four other archival partners, Data-PASS, for the purpose of the identification and acquisition of data at risk, and the joint development of best practices with respect to shared stewardship, preservation, and exchange of this data. During this time, the Dataverse Network (DVN) software was introduced, facilitating the creation of virtual archives. The combination of institutional collaboration and new technology lead the Murray to re-engineer its entire acquisition process; completely rewrite its ingest, dissemination, and other licensing agreements; and adopt a new model for ingest, discovery, access, and presentation of its collections.

Through the Data-PASS project, the Murray has acquired a number of important data collections. The resulting changes within the Murray have been dramatic, including increasing its overall rate of acquisitions by four-fold; and disseminating acquisitions far more rapidly. Furthermore, the new licensing and processing procedures allows a previously undreamed of level of interoperability and collaboration with partner archives, facilitating integrated discovery and presentation services, and joint stewardship of collections.

**Biographical Information:**

Dr. Micah Altman is Senior Research Scientist in the Institute for Quantitative Social Science in the Faculty of Arts and Sciences at Harvard University, Associate Director of the Harvard-MIT Data Center, and Archival Director of the Henry A. Murray Research Archive.

Dr. Altman's work has been recognized by the Supreme Court. His extensively-reviewed book, *Numerical Issues in Statistical Computing for the Social Scientist*, corrects common computational errors made across the range of social sciences. And his over thirty-five publications and five open-source software packages span political science, computer science, informatics, statistics, and many other disciplines in social and information sciences.

## **Introduction**

The Henry A. Murray Research Archive is the endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science, in Harvard University. The Murray preserves in perpetuity all types of data of interest to the research community, including numerical, video, audio, interview notes, and other types. The Center is unique among data archives in the United States in the extent of its holdings in quantitative, qualitative, and mixed quantitative-qualitative research.

The center's collection now contains over four-hundred and eighty research studies, comprising hundreds of quantitative electronic data files, millions of pages of textual data, and tens of thousands of hours (dozens of terabytes) of audio and video.

Founded in 1976, at Radcliffe, as the Henry A Murray Research Center. The Murray was given a mission to collect studies of lives over time. The Murray became the nation's leading archive for the study of human development, and its holdings have generated much additional research in the form of replications, reanalyzes, and follow-up studies (James and Sorenson 2000). It holds many studies on human development, including longitudinal studies of educational programs and policies, poverty programs and employment, family and children's issues, health behaviors and mental health (see Phelps, et. al 2002)

The Murray has been notable among data archives in the United States because of its focus on mixed-methods research and on studies that follow individual lives in historical and cultural context. Unlike most other archives, the Murray encouraged the archiving of original subject records including interviews, observational notes, test results, and administrative records, as well as the quantitative data coded from these sources. This qualitative component of the holdings offers richer opportunities for recoding and reanalysis. (Young, Savola & Phelps, 1991)

The Murray had a longstanding (although not exclusive) focus on the lives of American women. In 1995, the center began a project to increase the availability of data with a special focus on lives that have been historically underrepresented in social science research (see Phelps, Giele and Barbosa 2006). This effort resulted in the acquisition of several very rich studies being acquired, including thousands of hours of video interviews.

In 1999, Radcliffe, the Murray's home institution, merged with Harvard University, and began a transformation from college to research institute. Accompanying this transformation were large changes in the mission and budget of the institution. A number of significant organization and programs that had been within Radcliffe found new homes, the Murray among them.

In 2005, the Archive joined the Harvard-MIT Data Center (HMDC) within the Harvard Faculty of Arts and Sciences, in order to benefit from HMDC's research and infrastructure development in the field of digital libraries. During this time, HMDC became part of the Institute

for Quantitative Social Science (IQSS), which placed the Murray within a family of organizations that conduct research and develop infrastructure at the intersection of social science, research methods, and information science.

## **The Evolution of an Archive**

As the Murray departed Radcliffe for IQSS it was dealing with a number of substantial challenges. First, as Radcliffe completed its transformation to a research institute, it eliminated the subventions that made up a large part of the institute's budget. Second, the Murray held tens of thousands of hours of audio and video data that were stored only on aging analog tapes. Third, the processes and technologies at the Murray were cumbersome, slow, and based on obsolete technologies - these processes were in urgent need of retooling.

