

Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives

Ann Green and Myron Gutmann¹

Please do not cite or quote without permission. This paper is currently under review for publication..

Introduction

The digital library world has recently developed and debated both the concept and the implementation of institutional digital repositories. It has spent less time discussing discipline- or domain-specific digital repositories. Such repositories have been in existence for many years in the social sciences and have generated important lessons about the long-term preservation and sharing of academic work. The goal of this paper is to compare these two kinds of repositories and to suggest ways that they can help build partnerships among themselves and with the research community. All the parties share important goals and, by working together, can advance these goals.

As authors, we emphasize the role of these repositories in social science research, and we bring our own experiences and perspectives in dealing with digital preservation of research products generated by social science. We have both been involved in the leadership (Gutmann as Director, and Green as Chair of the governing Council) of the Inter-university Consortium for Political and Social Research (ICPSR)^[2], a consortium of 550 institutions worldwide that serves as the long-term steward and a primary channel for sharing a vast archive of social science data. At the same time, Green has worked extensively as a digital information specialist, while Gutmann continues to be an active researcher who has everyday contact with others in the social science fields (historical population studies) where he has worked for the past thirty years. ICPSR and other data libraries and data archives (throughout the US and worldwide) provide essential social science infrastructure, including online access to key datasets, stewardship and preservation commitments for long-term access to data, training in statistical methodology, and a range of support services for and by an international network of researchers, students, data support professionals, and data providers. Because of our experience and the complexity of trying to deal with the international situation, we focus most of our attention on digital repository development in the United States.

Acknowledgements: We are especially grateful to Cole Whiteman for helping us think through the relationships explained in the paper, and then providing the illustrations, and to Ruth Shamraj for editorial consultation. We also wish to thank Chuck Humphrey, Julie Linden, Caroline Arms, Jonathan Crabtree, Nancy McGovern, Ron Jantz, Linda Detterman, Diane Geraci, and Robin Rice, who read the paper in draft form and helped us clarify and improve it.

We begin by describing the life cycle of social science research. Next we turn to some of the key elements of institutional versus domain-specific repositories, before concluding with recommendations about ways that researchers and those operating the two different kinds of digital repositories can forge partnerships.

The Social Science Research Life Cycle

Many authors have described a life cycle through which social science research proceeds, and the simplified model we propose here does not differ materially from that of others.^[3] What this model allows us to do is to identify important stages in social science research, and by doing so point out times when the researcher may benefit from interacting with either institutional repositories or domain-specific repositories. We believe that constructive relationships between researchers and both types of repositories, as well as between the repositories themselves, will have genuinely valuable results for all concerned.

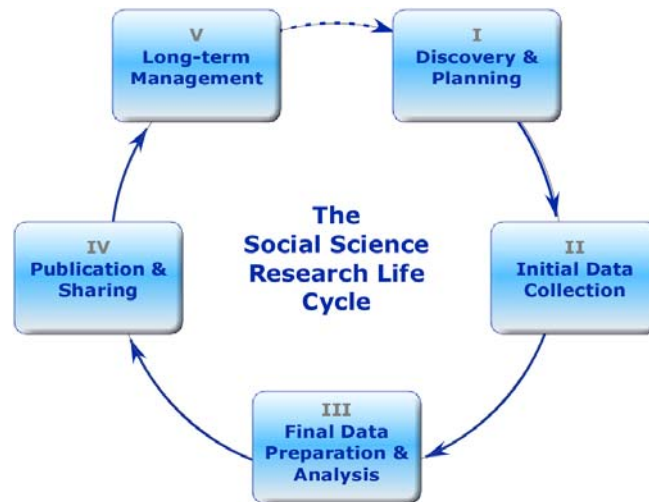


Figure 1. Social Science Research Life Cycle

- **Discovery and Planning.** The first step of the research process is one of discovery and planning. Through a theoretical and empirical perspective, the researcher seeks ways to enhance knowledge in her or his field. This is the stage where the researcher explores the possibility of using existing data (hence linking to archived data at the other end of the life cycle), and where she or he determines whether new data must be collected in order to best answer the scientific question. If the project requires external funding, the researcher writes a proposal for funds. Given requirements at the National Science Foundation^[4] and the National Institutes of Health^[5] for data sharing plans, this is also the stage where the researcher enters into discussions with digital repositories about steps required to ensure proper data sharing later in the project. Costs related to archiving and special considerations for data sharing, such as informed consent and confidentiality concerns, should also be featured in this stage.

