Building on the Work of Colleagues: A Moment of Reflection

by Jonathan D. Crabtree[1]

Abstract

The social science data community is fortunate to have a tremendous group of talented professionals. We work each day to build upon the ideals founded by those that came before us. This article recognizes the efforts of early IASSIST members whose pioneering efforts enabled our work today. In particular the work of Sue Dodd will be acknowledged. This article reflects on the many partnerships and data oriented projects the author has had the good fortune to be a part of over the last ten years in his work at the Odum Institute (Odum, 2014). Many of these projects are work performed under the aegis of the Data Preservation Alliance for the Social Sciences "Data-PASS" (Data-PASS, 2014). These projects are just a small subset of many achievements accomplished by the greater social science data community.

> Her work paved the way for the development of many tools...

Keywords: Collaborations, Digital Preservation, Data Management, Metadata, Catalog

A Personal Prologue

It is an honor to have the opportunity to reflect on the advancements our social science community has made in recent years toward metadata harmonization as well as the preservation of the materials we all hold so dear. I have had the pleasure of working for the Odum Institute, University of North Carolina UNC, for twenty-one years so it seems fitting for me to reflect on what has been accomplished as the Institute celebrates its 90th anniversary this year. I was fortunate that my service here at the Institute overlapped with the tenure of Sue Dodd, if only for a few years. The Institute has always provided a home for researchers and staff who share a passion for service to the social science community. Sue Dodd exemplified this ideal, and today as we build on her work, the Odum Institute Data Archive is dedicated to serving the social science community and its customers around the world who are seeking critically important data and information to support their research and data management services.

Foundations

Libraries and archives have been organizing information long before the advent of digital records. In early 1970 the investigation into a set of rules to catalog Machine-Readable Data Files, or "MRDF" began. (Dodd, 1982). Recognizing the unique properties of digital materials and having a keen eye towards both the potential challenges and affordances of cataloging these materials, Sue Dodd was instrumental in the evolution of cataloging standards for MRDF that first made their appearance in the second edition of AACR2 published in 1978 (Dodd, 1982). Her work paved the way for the development of many tools that simplify the discoverability, accessibility, and usability of vast amounts of social science data. Today's advancements would have been tremendously more difficult without the development of standard cataloging requirements and descriptive methodologies used to define these MRDFs.

My early work at Odum was in the information technology arena. I knew nothing of these early foundations and the valuable work of my new colleague Sue Dodd. I did not know that one day I would be tasked with the migration of thousands of catalog records from MARC format (MARC, 2014) to our current format the Data Documentation Initiative "DDI" (DDI, 2014) (Blank & Rasmussen, 2004). It was the thoughtful design and planning during the early days of MRDF catalog records that made my job much easier.

Little did I know at that time, the standardization of MRDF catalog records and the early efforts of the social science community to adopt and embed these nascent standards into their workflows have provided the bedrock upon which we build today's modern archive systems. Dodd asked in her writings "Where Do We Go From Here" (Dodd, 1982)? Ever prescient, she speculated that researchers and scholars would need these records to enable shared cataloging, authority control, acquisition systems, private file creations, products and a union list. As we know, these services and products we now take for granted are offered around the world today for a vast amount of social science data.

Behind this mountain of data is a network of researchers, archivists, librarians, information scientists, and administrators like Sue Dodd who work tirelessly to safeguard and provide access to valuable social science data that has helped to guide everything from public policy to education. We owe credit not only to Sue Dodd, but also to the whole of our international social science community for building these remarkable tools and services that continually add to the legacy of pioneers in our field. I value this opportunity to reflect on the enriching collaborations I have been involved with over the past ten years working to fulfill these earlier visions, and I encourage readers to do the same.

Building a Union Catalog

The Odum Institute was involved with one of the first projects following the founding of the National Digital Information Infrastructure and Preservation Program "NDIIPP" (Library of Congress, 2014). As part of the newly formed Data Preservation Alliance for the Social Sciences "Data-PASS" (Data-PASS, 2014) led by the Inter-university Consortium for Political and Social Research (ICPSR), we became a member of a voluntary partnership to archive, catalog and preserve valuable social science data that were at risk, in support of the NDIIPP agenda. The early Data-PASS partners – ICPSR, the Institute for Quantitative Social Science at Harvard (IQSS), the Odum Institute, the Roper Center, the National Archives and Records Administration (NARA), and the Murray Center -- were not strangers to one another. For many years, we had worked together on projects to provide access to quality social science data for our constituents. This familiarity, combined with a shared common goal, allowed the partnerships to grow and take root. Once Data-PASS was established, the group immediately began to survey the landscape and take action. By building on existing relationships, the Data-PASS partners were able to expedite the process. (Crabtree & Donakowski, 2006). The Data-PASS partners had four primary goals during the NDIIPP project: (1) archive at-risk social science content, (2) build a shared union catalog, (3) provide replicated preservation, and (4) advocate for best practices in digital preservation. We began to identify at-risk content almost immediately and developed strategies and best practices to manage this task. Jointly, we also began to develop a plan to take steps toward building the union catalog envisioned in the early days of MRDF catalog records.