These challenges had a substantial impact on the operation of the archive. The director had departed, and much of the staff had been laid off. All teaching and research seminars had been discontinued, and ongoing data acquisition had been suspended except for those studies already under agreement. Even use of the archive was impaired since the Murray had always relied on staff to duplicate both paper and electronic materials to provide remote user access.

The transfer of the Murray coincided with the start of the NDIIPP funded Data-PASS partnership. 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 goals of Data-PASS are 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.<sup>1</sup>

While the project had been defined primarily in terms of a shared preservation mission, the archival collaboration it embodied, in combination with enabling technology developed at IQSS, played a major and unanticipated role in the evolution of the Murray as a digital archive. The results were transformation in four areas: archival practices, processing, cataloging, and collection development.

Digital library technology developed at Institute of Quantitative Social Science (IQSS) co-enabled these transformations. In the years preceding the transfer, the Murray had been working with the Harvard-MIT Data Center (HMDC) to expose its holdings through HMDC's Virtual Data Center (VDC) digital library system (see Altman, et. al 2001). HMDC developed this as an open-source system to support federated libraries of data and virtual collections. The VDC system supported search, browsing, downloading, on-line analysis, and extraction of research data for individuals. As well as harvesting, virtual collections, access control and terms of use for groups.

The VDC was succeeded in 2007 by the Dataverse Network (DVN). The DVN was a complete code rewrite, and added support for the creation of virtual archives. These virtual archives, called *dataverses*, are self-contained data archives hosted using a single DVN installation, and

---

<sup>1</sup> For a detailed description of the Data-PASS project see Guttman, et. al, in this volume. The Data-PASS operations, best practices, and shared catalog are described in more detail in Altman, et. al (in press).

individually branded, curated, and managed by scholars and organizations (King, 2007) under their own terms of use (which are automatically managed by the system). Multiple dataverses are hosted by a single institution in a single Datverse Network (DVN), and multiple DVN's can be connected through harvesting.

Although initially developed at IQSS outside of the Data-PASS project, this technology has now been used throughout it: The DVN system powers the shared catalog, which allows searching and on-line analysis across the entire partners' holding; and runs the harvesting processes that populate the catalog. And the system is used by the Murray and the Odum institute (and is being tested by ICPSR) for the ingestion, management, dissemination and preservation of their holdings.

The DVN software continues to be developed by IQSS as an open source system. That means that anyone can download the software and use it to set up their own Network of virtual archives. The open source license also permits anyone to modify and adapt the software for their own purposes - so long as they contribute these code enhancements back to the developer community.

In addition, IQSS hosts a DVN which is open to all creators of data used for social science research. This DVN provides virtual archiving of data for scholars, journals, and research organizations, without their needing to install software or maintain servers. IQSS hosts the servers, ensures bit-level preservation, and uses the automated DVN tools to automatically migrate known formats to preservation friendly formats. This system hosts over a hundred and

forty "virtual archives", archiving and disseminating data for dozens of organizations and over a hundred scholars at dozens of organizations and over a hundred individual scholars.

The partner's adoption of the Dataverse Network software has provided useful feedback to the software development process. Not only did the Data-PASS project help ferret out bugs, it led to the incorporation of new features explicitly to support the preservation and dissemination activities of the partners. Some of these features include: providing a plugin to enable LOCKSS (Reich & Rosenthal 2000) preservation harvesting of DVN collections; integrating JHOVE (Abrams 2004) for identification and characterization of formats; and storing the core content of the DVN (metadata, data, and identifiers) in preservation-friendly form. This storage is designed so that in the case of a problem, the entire content of the archive could be recovered from a direct copy of the filesystem, without the need for DVN software to recover, reconstruct, or interpret the data.

### **Archival Practice: From "Ad-Hoc" to "Not-Bad"**

The potential volume of un-reclaimed social science data that could be acquired, and the need to make the most cost-effective use of limited resources, led the Data-PASS project to establish a coordinated approach to identification, appraisal and processing.<sup>2</sup> This has synergistic benefits for the Murray.

