# Virtual Archiving for Public Opinion Polls

Jonathan Crabtree
University of North Carolina
Odum Institute
22 Manning Hall
Chapel Hill NC USA
+1 919 428 6112
Jonathan_Crabtree@unc.edu

## ABSTRACT

The Odum Institute for Research in Social Science Data Archive at the University of North Carolina, and partners from the National Network of State Polls present progress on a two year demonstration project using the Dataverse Virtual Archiving technology [1]. The goal of the Virtual Archiving for Public Opinion Polls: A Demonstration Project aims to streamline the ingest process and increase timely submission to data archives. Bridging this gap between producers and archives will increase the overall submission rates and ultimately preserve many data sets that would otherwise be lost.

Around the world researchers and scientists collect vast amounts of data, which often are not archived after the completion of the project or task. As researchers move on to new projects, past data they have collected are seldom documented and preserved [6]. Until the tools for data curation are integrated into the research lifecycle of data, we will continue to experience this problem [2]. This project seeks to provide a solution for this problem. Although the virtual archiving technology needed to bridge the gap between the data producers and archives already exists, the availability of this tool and its value needs to be communicated to the scholarly community.

The technology we use for this demonstration can be applied to many disciplines and data types. In this demonstration, we use public opinion data producers because these data serve as a useful, readily recognized example that will be widely replicated. Public opinion survey data are the most prevalent single kind of social science data and usually what most scientists first encounter.

The Odum Institute's relationship with the various state polling agencies and the National Network of State Polls make it an ideal candidate to propose new and innovative changes in the data life cycle of public opinion polls. This projects builds on our previous work with the Dataverse Network (DVN), developed at Harvard University. The Odum Institute has been an active partner in the DVN development and has recommended system modifications to allow for the maximum flexibility in public opinion preservation, distribution, and analysis. The DVN provides the tools necessary to implement this change. In this project, we use the DVN technology to aid data producing agencies in the ingest, curation and preservation of public opinion data, election polls and reports.

Automated ingest tools specifically designed for quantitative data are used to create metadata automatically on ingest. This is critical for two reasons. First, metadata are essential in making data accessible to the scholarly community beyond those who were involved in the data collection. Second, the creation of metadata — itself a technical field with its own set of tools and norms — is a specialty that lies beyond the expertise of most research teams.

The project is creating customized Web interfaces or "virtual archives" for each of the participating data producers. These interfaces are created to allow the researchers to seamlessly upload their data. The data will be transparently archived, preserved and curated by an organization trained in the field of long-term preservation. Once in the Dataverse Network, the data can be discovered and accessed via the existing federated search capabilities of the DVN. We are designing new workflows and help train the participating data producers to use these new and efficient methods of data ingest.

## Categories and Subject Descriptors

H.4.1 [**Office Automation**]: Workflow management

## General Terms

Management, Documentation, Design, Human Factors, Standardization.

## Keywords

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


*iPRES2011*, Nov. 1–4, 2011, Singapore.


## 1. INTRODUCTION

Researchers and scientists collect vast amounts of data worldwide. Often these data are not archived or preserved following completion of the primary task for which they were intended. A number of scholars in their respective disciplines are champions for the difficult tasks of documentation and archiving; however, these tasks seldom receive funding and are often the first items cut from budgets.

Though the list of explanations can be expansive, the research community will continue to experience this issue until the tools for curating data are integrated into the research lifecycle of data.

The virtual archiving technology required to bridge the gap between the data producers and archives already exists [3]; the necessary next step is enhanced awareness of the accessibility of this technology and communicating its value to both scholars and the academic community.

Data-producing organizations are typically supportive of archiving the materials produced. Economic and workflow issues tend to inhibit attempts at comprehensive archiving. In order to maintain the economic feasibility of their data collection organizations, researchers must often move from one project to the next with minimal downtime often undermining attempts to adequately archive valuable data. The costs associated with rehiring qualified staff when new projects arise and retaining staff on the payroll with no outside financial support are considerable. Organizations must consistently maintain a queue of new projects and opportunities in order to justify their continued existence; a strategy that reserves little time and resources for data archiving.