Our strategy was to utilize standard harvesting methodologies like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, 2014) to collect metadata records from partner repositories. This would provide a standardized interface and allow the integration of existing and diverse technologies across the partnership into the new common catalog. We were very fortunate that our partners at Harvard IQSS (IQSS, 2014) were already developing open source archive technology that used these standards and thus had experience in this area. The Odum Institute took advantage of Harvard's success in building and implementing their virtual archiving platform and became one of the first outside implementers of what is today the sophisticated Dataverse Network, DVN (Crosas, 2011). Because the Data-PASS common catalog design was platform agnostic, partners not having adopted the DVN could still contribute to the catalog via a simple OAI-PMH interface server. This low barrier of technical entry was essential to the success of the partnership.

Of utmost importance to the success of the Data-PASS common catalog was the standardization of each partner's metadata—made feasible by the groundbreaking work of Sue Dodd many years earlier. The Odum Institute during this period was also migrating MARC records to the new DDI standard and looking for a replacement for the soon-to-be outdated version of the Stanford Public Information Retrieval System (SPIRES) database (SPIRES, 2014). This is not to say it was without challenges, but working with our partners at Harvard we were able to complete the migration of the metadata and ingest the contents of the Odum Archive into our newly built Dataverse Network.

Aside from the Odum Institute's collection, the Data-PASS partnership brought together a wide collection of social science catalog records from six major U.S.-based social science repositories and for the first time allowed discovery of these social science data from a single common catalog. The Dataverse Network has since expanded to include collections from all over the world and continues to grow daily. The power of a standardized metadata catalog record has been exploited to provide discoverability for a vast amount of social science data worldwide. Replicating and preserving the catalog records of our joint institutions was an important first step, but our partnership also sought to create a distributed preservation system that our partners could leverage to provide geographically distributed preservation for the group.

Collaboration for Preservation

The formation of the Data-PASS partnership established the groundwork for a distributed preservation project. Building on the success of the union catalog, the group identified the need to distribute our joint content in addition to our metadata as a means to better protect our data--despite disparities in repository size and resources among the partners. This is a challenge many organizations face. The expense of maintaining multiple machine rooms and backup systems in multiple geographic regions is prohibitive for small to mid-sized repositories. I would argue that it is equally a burden on larger repositories that would rather spend their ever shrinking resources in more fulfilling areas. Both of these circumstances were present among Data-PASS partners, which created the need for our preservation system to deal with the asymmetrical size of the collections (Altman *et al*. 2009).

Rather than reinvent the wheel, we decided to borrow from the work of other NDIIPP partners working in this space. The MetaArchive (MetaArchive, 2014) project had been working on defining Private LOCKSS Networks (PLN) to adopt solutions already implemented at Stanford University (LOCKSS, 2014). We were able to build our preservation network using tested strategies. We had additional challenges along the way due to our content types and sizes but by leveraging the work of fellow NDIIPP partners we were better prepared to tackle these challenges. The asymmetrical nature of our PLN layered additional challenges on top of our more distributed administration approach. Each partner had primary responsibility for running their independent LOCKSS node, and because we had no one central administrator for the network, it was essential that we developed a reporting structure that would generate audit reports of the network. These tools did not exist, so we sought additional funding to build auditing tools for our PLN.

Trust but Verify

Data-PASS members needed the ability to audit the new preservation network if it were to demonstrate compliance with standards for trustworthy repositories. The members all had diverse plans for preservation of content in place already, but the addition of a remote copy of each repository under the administration of other members is something that not only needed legal policies in place, but also the ability to audit the performance of the network. This prompted the design of the asymmetrical audit system prototype developed by Data-PASS during the NDIIPP project extensions (Altman *et al*., 2009). Follow up funding from the Institute for Museums and Library Services (IMLS) had allowed the prototype to mature into the current open-source offering, the SafeArchive Audit System (SafeArchive, 2012). Utilizing the TRAC audit framework (CRL, 2007) allows the SafeArchive to enable a PLN to define preservation policies in both qualitative and quantitative means. These user-defined policies are stored in a schematized XML format and used to compare the actual performance of the LOCKSS PLN to their policies. The result is an audit report that can be provided for each of the members on the status of their content as it compares to the preservation policies they have specified.