---

<sup>2</sup> A written set of "Articles of Collaboration" guide this collaboration at the strategic level. We have established a steering committee and an operations committee to coordinate collaboration across the institutions. We have also developed joint deposit agreements, metadata standards, and recommended practices. All of these are published on the partnership web site: <http://data->

The identification and selection process developed in Data-PASS is decentralized, yet coordinated. Each archive 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.<sup>3</sup> At the same time, the partners share a database of leads, confer regularly, and refer potential contributors to other partners based on the fit with the partners' collection development policies.

This collaborative acquisition process has reduced duplicative effort across the partnership. It has been particularly beneficial to the Murray, which has moved from an expensive and labor-intensive acquisition process, involving regular travel and multiple levels of internal and external committee review to a more light-weight processes. The collaboration with Data-PASS partners has allowed us to continue to cast a wide net for data through the partners. And this collaboration has provided a pool of expert reviewers for those acquisitions requiring more extensive review.

To ensure consistency throughout this process, Data-PASS developed several sets of documented best practices including: appraisal guidelines to aid in prioritizing these acquisitions; processing guidelines for making these acquisitions; confidentiality standards; data security standards; a standardized deposit agreement

---

[pass.org/](http://pass.org/) . See Altman, et. al, in press for a detailed discussion of the collaborative agreement and structure.

<sup>3</sup> While the partners have not during most of this time had formal written collection development policies, we have regularly communicated our informal development policies through the operations and steering group. Recently, we have summarized our ongoing collection development policies in Guttman, et al (this volume.)

for materials shared in the project; and a standard set of core metadata for the common catalog.

The partnership-wide creation of best practices has dovetailed with the Murray's efforts to update its own processes and practices. The resulting new practices helped us to separate our historical practices from the core archival requirements they served, to re-examine these practices in the light of our partner's practices and externally developed standards, and to develop a new set of standards that still incorporated the Murray's archival experience.

For example, the Murray took the lead in the development of confidentiality standards for the partnership, because of its extensive experience archiving sensitive data. Historically, the Murray would individually tailor the deposit agreement for each new acquisition, resulting in a quite heterogeneous set of restrictions on dissemination, use, and preservation of the data. Creating a best-practices document prompted us to revisit the underlying principles ethical principles motivating confidentiality, the laws governing it, and the technical methodologies used to protect it. Because the document was partnership-wide, completing it required reviewing not only with legal and archival experts at our own university, but also incorporated the expertise of the partners and of the Library of Congress. The result was a more modern, coherent, and systematic set of practices.<sup>ii</sup>

We have since used this best practice document to replace our older confidentiality policy. Based on these identified practices, the Institute for Quantitative Social Science extended the software to

support these new policies. The Murray, through the Dataverse Network system, now provides a standard way to document and present terms, and the ability to automatically allow access once the requestor agrees to appropriate usage terms. It also provides a standardized way to extend terms, such that additional terms and modifications are incorporated into the metadata for the study (rather than being stored in a separate document) and clearly distinguished from the standard terms. At the end of last year, we were able to incorporate all of our new legal deposit agreements into the automated deposit system. This eliminated the delay in processing and availability that occurred when paper agreements had to be signed and mailed.

Similarly, the efforts of Data-PASS to develop a core set of common metadata standards, data security standards, and deposit terms have dovetailed with the Murray's efforts to adopt new practices in these areas. The deployment of a dataverse (an individual archive within the Institute of Quantitative Social Science's Dataverse Network) for the Murray archive has provided the technical foundation that the Murray Archive has used to re-engineer the entire acquisitions and disseminations processes at the Archive to be consistent with the "not-bad" practices identified by Data-PASS as part of the NDIIPP project.<sup>4</sup> This has also allowed us to automate all stages of our processing workflow: managing metadata and metadata exchange, ingestion, and computation of descriptive statistics.

---

<sup>4</sup> See Altman, et. al (in press).

## **Processing: from 'Analog' to Digital**

Over the decades, the Murray has acquired a large amount of qualitative social science data that were delivered in analog formats: tens of thousands of hours of audio and videotaped interviews; tens of thousands of pages of documentation; and millions of pages of paper data (survey responses, administrative records, interview transcripts, etc.). Much of the Murray's ingestion and dissemination activities were 'analog' as well: photocopying material (e.g., for dissemination, or onto acid-free paper for storage); retrieving media from depositories (e.g., for use by local visitors); physically masking text (for confidentiality); and mailing media to patrons.

