

Building Relationships “A Foundation for Digital Archives”

Jonathan Crabtree

The Odum Institute
University of North Carolina
22 Manning Hall
Chapel Hill N.C. USA
(919) 962-0517

Jonathan_Crabtree@unc.edu

Darrell Donakowski

ICPSR
University of Michigan
P.O. Box 1248
Ann Arbor, MI
(734) 647-5000

dwdonako@UMich.edu

Abstract

In the digital age, what is the best way to build and sustain data archives? Collaborating with other archives could improve how they operate and grow, but building relationships consumes time and scarce resources. Is it worth the effort? Six major digital archives are exploring this issue through a partnership funded by the Library of Congress's National Digital Information Infrastructure and Preservation Program. This partnership's efforts to develop a federated approach to data archives require building relationships at the producer, administrative, and program application development levels. Now two years into its development, this alliance has accomplished a number of important objectives in each of these realms. This paper highlights the experiences of two alliance members, the ICPSR and the Odum Institute. Our interactions and accomplishments lead us to believe that the benefits of partnerships such as this one far exceed the costs they entail. More particularly, establishing collaborative relationships between archives achieves four objectives, as follows: (1) It facilitates communications between archiving professionals, enhancing efficiency by creating a common pool of knowledge and a framework for ongoing interactions and education; (2) It improves our relationships with data providers by enabling us to provide a better and more durable quality of service; (3) It allows archivists to build networks of relationships with software developers, increasing the probability of identifying, developing, and adopting broadly functional applications serving a multiplicity of needs and audiences; and (4) In promoting development and adoption of common standards, it dramatically improves the probability of effectively networked collections and diminishing the costs involved in creating them. While our focus is social science data, the approach would work in many fields.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Collections, Dissemination, Standards, Systems issues, User issues.

General Terms

Management, Design, Human Factors, Standardization.

Keywords

Digital Archives, Alliances, Federation, Harvesting, Social Science Data

1. Introduction

Throughout the world, universities, businesses, governments, and archives are facing the task of building as well as maintaining digital archives. Archive administrators must navigate increasingly complex and technical waters to accomplish this task. They must make decisions on metadata standards, database standards, and interface standards. IT managers and archivists must choose between a growing number of innovative open source options and the security and support traditionally provided by more expensive proprietary systems [1]. A common thread among these options is the relationships needed to be successful. Regardless of the technical solution chosen, the need for collaborative relationships is constant.

2. Data-PASS Experiences

The Data Preservation Alliance for the Social Sciences (Data-PASS) was created with an award from the U.S. Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP) [2]. The NDIIP mission is to develop a national strategy to collect, archive and preserve digital content, especially materials created in digital format [3]. Data-PASS is led by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. Other partners include the Roper Center for Public Opinion Research at the University of Connecticut, the Howard W. Odum Institute at UNC Chapel Hill, the Henry A. Murray Research Archive, a

member of the Institute for Quantitative Social Science at Harvard University, the National Archives and Records Administration, and the Harvard-MIT Data Center, also a member of the Institute for Quantitative Social Science at Harvard University. The partnership's primary goal is to identify and preserve historically-significant digital social science data at risk of being lost. However, the process of creating the partnership has itself provided important benefits and taught us a great deal about the benefits of formalized collaborative relationships.

Building relationships is critical to successful data archiving, but it is time consuming. Although relationships come in many shapes and sizes, the central ingredient necessary for success is trust. To develop partnerships, archivists working in different organizations must trust the quality of each institution's data and documentation, as well as the hardware and software that make those data usable. Principal investigators must trust that archives will value and preserve the information they have collected, and provide systems that make those data accessible. Archivists must trust software developers to deliver reliable, innovative platforms for storing, retrieving, and analyzing important digital data. External organizations contemplating working with existing partnerships must trust the value of what membership in such partnerships can offer them. And ultimately, researchers must trust the quality of the archived data they use to conduct secondary research as well as of the documentation that makes such usage possible.

2.1 Developing Inter-archival Relationships

From the beginning, the founding members of Data-PASS have had a good working relationship. Our similar goals and missions provided the incentive to work together on this project. Despite the competitive academic world each of us lives in, we have each been able to find common ground by contributing to the effort in our individual areas of specialty and strength, and in our ability to build quality relationships with researchers with diverse backgrounds and levels of expertise.