The collection of social science research data — and particularly public opinion data — is ensnared in this problematic process, resulting in the loss of numerous valuable datasets [5]. Access to empirical social science data is fundamental to successful social science policy development, research and education. For example, students and teachers who wish to gain a deeper understanding of the findings in economics, psychology, political science, sociology, educational research, and other social sciences must be able to discover and access the data that constitute these studies. Teachers and students in the natural sciences also routinely encounter the products of empirical social science in surveys, newspaper editorials, magazine articles and other academic research products. The data that underpin many social science research studies discoveries and theories have not been consistently archived despite mandates from funding agencies such as the National Institutes of Health and National Science Foundation. This is due primarily to the enormous degree of post-project effort required to prepare data for archiving. To help ease this burden, data archive managers and data producers must work cooperatively.

## 2. VIRTUAL ARCHIVING

The Virtual Archiving for Public Opinion Polls project demonstrates a streamlined process for data submission from polling agencies across the country. The Institute will develop virtual archives for data producers to facilitate simple, direct access to the submission process. In traditional research data archival workflows, materials arrive at the archives after the project is completed. Ideally, the research teams would assemble the data, materials and any existing documentation post-project. Then, the materials are forwarded to the archive for ingest processing. Ingesting involves preparing data for archiving, de-identifying personal and confidential information, creating standard file formats, building any necessary metadata and documenting this process. The depth and quality of the ingest process varies greatly in each situation, and the effort required to assemble the components often limits the amount of materials archived. In most cases, the researchers have already moved onto new projects and do not have the time to follow through with the archiving steps.

With virtual archiving, the researchers begin using the archival tools earlier in the research process. These simple Web-based tools allow researchers and their staff to manage their data and documentation throughout the life cycle of the project. The virtual archives that result recreate the look and feel of the home institution Web sites. Simple ingest procedures provide metadata validation routines that assist in documentation and even prompt researchers to enhance their metadata. When quantitative data is ingested, automated routines create detailed variable-level metadata without requiring costly manual procedures.

The goal of virtual archiving is to provide simple tools that research teams can use to easily manage their data. As these research teams begin submitting their datasets, the ingest tools collect and verify much of the required metadata. When the research team releases the dataset by setting appropriate permissions, the archival submission process is complete. Although the process seems to the research team to be local on their Web site, the data is stored in the remote archive site. The data is backed up and preserved in a trusted replicated network from the moment it is ingested until it is released to the public. Trained archivists manage the process and ensure that important documentation and archival formats are created to ensure proper preservation. After datasets are archived, users will be able to search for the data from the producer's local Web sites and other scientists will be able to discover these studies and harvest the metadata using the Dataverse Network. Credit and acknowledgement for the data will remain with the research teams who produced the original data. The virtual archives are part of a national federated network of social science archives that aid in the dissemination of the work.

The demonstration project is developing archival and ingest workflows for five social science polling centers: the University of South Carolina Institute for Public Service and Policy Research, Monmouth University Polling Institute, the University of Georgia Survey Research Center, the University of Arkansas and the University of Indiana Center for Survey Research. Once these demonstration sites are complete, the Odum Institute plans to implement this technology at other national state polling centers.

This project focuses on public opinion or election data and the polling agencies that collect them, but the virtual technology involved — as well as our open source and distributed acquisition method — can be applied to other disciplines and data producing organizations. Though many of the features and analysis components are geared toward quantitative data, the virtual archive process remains applicable to qualitative data, documents, and images. A virtual curated preservation environment will promote effective and timely archival dataset preservation. The ability to customize virtual archives to meet the needs of individual data producers adds to the value of this workflow system. The benefits of a simple ingest workflow for datasets is considerable. Breaking down the barriers to ingesting data will ensure that a significantly larger portion of research data are archived properly. Though the precise impact of this virtual ingest demonstration may be difficult to estimate, we anticipate that archive submissions will increase by more than fifty percent.

## 3. PROJECT DESIGN

The project is developing virtual archives of election and public opinion poll data, a versatile system that will assist data producers across multiple disciplines.