Data Management Services

It seems that today we are living in the "Age of Data Management Enlightenment." Everywhere you turn governments, funders, publishers and research institutions are seeking assistance for data sharing, data management, and data science (OSTP, 2014). As I reflect on my time here at the Odum Institute, I want to scream "Social Science Archives Already Do This!" When I calm down, I am thankful that the early work on MRDF has positioned the social science data community at the forefront of modern data management. Our community is comprised of many individuals like Sue Dodd who have the insight, ingenuity, and enthusiasm to contribute to new initiatives. Joint efforts to adopt a common metadata standard like the MRDF metadata grandchild, DDI, along with sophisticated approaches for handling confidential data and the experience of building partnership for preservation and access of complex data files, all provide a wealth of expertise as our society embraces open data policies (Data Transparency, 2013) and builds massive indexes of health-related data (NIH, 2014).

We should embrace new partnerships with libraries and library educators as they tackle the monumental task of managing a research output that is growing exponentially. The Odum Institute is currently working with the UNC School of Information and Library Science and the UNC Libraries in a joint effort to design data management curricula that are flexible enough to be delivered as online content via a Massive Open Online Course "MOOC" yet grounded enough to allow students to develop local support networks within their own institutions. As a result of this new Curating Research Assets and Data using Lifecycle Education (CRADLE, 2014) grant, we hope students will share their new knowledge and experiences as they enhance their local data management networks.

The international social science data community is graced with many great organizations that are working to educate researchers on proper data management practices. The current data-sharing climate has prompted the research community to seek these services around the world. Journal publishers are encouraging and in some cases are requiring authors to submit data supporting their findings alongside their manuscripts. This push toward such a replication data requirement will provide a solid foundation for future scientific discovery as new research is designed around previous discoveries. The Dataverse Network is working with journal publishers to help satisfy this new requirement. Efforts like the Open Source Journal (OSJ) deposit API for the Dataverse seek to streamline and simplify this process for the authors and publishers (OJS, 2014).

Where Do We Go From Here: Hello, Big Data

In the spirit of the 1982 Dodd manual (Dodd, 1982), "Which direction do we go from here?" I hereby declare that the social science data community has come a long way in standardizing data descriptions to make data accessible and understandable for secondary use. Pausing to reflect on our past is indeed a worthy exercise, but we should not rest in our efforts to seek improvements for managing the growing collections of data under our stewardship. Sue Dodd would not be surprised that today the data we are entrusted to are increasingly larger and more diverse than those that came before them. The need for tools and services to visualize and analyze new data types has never been greater. Social science is becoming more and more interdisciplinary, and the community will be facing more complex and larger data types like those used in social network analysis and mixed methods studies. Relationships between social science datasets will become increasingly complex and require a complex object model to describe. This is not a revolutionary notion, and new standards like DDI version 3 are already designed to handle these relationships (DDI, 2014). The challenge will be to integrate these new models into large preexisting relationship among data sets within archives.

As the sheer volume of research data becomes much more massive, we will be forced to seek the council of those in other disciplines that have become accustomed to handling dataset in the petabyte range. The Odum Institute has begun working with the Data Intensive Cyber Environment "DICE" group to begin leveraging the iRODS (iRODS, 2014) rules-based grid system. Tools like iRODS that have the ability to manage multi-petabyte collections and apply active policies will be needed as we begin embracing the new and larger data formats in the future. We have initiated the integration of the Dataverse Network and iRODS that seeks to provide data archiving at scale and allow the federated Dataverse Network access to discover the massive amounts of data existing in data grids around the world. Through our work on the National Science Foundation DataNet Data Federation Consortium project we hope to link diverse communities of data users ranging from oceanographic and hydrologic disciplines to temporal dynamics and plant genomics communities (DFC, 2012).

As social science researchers are encouraged or required to share their data, we must always remember our dedication to protecting human subjects. This will require archives to provide tools to assist in this process. The Odum Institute is closely monitoring the progress of and learning from projects like the Data Privacy Center at Harvard's Data Tags initiative, which will be critical in providing new tools to share these data while protecting our human subjects (Privacy Tools, 2014). We should also seek to partner with computer science and data science initiatives like the National Consortium for Data Science (NCDS, 2014) to better understand our security risks and provide input into the next generation of secure data transmission systems.