In other words, we as archivists quickly recognized many of the benefits that we could achieve through the partnership. We have not been disappointed; formalizing our relationships increased our awareness of each others' activities, improved our communications, and provided valuable and ongoing professional development opportunities at low cost. It has also generated an important result whose full potential we did not fully recognize at the outset of this partnership, namely, enhancing our relationships with data producers. Building and maintaining the relationships between data producers and data archives are among the most important tasks an archivist has, and these tasks merit special attention.

2.2 Enhancing Relationships with Data

Producers

A key factor guiding the activities of our partnership is our shared and growing recognition that the relationships developed with researchers, i.e., data producers, are critically important. As our partnership exists in order to identify and preserve historically significant social science data, it is essential that each of us work effectively with data producers if the partnership is to succeed in its mission. Strong relationships often begin with education. Researchers need to understand the benefits of archiving data. Among these benefits, archiving data reinforces open scientific integrity and encourages diversity of analysis and opinions. To further highlight the benefits and assist researchers in the process, ICPSR has produced a third version of its "Guide to Social Science Data Preparation and Archiving". The guide provides information that is of use throughout the cycle of research, from applying for a grant to preparing the data for deposit in a public archive. While it does not attempt to address policies and procedures of specific archives, it can serve as best practices in the creation and documentation of datasets [4].

It is also important to remember your audience when discussing the benefits of archiving data. Many researchers come from academic institutions, however, the acceptance and recognition of the importance of data sharing can depend on the culture of their particular academic field. Most recently, an increasing number of academic disciplines recognize that sharing data promotes new research and allows for the testing of new or alternative methods. They are also becoming more open to the scrutiny of others using their data, recognizing that with scrutiny can come improved methods of data collections.

It should also be noted that private non-profit research organizations also 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 above 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 archive can work with the data producing organization to alleviate perceived burdens.

Another factor influencing our interactions with data producers – academic and non-profit alike – is the greater attention granting agencies are paying to archiving funded research. 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.

Building and maintaining the relationships between data producers and data archives are among the most important tasks an archivist has. As with many relationships, a key aspect is trust; trust in the capabilities and integrity of the archivists, leading to trust that the data will be securely stored and reliably preserved. Data-PASS has significantly augmented our ability to ensure that

the data are secure safe, because the Data-PASS partners are committed to archiving multiple copies of each data collection in multiple locations. To ensure that the data are not only preserved, but also accessible in the future, the partners also recognize that current best practices need to be implemented and consistently revisited to keep up with advances in IT. Thus, in the short term, Data-PASS increases what each member organization can offer the data providers whose collections it archives. In the longer term, we plan to increase researchers' awareness of the exceptionally high quality of the services the partnership provides, including its assurance of data security, preservation, and accessibility. Becoming recognized as an existing, durable partnership consistently offering the highest level of service will reduce the costs we face in securing, documenting, and preserving data, enabling us to increase the scope of our offerings without a proportional increase in staffing and effort.

2.3 Interacting Effectively with Software Developers

Another important increase in efficiency realized through Data-PASS has been achieved through our efforts to identify – and work to develop – broadly functional software applications serving a multiplicity of needs and audiences. Our joint needs are both broad and deep, including the storage, updating, and transport of a wide assortment of data collections of different sizes, structures, types, and uses; the need for a variety of search functions, some simple, some complex; and growing demand for a wide array of descriptive and statistical tools capable of interacting dynamically with the datasets we provide. All of these resources must be available and accessible for use by researchers of varying backgrounds and skill levels.

Exploring different archival software packages leads down many paths. Custom building a system might require collaborating with a team of developers. Sometimes the most affordable option is to adopt technology developed at other institutions. Regardless of the approach, these collaborations call for strong communication between developers and archive administrators. For example, the Odum Institute is testing the Virtual Data Center (VDC) as an open source platform to upgrade our archive, and we have been working closely with its developers, the Harvard-MIT Data Center [5]. Although the Odum Institute has one of the oldest and largest archives of machine readable data in the U.S, it is important to keep adapting to a constantly changing IT environment [6].