The primary project goals are:

- Demonstrate use of the Dataverse Network and virtual archives to streamline the submission workflow process. These virtual archives will provide data producers tools for ingest, automated metadata creation and validation using Web-based client-server technology while preserving the look and feel of their local Web site storage. These Web-based workflows will allow data producers the ability to upload their datasets in the archive and document them in a seamless client-server environment. Once the data is archived, producers can assign rights, analyze and manage their data using the many Web-based services offered by the DVN;
- Build generic virtual archive template models to allow simple adoption of DVN technology. Templates are a predetermined set of code generic enough to fit applications across repositories and domains. They will be built using Java code, XML and HTML and can be easily modified to accommodate colors, logos, headers and footers to readily create the look and feel of the home institution Web site. These templates will reduce the cost of future virtual archive creation and integration;
- Work with polling data producers in an effort to increase archival rates;
- Train data producers in the use of quantitative automated ingest tools; and
- Disseminate findings and experiences to the preservation and data producing communities.

In addition to the initial, in-depth evaluation processes, the project consists of four major areas of effort: research and design, programming, training and reporting. The work plan begins by evaluating producing agencies to ensure that programming and design take full advantage of similarities across sites. Evaluation is not necessarily a onetime, linear process; findings from later phases will inform the ongoing design and programming phases.

Phase I: Research and Design

Odum programming staff and research assistants are working in conjunction with, and seeking input from, individual data producing agencies to understand existing workflows and processes local to each demonstration site. The mission for this phase is to understand the local environments of the demonstration sites, design individual virtual archives for these sites and seek commonalities for use in designing templates to reduce the cost of future virtual archive design. Odum staff seeks to find similarities and efficiencies in the design of the virtual archives for the demonstration sites. These commonalities will be used to create base programming templates during the programming process. This phase will also involve collecting baseline quantitative information from each producing agency on how many datasets they have archived. This will allow us to compare the numbers of datasets archived (and how long archiving took) by each agency before and after implementing the virtual preservation process. This information will allow us to ultimately quantify our results and identify a metric of success.

Phase II: Programming

Odum archive staff are creating virtual archives within the Odum Dataverse Network archive software for each demonstration site. These archives are customized portions of our archival system designed to house the producer's data. Once a dataset is part of the system, data producers can define access controls, download data, analyze data using statistics and promote international discovery of their data though the Institute's federated archival network. Odum programmers work with designers to construct Web-based interfaces for the new virtual archives. These interfaces integrate the demonstration partners' existing Web sites and provide continuity of appearance and function for the researchers and users. The Institute will incorporate ongoing recommendations from the individual data producers to provide a streamlined ingest workflow and minimize barriers to submission. Developing these Web-based interfaces requires customized programming that can be costly and inhibitive to widespread adoption of the technology. For this reason, we will create programming templates that build on the commonalities among the participants and will use these similarities to create code templates reducing the future cost of virtual archive integration. Templates will reduce costs and will provide a base for future dissemination of the technology to a wider community. We are document the programming process to assist in developing training materials in the next phase.

Phase III: Training

Institute staff are designing training documents and visual aids for data producers and will help train remotely the data producers at the demonstration sites in using the automated ingest tools and metadata templates. In addition to hands on training, Web-based video instruction will be used to provide economical and effective training. The Institute will produce and disseminate live Internet streaming of Institute short courses designed to educate data producers on the virtual archive workflow process. Once training is complete, project staff will supervise and provide ongoing technical support for the demonstration sites.

Phase IV: Reporting

This final phase will assemble reports for the sponsor as well as develop and execute the final evaluation surveys. Project staff will work with the Odum Institute survey design and methodology staff to develop these surveys, which will gather information on how many datasets, are being archived and how long the process takes for each data producer. Following this phase, the Odum staff and data producers will leverage the existing social network within the National Network of State Polls to help disseminate the findings both nationally and internationally to archivists and data producers.