- **Initial Data Collection.** Once the project has been designed and funding secured, it is time for the researcher to collect data. In a survey or an experiment, new data are collected from respondents or experimental subjects. In a project that makes use of previously collected data, those files containing the data are acquired and potentially reformatted or linked to other useful data. This is also the stage at which essential data management strategies are formulated and instituted, including decisions about documentation content and formats. At the end of this step, a clean initial dataset is in place. To ensure that this dataset is available for her own or others' future research, the researcher should secure the long-term preservation of these preliminary data. This is also the time to inform the research community about the structure and aims of the project by providing high-level metadata that can be located within domain-specific research aids. The delivery of this high-level metadata should trigger a discussion between the researcher and a domain-specific repository about the format required for effective sharing of the data.
- **Final Data Preparation and Analysis.** The third stage in the process occurs when the researcher is performing final verification and modification of the data, undertaking analysis, and beginning to write results. At the end of this stage, the process of data preparation is complete and a copy of the final dataset should be preserved, if only at the researcher's institutional repository.
- **Publication and Sharing.** The fourth stage for the researcher consists of communicating research findings. When publications appear, the sharing of data and publications with the broader research community is triggered. By the time this stage is well underway, the intellectual productivity of the project is fully demonstrated, and the researcher would normally share her data with the community by making data and full metadata discoverable and available through a domain repository.
- **Long-Term Management.** During the final stage the repository community has two critical goals -- first to ensure that the data and other intellectual products are exposed for use and learning by others, and second to ensure long-term preservation. As more time passes, and the ability of the original researcher to manage any part of the relationship with secondary data users diminishes, more and more of the sharing and preservation activity resides with the repositories. Once exposed for secondary use, data that have reached the final stage of the research cycle become the seeds of new projects that begin with their own discovery and planning, thus beginning the cycle anew.

Digital Repository Types

The current digital repository landscape is made up of a blend of repository types. For the purposes of this discussion about social science data and digital repositories, we have grouped repositories into these two broad categories:

- Institutional digital repositories with no specific discipline focus. By definition, these are found at academic institutions, and have goals of preserving and making available some portion of the academic work of their students, faculty, and staff.
- *Discipline* or *domain*-specific data archives. All of these institutions currently share the attribute of focusing on data preservation and sharing. These include Social science data libraries maintained by a single institution, broad based social science data archives such as ICPSR or the Roper Center for Public Opinion Research, and topic-specific domain-focused data libraries and archives, such as Harvard's Henry A. Murray Research Archive or Princeton's Cultural Policy and the Arts National Data Archive (CPANDA).[⁶]

These repositories vary widely in their missions and roles in supporting the different stages of the research life cycle. Clarifying those missions and goals helps to understand the potential relationships, dependencies, and overlaps among repositories. It also helps to identify gaps and anticipate needs.

Institution-Based Repositories

The majority of institution-based repositories currently planned or already in operation emphasize the collection, cataloging, and sharing of journal-type articles and monographs. Their goal is to ensure that the scientific production of their faculties and scientific staffs are collected in a single location, one that will ensure their availability and bring credit to the institution. While not always explicitly stated, another repository goal is to reduce the institutional or community-wide cost of scientific dissemination by drawing that role back from scientific publishers and keeping it closer to the producers of scientific output. Moreover, many eprint digital repositories strive to supplement or even replace the peer-reviewed scholarly journal publication process. (Cervone, 2004) Long-term sustainability is an important component of the institution-based repository movement: university libraries (where most are located) have a long record of sustainably preserving and delivering research materials. Institution-based repositories have emerged from financially stable and dependable organizations.

Most repositories hope to be more than a stockpile of eprints. They seek to provide safe harbors for a more inclusive interpretation of the intellectual output of local faculty-driven research and teaching, by including pre- and post-prints, working papers, research reports, datasets, course materials, personal image collections, among other types of content. As with eprint collections, these repositories retain institutional identity, but may choose to limit access to specific sub-communities within the institution instead of public "webwide" open access.

Institution-based repository collections primarily grow through voluntary deposit, but in some cases collections are developed by selection criteria through the assistance of professional librarians, such as the work done by the Harvard Science Libraries.[⁷] Domain-specific collections may be developed within the overall repository structure, but support services for data processing, metadata production, or analysis are not usually offered as part of the repository service. Where disciplinary focus exists, it is more likely

that specialized units within institutional libraries typically are more able to work with and support faculty in their areas of interest.

Whatever the collection policy they develop, institutional repositories attempt to limit the amount of effort that repository managers have to spend to acquire any single item by developing automated or nearly-automated deposit processes. Services are frequently quite basic and user-friendly: authors can *put* (place content in the repository), users can *search* and *get*, all within a framework of *access control* defined either by the author, the repository, or both. Intellectual property rights and copyright clearances are part of the deposit process, often negotiated at the level of the institution, department, or individual. With these procedures, faculty who contribute to the institutional repository upload content and define the high-level metadata describing their submission. Little detailed metadata are called for, and validation of metadata and content is restricted to checking file formats and ensuring that all required high-level metadata fields have been provided. Knowing the author, affiliation, title of the work and employing a file format (ASCII text, Microsoft Word, Adobe PDF, etc.) that complies with policies adopted for the institutional repository are often all that is required.