The Odum Institute is currently looking to migrate away from the legacy systems currently used in its archival process. After many months of research and investigation it became apparent

that regardless of the technology chosen, the relationships with the software developers would be critical. The job of domain specific archives like Odum requires processes and systems that require many customizations. Our choice to focus on open source solutions has allowed us the opportunity to work with a developer such as the Harvard-MIT Data Center to better achieve our goals. We still have diligent work to be carried out but we are confident that we will achieve the high standards we seek.

The results of collaboration are evident in the Data-PASS shared catalog. This requires extensive cooperation with software developers as well as archival partners. Under a common catalog approach, users will be able to search and download data from all partnering archives. We feel that this will be an enormous benefit for social science data users. The ability to search a common or union catalog from any of the partner sites will allow the social science community nearly instant access to a vast and diverse array of datasets. This will not only benefit the seasoned principle investigator but scores of brilliant graduate students that are the future of our community. Data-PASS is currently seeking to develop the coalition in hopes of expanding the amount of information provided in this catalog. The partnership's choices of transfer and metadata standards are at the heart of this ability.

2.4 Adoption of Common Standards

Adopting common standards for information systems lays the ground work for relationships to grow and prosper. The partnership permits a much higher level of inter-archival cooperation, including mutual support in backing up the archives and an agreed upon acquisitions policy. Common metadata standards among partners are also essential, and the social sciences have chosen DDI as a common metadata standard. DDI provides the in depth tagging structures needed for complex datasets and is compatible with the Dublin Core [7]. Other organizations that use DDI and wish to share in the common catalog only need to expose it for harvesting. The use of common standards removes the barriers that sometimes prevent cooperation within groups with similar goals and interests.

The metadata standard is just the first technical step to removing the barriers. The world of digital archives has many choices when it comes to an archival platform, even within the open source arena. One theme among some of the technologies available is the concept of federation. The Harvard-MIT development team used the Open Archive Interface (OAI) to allow the federation of archive information between partners [8]. Using these common technical standards has allowed partners that expose their metadata using an OAI server to join Data-PASS and keep technical costs within acceptable limits. To further prevent restrictions the partnership allows different levels of participation. This can be as simple as exposing a catalog for harvesting or placing and managing ones holdings in a VDC node located on site. These moves to federate will strengthen all the data archives involved and have the potential to make them more efficient and more innovative in the eyes of their constituents.

3. Conclusions

The Data-PASS partnership has allowed its members to do many things that would not have been possible alone. From upgrading technical infrastructure to adopting the same metadata standards, partners have benefited far beyond their investment of time and energy. Digital archives are, by definition, connected to the world in a way never before possible. Future collaborations should also include the emerging group of institutional repositories. Data archives around the country are poised to assist these new repositories with metadata creation and dataset processing tasks that even the largest of library staffs have trouble managing. Divisions of labor will have to be drawn but in the new culture of open source software and federated approaches these should be possible. New developments in software and hardware present challenges, but also opportunities for building and strengthening digital archives. In the future, those who succeed will likely do so through collaboration and through building enduring relationships. One of the key ideals we must collectively remember is that we must build not only sustainable archival technology but sustainable organizational cyber infrastructure. In doing this, we support each other as we move through the ever shifting sands of technology.

4. Acknowledgements

Our thanks to The Odum Institute and ICPSR as well as our Data-PASS partners.

References

1. Altman, M., 2003. "Open Source Software in Libraries: from Greenstone to the Virtual Data Center, and Beyond", *IASSIST Quarterly* 25(1):5-12
2. Data-PASS. *Web site*. <URL: <http://www.icpsr.umich.edu/DATAPASS/>> (04/26/06)
3. NDIIPP. *Web site* <URL: <http://www.digitalpreservation.gov/about/index.html>> (05/19/06)
4. ICPSR (Inter-university Consortium for Political and Social Research). (2005). *Guide to Social Science Data Preparation and Archiving*.
5. The Data.Org Project. *Web site*. <URL: <http://thedata.org/index.php/Main/AboutVDC>> (04/26/06)
6. The Odum Institute. *Web site*. <URL: <http://www.odum.unc.edu>> (04/26/06)
7. The DDI Initiative. *Web site*. <URL: <http://www.icpsr.umich.edu/DDI/codebook/index.html>> (04/26/06)
8. OAI Project. *Web site*. <URL: <http://www.openarchives.org/>> (04/26/06)