## 4. CURRENT PROGRESS

The project is nearing the end of the first year and has been very successful to date. Qualitative interviews with each polling agency have been conducted and provided to the design team with information to begin the programming process. Programming has been completed for four of the NNSP partners with one of the virtual archives already in use. The team has designed the Web interface for the individual virtual archive and developed ingest templates for the different survey methodologies used by the agency. These templates record commonly used metadata responses to enable polling agencies to streamline the ingest process and easily train their staff to use the virtual archive for ongoing data management and documentation by simplifying the process. Initial reactions to the interface and processes have been very favorable.

Current efforts include finalizing programming on the remaining virtual archive interfaces and developing training documentation using the feedback from the initial deployments. We are developing training videos for web-streaming application to assist in the education of our participating agencies and for continued use in the dissemination of this technology.

## 5. CONCLUSION

The research community is faced with expanding burgeoning collections of digital data. As researchers and scientists struggle to deal with this vast amount of information, they still have to continue their primary scientific work and would like assistance in this process [7]. A recent poll of *Science's* peer reviewers shows that 20% of those asked were creating data sets larger than 100 gigabytes and 7% used data sets greater than 1 terabyte [7]. When asked where the respondents archive the data created by their research, over 50% claimed they stored the data in their labs [7]. Additionally, 38.5% reported that they archived their data on university servers while only 7.6% used community repositories [7]. This leaves most data to reside outside of archival repositories and beyond the care of data curators. This lack of stewardship places much data at risk and raises the questions of "what roles can digital archives play in the preservation process and when should they become involved in the data lifecycle". This project seeks to insert archival processes earlier in the research data lifecycle in the hopes of ingesting a larger portion of these valuable projects. Early indications are very positive and data producers are very open to examining this change in workflow. Current demands by funding agencies for research data management plans have forced researchers to think about the preservation of their data during the proposal development process [4]. Archives should provide assistance in this process and need to provide tools that aid in reducing the efforts require managing these data. Virtual archives are a potential solution for many researchers. This project seeks to demonstrate the usefulness of this technology and document the workflow process. Early responses are very favorable and we have been approach by additional agencies wishing to examine the tools.

## 6. ACKNOWLEDGMENTS

I would like to convey a world of gratitude to the Odum Institute for Research in Social Science for the freedom to work on projects advancing digital archival research. I would also like to thank all our Data-PASS partners for assistance in the common goal of better social science data preservation and access. In addition, I would like to thank the Institute for Museums and Library Services for the funding of the deployment of virtual archives in a demonstration project for public opinion data collection centers. IMLS National Leadership Grant 2010: Award Number LG-07-10-0240-10

## 7. REFERENCES

[1] DVN, Dataverse Network Repository Software,www.thedata.org

[2] Green, A.G. and Gutmann, M. (2007). "Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives." *OCLC Systems & Services: International Digital Library Perspectives* 23:35-53.

[3] King, G., (2007), An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, *36*(2), 173-199. Retrieved March 21, 2008, from http://gking.harvard.edu/files/dvn.pdf

[4] NSF, National Science Foundation Data Management Plan requirements, Retrieved March 2011 from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[5] Parry, J., Kisida, B., and Langley, R. (2006). "What Would (Mac) Jewell Do? The State of State Polls" Presented at the Annual Meeting of the Southern Political Science Association Atlanta, Georgia January 2006. Accessed January 24, 2010 from http://www.allacademic.com/meta/p_mla_apa_research_citation/0/6/8/7/7/pages68776/p68776-1.php

[6] Pienta, A. M., Gutmann, M. P., Hoelter, L. F., Lyle, J. and Donakowski, D. (2008). "The LEADS Database at ICPSR: Identifying Important "At Risk" Social Science Data." Paper presented at the annual meeting of the American Sociological Association Annual Meeting, Sheraton Boston and the Boston Marriott Copley Place, Boston, MA. Accessed January 15, 2009 from http://www.allacademic.com/meta/p242699_index.html.

[7] Science (2011), Challenges and Opportunities, *Science 11* February 2011: Vol. 331 no. 6018 pp. 692-693 DOI: 10.1126/science.331.6018.692, Retrieved March 26, 2011 from http://www.sciencemag.org/content/331/6018/692.short