Data volumes are almost guaranteed to increase exponentially into the future. The social science data community will need to leverage as much as possible automated metadata generation technologies to help reduce the burden on depositors and archive staff. Automated ingest tools that create variable level metadata are already being deployed in tools such as the Dataverse Network. Projects such as the NSF-funded DataBridge project (Rajasekar et al., 2013) seeks to use sociometric analysis techniques used in social networking to help determine relationship between users, data, and methods. These relationships could be used to produce multilevel object relationship models to aid in data discovery and population of DDI 3 object relationship models. Tools like these, combined with advanced commercial indexing of datasets, will be important to the sustainability of data sharing.

If the Odum Institute and other organizations dealing with data are to contribute to Sue Dodd's legacy, we must recognize that the complex problems we contend with today often warrant complex solutions. These are solutions that likely cannot be generated by any one individual or organization alone. Members of the social science data community must be willing to reach out beyond their own walls to forge partnerships that take full advantage of the vast amounts of talent that are dispersed throughout our community. To answer Sue Dodd's question today, "Where do we go from here?" I would suggest that "wherever we go, we go together."

Conclusion

The social science data community has made great advancements over the years since Sue Dodd and others began defining bibliographic control over computer information in the late 1970s. I have been fortunate during the past ten years to work with wonderful collaborators and colleagues to build on the work of early IASSIST members.

Our community has been fortunate to have strong foundations that have placed us ahead of the game. The social sciences are becoming increasingly interdisciplinary, and we should make every effort to help other disciplines that could learn from our experiences. Sharing knowledge will enhance our ability to deliver quality data management and archiving to the diverse social science researchers we will encounter in the near future. Building new relationships takes valuable time and effort, but the rewards are great. We have a wonderful data community and we should promote open exchange of knowledge to other disciplines.

Social science data specialists are seeing the demand for our assistance increase exponentially. Our workflows are becoming increasingly complex with the introduction of innovative data formats and expanding data sizes. Today's modern services and tools for managing the outputs from social science research are grounded in the early works of IASSIST members like Sue Dodd. Without these tools, we would not be equipped to handle our growing set of responsibilities as data stewards. New challenges for the social science data community evolve everyday. As we design services to address these needs, we should encourage new collaborations, encourage open exchange of knowledge, and build on past experiences. The social science data community has tremendous knowledge and experience in its ranks. We should share these with the world.

Acknowledgements

I first would like to thank Peggy Adams and Libbie Stephenson for asking me to reflect on some of the current projects that are building on the early work of Sue Dodd. It would be overwhelming to mention all of the exceptional colleagues I have worked with on these projects but I feel compelled to mention a few. An overwhelming thank you goes to the Odum Institute that has encouraged me to build these relationships. Working with the world class Odum Institute Archives and Information Science staff has been a great pleasure. Dr. Kenneth Bollen and Dr. Myron Gutmann allowed me to work with the fabulous partners in the early days of Data-PASS. I owe a tremendous debt to these two great researchers and dear friends. I also would like to thank the current Odum Institute director Dr. Thomas Carsey for his support and assistance as we continue to support archive development and data management services. A personal thank you also goes to our Deputy Director Peter Leousis and our current Data Archivist Thu-Mai Christian for all the support.

The projects I have described are the products of many great relationships and wonderful partner organizations. Without the collaborations within the Data-PASS partners few of these projects would have been possible. Many other great partners across the social science data community also deserve mention: Educopia, California Digital Library, LOCKSS, DICE, RENCI, UNC School of Information and Library Science, UNC Libraries, and the Australian Data Archive have been valuable partners in our efforts.

None of this would have been possible without great funders. Many thanks go to the Library of Congress, the Institute for Museums and Library Services, the National Science Foundation, the National Institutes for Health, and the Sloan Foundation. Finally, I wish to thank the wonderful IASSIST community. We have a great history and community on which to build.

References

Altman, et. al (2009). Altman, M., Beecher, B., Crabtree, J., Andreev, L., Bachman, E., Buchbinder, A., Burling, S., King, P., & Maynard, M. "A Prototype Platform for Policy-Based Archival Replication" Against the Grain 21(2). Forthcoming.

Blank, Grant, and Karsten Boye Rasmussen (2004). "The Data Documentation Initiative: The Value and Significance of a Worldwide Standard." Social Science Computer Review (22): 307–318. doi:10.1177/0894439304263144.