As part of the migration of the Murray to Harvard, the Murray launched a university-funded project to digitize documentation and at-risk analog formats. Although the Data-PASS project efforts (and all NDIIPP projects) were restricted to "born-digital" materials, working as part of Data-PASS aided the Murray's migration to a digital process. The Data-PASS project and the Dataverse Network software helped to establish a complete set of processes and infrastructure for all new collections, which we then adapted to existing analog collections.

Now, our pre-existing analog collections materials are discovered, catalogued, documented, and disseminated through the Institute of Quantitative Social Science's Dataverse Network along with the born-digital collections. The main remaining difference between the two is that when a user applies to access a collection

that is currently in analogue format, we digitize and upload the materials at that point.

### **Dissemination: from Local to Global**

As a result of the Data-PASS project, for the first time the Murray is participating in a federated cataloging system, powered by the Dataverse Network software. The Data-PASS catalog is itself a milestone for the data-archiving community, containing the tens of thousands of studies that comprise each partner's entire data holdings.

Participation in the catalog also has direct benefits for the Murray. With no additional cataloging effort, the Murray collections garner more use because they are automatically incorporated into the Harvard University library systems, the Data-PASS catalog, and into other independent federated library catalogs such as OAISTER (Hagedorn 2003). Indirectly, participation in the shared catalog has lead us to standardize our cataloging records and process.

### **Collections: from Curator- to Depositor-focused collection development**

Through the Data-PASS project, the Henry A. Murray Research Archive has successfully acquired more than 140 significant research studies in social science (comprising over one hundred and seventy five datasets), most of which will be processed and available through the joint shared catalog by the time this article is published. In the previous thirty-year history of the Murray, the archive had collected on average nine studies per year: During the Data-PASS project we were

able to multiply our normal rate of acquisitions by more than four fold.

Some notable collections include:

- The landmark Longitudinal Study of Personality Development, conducted by Jack and Jeanne Humphrey Block. (Block and Block 2006) This is the most intensive study of human personality development in existence. Spanning a 30-year sweep of American history from the start of the Nixon Administration to the end of the 20th century and the dawn of the Digital Age, the study follows a sample of 128 children from age 3 through age 33 years, revisited on nine occasions. Assessments evaluated the domains of social, emotional, moral, cognitive, and ego development using more than 100 different psychological instruments measuring thousands of psychological variables. This data is also unique in linking an extraordinarily broad set of quantitative measures to qualitative transcripts and hundreds of gigabytes of videotaped interviews. This study has already been the subject of over one hundred publications, and contains a wealth of data that has yet to be analyzed, and can be expected to yield insights into human personality for decades to come.
- The International Food Policy Research Institute (IFPRI) has agreed to make its extensive collection of social accounting matrices, institutional surveys, and individual surveys available through the Data-PASS and the IQSS

Dataverse Network. These studies span over 10 years and 20 countries, and have been the source of numerous articles in economics and related surveys. (Most of these studies have not yet been received and/or processed and not included in the totals above.)

- Collections containing the holdings of the "Issue Correlates of War" (ICOW) project, a systematic collection on contentious issues of world politics (see Hensel 2001), and the highly-cited "Correlates of War" project. (See Singer and Diehl 1990) Although the latter had already been partially archived at partner institutions, the Data-PASS project supported the preservation of additional studies from that source, preservation of the related ICOW project data, and unification of these collections for the first time in a single catalog.
- Replication archives for the two journals, *International Studies Quarterly* and the *Annals of Applied Statistics*, which contains data associated with published articles. The journal staff will add data related to new data as articles are published.

In addition, there are now more than an equal number of studies deposited, processed, and available through IQSS DVN that are not archived elsewhere. More are added regularly.

Much of the success of our collections activity was a result of development, through the Data-PASS project, of a shared database of

leads, a streamlined set of acquisition agreements, and processing workflow.