Crabtree, Jonathan and Darrell Donakowski (2006). "Building Relationships: 'A Foundation for Digital Archives.'" Accessed on February 6, 2012, <http://www.ils.unc.edu/tibbo/JCDL2006/Crabtree-JCDLWorkshop2006.pdf>.

CRADLE (2014). "Dr. Helen Tibbo Receives IMLS Grant for CRADLE Project | Sils.unc.edu." Accessed February 2, 2014, <http://sils.unc.edu/news/2013/tibbo-odum-imls-cradle>.

CRL (2007). Center for Research Libraries (CRL), OCLC. "Trustworthy repositories audit & certification: Criteria and checklist." <http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf> .

Crosas, Mercè (2011). "The Dataverse Network®: An Open-source Application for Sharing, Discovering and Preserving Data." D-Lib Magazine 17 (1/2) (February). doi:10.1045/january2011-crosas. <http://www.dlib.org/dlib/january11/crosas/01crosas.html> .

Data-PASS (2014). "About Data-Pass | Data-Pass." Accessed February 2, 2014, <http://www.data-pass.org/>.

Data Transparency (2013). "The Future of Open Data Policy." Accessed February 2, 2014, <http://www.datatransparency2013.com/>.

DDI (2014). "Welcome to the Data Documentation Initiative | DDI - Data Documentation Initiative." Accessed February 2, 2014, <http://www.ddialliance.org/>.

DFC (2012). "DataNet Federation Consortium." Accessed June 12, 2012, <http://datafed.org/>.

Dodd, Sue A. (1982). CATALOGING MACHINE-READABLE FILES: AN INTERPRETATIVE MANUAL. Chicago: American Library Association.

IQSS (2014). Accessed February 2, 2014, <http://www.iq.harvard.edu/>.

iRODS (2014). Accessed February 2, 2014, <http://www.irods.org>

Library of Congress (2014). "Digital Preservation (Library of Congress)." Accessed February 2, 2014, <http://www.digitalpreservation.gov/index.php>

LOCKSS (2014). Accessed February 2, 2014, <http://www.lockss.org/pln/>

MARC (2014). "MARC STANDARDS (Network Development and MARC Standards Office, Library of Congress)." Accessed February 2, 2014, <http://www.loc.gov/marc/>

MetaArchive (2014). Accessed February 2, 2014, <http://www.metaarchive.org/>.

NCDS (2014). "National Consortium for Data Science." Accessed February 2, 2014, <http://data2discovery.org/>.

NIH (2014). "RFA-HL-14-031: Development of an NIH Data Discovery Index Coordination Consortium (U24)." Accessed February 2, 2014. <http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-14-031.html>.

OAI-PMH (2014). "Open Archives Initiative Protocol for Metadata Harvesting." Accessed February 2, 2014, <http://www.openarchives.org/pmh/>.

Odum (2014). "The Odum Institute: Advancing Social Science Teaching and Research." Accessed February 2, 2014, <http://www.odum.unc.edu/odum/home2.jsp.>

OJS (2014). "And So It Begins: OJS Dataverse Plugin Testing | PKP-Dataverse Integration Project." Accessed February 2, 2014, <http://projects.iq.harvard.edu/ojs-dvn/blog/and-so-it-begins-ojs-dataverse-plugin-testing>.

OSTP (2014). "OSTP Announces Plans to Increase Access to Federally Funded Research." Accessed February 2, 2014, <http://www.aera.net/ResearchPolicyAdvocacy/AERAandOpenAccess/OSTPAnnouncesPlanstoIncreaseAccesstoFederal/tabid/14784/Default.aspx>.

PLN (2014). "Private LOCKSS Networks." Accessed February 2, 2014. <http://www.lockss.org/pln/>.

Privacy Tools (2014). "Project Description | Privacy Tools for Sharing Research Data." Accessed February 2, 2014, <http://privacytools.seas.harvard.edu/project-description>.

Rajasekar, et.al. (2013), Rajasekar, A, H. Kum, M. Crosas, J. Crabtree, S. Sankaran, H. Lander, T. Carsey, G. King, and J. Zhan. . "The DataBridge," Science Journal. ASE (in press).

SafeArchive (2012). Accessed May 17, 2012, <http://www.safearchive.org/>.

SPIRES (2014). "About SPIRES." Accessed February 2, 2014, <http://www.slac.stanford.edu/spires/about/>.

Notes

1. Jonathan D. Crabtree is the Assistant Director for Information Technology and Archival Research for the Odum Institute for Research in Social Science at UNC Chapel Hill and can be reached at Jonathan_Crabtree@unc.edu.