The successful collaboration with Data-PASS partners for identification, acquisition, and best practice development has prompted us to re-engineer the entire acquisition process of the Institute for Quantitative Social Sciences and Murray Archive. Building on the joint work in the project to develop a Data-PASS shared deposit agreement, we have rewritten the Murray Archives' standard deposit terms to incorporate the Data-PASS deposit terms. Thus, all of the data collected in the future will be eligible for inclusion in the shared catalog, and for preservation by the partners (subject to operations committee review and approval). We have also revised our dissemination terms so that all new incoming data is available for public dissemination through Data-PASS. Furthermore, the IQSS Dataverse Network (DVN) has also incorporated these terms. Thus the hundreds of collections from individual scholars, institutions, research groups, and journals being hosted in the DVN are also available for preservation through Data-PASS.

The overall impact of these changes to the user are access to more data, more easily. Data in the Murray is now easier to find since more metadata is exposed to users, the studies can be found in multiple catalogs and virtual collections. Data is easier to access because the digital resources are now uniformly made available online, the DVN system facilitates access in convenient formats, and data is available under consistent terms of use. More data is available because organizations deposit the data themselves. In

addition, these changes have lowered the cost of archiving by distributing some of the cataloging effort, reducing the effort needed for acquisitions, and streamlining the processing workflow.

Another lesson learned was the importance of building archiving into institutional processes. Through Data-PASS we have established ongoing archives of dissertation data from university graduate programs, journals, and research groups. These ongoing archives capture data as it is produced, avoiding the need to “rescue” it later.

We have incorporated these lessons into the Murray’s operations policy by broadening our collection development policy. In doing so, we’ve adopted the core principle of the Data-PASS content selection criteria, which are that any social science data that is not currently managed by a permanent archives is considered to be at risk of loss and potentially worth preserving. As long as the data were collected or analyzed as part of a research design aimed at answering social science questions, it is now a candidate for selection into the Murray.

The Murray’s forward-looking collection development strategy is built upon finding ways to use the Dataverse Network software and its virtual archiving functionality to integrate archiving into the institutional processes of organizations that produce data and intellectual works based on data. In many cases, the end result is a living collection of data that is directly managed by the publisher, with preservation safeguarded by the Murray.

Complementing this approach, the Murray initiated an effort to establish lead to the development of partnership wide guidelines for living collections. These guidelines are now publicly available on the Data-PASS website (<http://www.icpsr.org/DATA-PASS>) as "Appraisal and Acquisition of Actively Curated Collections".

## **The Future**

### **Sustainability**

The Murray has now moved to an entirely digital workflow, supported by the Dataverse Network software. Much of the original analog holdings of the Murray have been converted to digital form, and are disseminated and preserved in that form. Furthermore, as part of our virtual archiving strategy, the Murray endowment ensures permanent bit-level preservation of data stored in the Institute for Quantitative Social Science's DVN Network, and the Murray staff takes supplementary professional cataloging, documentation, and preservation actions for selected collections.

The archivist must attempt to predict the long-term cost of digital preservation in order to develop a rational collection development policy. With a large collection of thousands of hours of digital video and audio, and regular opportunities to acquire thousands more, digital storage costs by themselves are a significant part of the Murray's planning. (As discussed below, however, storage costs are only a part of total digital preservation costs.)

Currently, high-availability, managed, locally-replicated<sup>5</sup> online storage costs approximately \$1500 per terabyte annually (all inclusive) for the Institute of Quantitative Social Science to provide, using its own technical staff and facilities. This is somewhat smaller than the cost of commercial online storage through Amazon S3 service which costs approximately \$1800 per TB annually<sup>6</sup>, and much smaller than the \$21000 per TB annually charged by Harvard's central IT for comparable storage.

Based on our current cost of storage, and with existing endowed funding, we could permanently store a collection several times that of our entire current holdings (including all of the virtual collections currently IQSS's Dataverse Network). And we can reasonably anticipate that storage costs will decline precipitously over time, as they have done over the last three decades.

However, storage costs are only a lower bound for total digital preservation costs. Ensuring preservation in the long term requires active monitoring and risk assessment and taking actions to address these risks. One well-known risk is format obsolescence. To protect against this, the curator must monitor the risk that formats represented in the archive will become obsolescent, identify or acquire tools to migrate objects to less risky formats, perform this

---

<sup>5</sup> The Data-PASS project is developing an experimental syndicated storage system that aims to provide collaborative archival replication across the different partners. When this is successful, we plan to migrate from locally replicated storage to geographically distributed and replicated storage. Over time, we expect the costs to go down.

<sup>6</sup> In addition Amazon charges incrementally for depositing and retrieving data from outside its EC2 computing service. This effectively raises the cost of using S3 over local storage for data delivery.

migration, and verify the results. Curators will also need to support users in accessing the collection, monitor and address other risks, including institutional risks to preservation, and risks to discoverability and usability of their collections. The long-term costs of digital preservation and access involve many components, and although some case studies exist for journals and related material (see Chapman 2003; Kenney, et al 200; Palm 2006; Rusbridge 2006; Davies, et. al 2007; Wheatley, et al. 2007); and there is no well-established body of published knowledge on the costs of preserving digital numeric data.

Nevertheless, storage costs have been a large proportion of the Murray's core archiving costs, historically. In addition, we have practically eliminated the duplication costs that were historically another major proportion of core archiving and dissemination costs for the Murray Archive. Moreover, we have reduced other cost components with these changes: The Murray's overall costs of maintaining collections are reduced by automating the entire workflow, and distributing the acquisitions and cataloging efforts among the various managers of the DVN virtual archives. So, even if, in our conservative estimates, up to half of the archive's total operating income must still be devoted to all of the other core preservation and dissemination activities, and even without a decline in storage cost or an increase in funding (both of which are likely over time), the Murray can support collections of twice its current size. If storage costs follow their historical trends, the Murray will be able to increase its holdings many times over. Overall, during the course of

the NDIIPP partnership, and in part because of the automation and standardization of processes catalyzed by partnership, the Murray has greatly reduced the costs of disseminating and preserving its collection.

### **Where do we go from here?**

Elder, et. al, wrote that the "growth of longitudinal data archives is one of the most dramatic recent developments in the behavioral sciences" (Elder, et. al 1993, pp. 1). They noted the huge potential to ask new questions of old data, but note that archival data are "never precisely what one wants or expects." (pp. 11)

Longitudinal data is changing, and data archiving is changing as well. In the near future, the evidence base on which the social sciences rests will rely more on "ambient" data that is generated (more or less continually) by people and organizations for their own purposes, and only later subject to scientific analysis.

Increasingly, social scientists will analyze web pages and links, Youtube videos, blog postings, GPS coordinate trails from highway-toll transponders, mobile telephone records, online social networks such as Facebook, and data from a variety of other new sources. These data too, provide a window into the study of individual's lives over time, but it is a window offering a very different perspective from traditional longitudinal studies, and a source of information requiring different tools and strategies to acquire, preserve, disseminate and analyze.

New forms of data archiving are emerging. With the explosion of the 'Web', 'self-archiving' where data producers distributing data

through their own sites (at least for a time) has become increasingly popular. At a different organizational level, hundreds of institutions have now established "institutional" archives to preserve the intellectual creations of the institutions.

We are also seeing the emergence of a new model of archiving, "virtual archiving", that separates, to a great extent, collection development and management from the infrastructure for data storage, preservation and dissemination. This development makes it possible for individuals, institutions, and virtual organizations to have great control over the content, dissemination, and branding on collections, and has the potential to support more seamless dissemination, more reliable preservation, and wider discovery than institutional and self archiving. This potential is most likely to be realized when the "virtual" archive is backed and supported by professional, permanent archives.

As this article has described, the Murray has made transitions from an analog to a digital workflow, from ad-hoc to 'best' archival practices, from local to global cataloging and dissemination, and from a curator- to a depositor- focused collection development policy. These transitions were shaped by, and in many cases, catalyzed by the collaboration with the NDIIPP Data-PASS partnership, and the use of the Dataverse Network infrastructure.

We are now working towards widespread high-quality virtual archiving. This is an approach made possible by technological advances such as the Dataverse Network, and institutional advances such as the Data-PASS partnership. Both provide essential forms of infrastructure

-- the first enables the creation of virtual archives, and the second helps to ensure that the content of those virtual archives can be preserved.

---

<sup>i</sup> This project is supported by the Library of Congress's National Digital Information Infrastructure and Preservation Program. We are also grateful to our institutions Harvard University for their part in the Data-PASS project's cost sharing. We thank Sonia Barbosa, Gary King, Nancy McGovern, and Copeland Young for their comments and suggestions.

<sup>ii</sup> This confidentiality document, and all the other best practices mentioned in this paper have been published on the Data-PASS project website: <http://www.icpsr.org/datapass/> .

## References

- Abrams, S.L., (2004). "The role of format in digital preservation", *VINE*, Vol. 34, Issue 2, pp. 49 - 55.
- 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*, Vol. 19 Issue 4, pp. 458-470.
- Altman, M., & King, G. (2007). A proposed standard for the scholarly citation of quantitative Data. *D-Lib Magazine*, Vol. 13 Issue 3/4. Retrieved June 29, 2008, from <http://www.dlib.org/dlib/march07/altman/03altman.html>
- Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (in press). Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences. *The American Archivist*.

- Block, J., and Block, J.H. (2006). "Venturing a 30-Year Longitudinal Study", *The American Psychologist*, Vol. 61, Issue 4, pp. 315-327.
- Chapman, S. (2003). "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?", *Journal of Digital Information*, Vol. 4. Issue 2.
- Davies, R., Ayris, P., Mcleod, R., Shenton, H. and Wheatley, P. (2007). "How much does it cost? The LIFE Project" *LIBER Quarterly*, Vol. 17, Issue 3/4.
- Elder, G. H. Jr., Pavalko, E.L., Clipp, E.C. (1993). *Working with Archival Data: Studying Lives*, Newbury Park: Sage Publications.
- Gutmann, M., Abrahamson, M, Adams, M.O., Altman, M, Arms, C., Bollen, K., Carlson, M., Crabtree, J., Donakowski, D., King, G., Lyle, J., Maynard, M., Pienta, A., Rockwell, R, Timms-Ferrara L., Young, C. (in press). From Preserving the Past to Preserving the Future: The Data-PASS Project and the challenges of preserving digital social science data. *Library Trends*.
- Hagedorn, K., (2003). "OAIster: a 'no dead ends' OAI service provider", *Library Hi Tech*, Vol. 21, Issue 2 pp. 170-18.

- Hensel, P. R. (2001). "Contentious Issues and World Politics: The Management of Territorial Claims in the Americas, 1816-1992." *International Studies Quarterly* Vol. 45, Issue 1 pp. 81-109.
- James, J.B., Sorenson, A., (2000). "Archiving Longitudinal Data For Future Research: Why Qualitative Data Add to a Study's Usefulness", *Forum: Qualitative Social Research* Vol 1. Issue 3.
- Kenney, A, Entlich, R, Hirtle, P., McGovern, N. and Buckley, E., (2006). *E-Journal Archiving Metes and Bounds: A Survey of the Landscape*. Washington, D.C. : Council on Library and Information Resources.
- King, G. (2007). An introduction to the Dataverse Network as an infrastructure for data sharing. *Sociological Methods and Research*, 36(2), 173-199.
- Palm, J. (2006). The Digital Black Hole, TAPE project: Training for Audiovisual Preservation in Europe, <<http://www.tape-online.net/>>.
- Phelps, E., Giele, J.Z., Barbosa, S. (2006). "Studying Diverse Lives", *Research in Human Development* Vol. 3, Issue 4, PP: 185-190.

- Phelps, E. Furstenberg, Jr., F.F., Colby, A, (eds). 2002.  
*Looking at Lives*, Russell Sage Foundation: New York.
- Reich, V. & Rosenthal, D.S. (2000). "LOCKSS (Lots Of Copies Keep Stuff Safe)", *Preservation 2000, The New Review of Academic Librarianship* Vol 6, pp. 155- 161.
- Rusbridge, C. (2006). "Excuse Me... Some Digital Preservation Fallacies?", *Ariadne*, Issue 46,  
<http://www.ariadne.ac.uk/issue46/rusbridge/intro.html>
- Singer, J. D. and P. Diehl (eds). (1990). *Measuring the Correlates of War*. Ann Arbor: University of Michigan Press.
- Wheatley, P., Ayris, P., Davies, R., Mcleod, R. and H. Shenton, (2007), "Full Report from the LIFE Project",  
<http://www.life.ac.uk/2/documentation.shtml>
- Young, C., Savola, K.L., Phelps, E. (1991) *Inventory of Longitudinal Studies in the Social Sciences*, Newbury Park: Sage Publications.