Given this approach and their emphasis on capturing final or near-final forms of scholarly productivity, these repositories position themselves at or near the end of the scientific research life cycle. Their goal is less to partner with researchers or with domain-specific repositories throughout the research life cycle than it is to garner the value of the institution's productivity, to gather this productivity, and possibly to lower the local or community-wide cost of scholarly publications.

The success of institutional repositories, when measured by the size and use of the collection, depends upon faculty participation and their willingness and ability to contribute digital materials into open access environments. Based upon data from a JISC sponsored survey of authors, Steven Harnard (2006) reports that unless deposits in repositories are mandated by institutions and funding agencies, submission rates will remain low.

Domain-Specific Digital Repositories

Domain-specific repositories share important goals with institution-based repositories, including the objective to provide access to research materials. At the same time, there are significant differences between the two types of digital repositories. Rather than focusing on publication-related materials from multiple subjects areas within a single organization, domain-specific digital repositories hold collections of materials grouped by type, subject, or purpose and intrinsically support domain- or discipline-oriented research needs. Domain-specific digital repositories in the social sciences have a history of providing infrastructure for data sharing and strive to provide support throughout the data life cycle.

These data archives hold the raw materials that faculty and students can reuse, repurpose, analyze, and recompile in teaching, learning, and research environments. Part of an

international network of data archives and libraries, they share a mission to acquire and manage social science data (both quantitative and qualitative) and provide support for data users. The 2005 report from the U.S. National Science Board on Long-Lived Data Collections (Recommendation 3, p. 6) emphasizes the community proxy role that domain-specific repositories play in contrast to the heterogeneous roles of institutional repositories. Domain specific repositories act and speak on behalf of their designated communities. As part of their key missions, they seek to know what the community wants and expects in terms of content, format, delivery options, support, and training.

Domain-based digital repositories in the social sciences share with institution-based repositories the goal of contributing to a common infrastructure that seeks to blur boundaries that divide types of research sources, institutions, researchers, disciplinary domains, geographic borders, and funding constraints. This is in line with how faculty researchers think of their academic community orientation. As Cliff Lynch (2003) has written, rather than having strictly an institutional orientation, faculty “often don’t stay at a single institution for their entire career, and they frequently disregard institutional boundaries when collaborating with other scholars. Federation of institutional repositories may also subsume the development of arrangements that recognize and facilitate faculty mobility and cross-institutional collaborations.”

Recently, the boundary has blurred between the domain-specific social science archive as data repository and other institutional repositories of published reports about data and research results. For many years, important single-source data disseminators such as the General Social Survey, the Panel Study on Income Dynamics, the Health and Retirement Survey, and the University of Minnesota’s Integrated Public Use Microdata Series of Census data have maintained their own bibliographies of publications related to their data. Since the early 2000s, ICPSR has systematically compiled bibliographies of books, chapters, articles, and other publications that make use of the items in its collection; that bibliography now numbers over 40,000 items, with many of them available as links to full-text publications. This fuller integration of resources in the domain repositories has increasingly positioned them as research aids that are among the most commonly used by scholars. Combined with more general tools like Google Scholar, these combinations open up new seamless points of access for researchers.

The social science domain repositories have also moved toward partnership across archives, leading to new opportunities for improvements in research efficiency. Beginning in 2004, and as part of the Library of Congress’s National Digital Information Infrastructure Preservation Program (NDIIPP), a group of the largest social science repositories came together to form the Data Preservation Alliance for the Social Sciences. This Alliance includes ICPSR, the Roper Center, the Odum Institute for Research in Social Science, the Murray Research Archive, the Harvard-MIT Data Center, and the US National Archives and Records Administration.^[8] These collections contain data that are widely used by researchers around the world, connected both by the emerging DataPASS infrastructure and by well-used guides to data such as that prepared by the University of California, San Diego.^[9] This parallels the European situation, where the

Council of European Social Science Data Archives^[10] maintains a common catalog and policies for cross-border data sharing.

Resource discovery in the social sciences now extends far beyond consulting a stand-alone research aid or search tool. We now see alliances across repositories and the emergence of a new set of tools allowing researchers to do complex and innovative searches to locate and explore data. The possibility of cross-database and cross-site searching has been made much easier by the emergence of a standard XML-based markup language for social science metadata called the Data Documentation Initiative, or DDI ^[11], now in its third version. The existence of common metadata markup has enhanced the development of software tools for exploratory data search and analysis, including Berkeley's Survey Data and Analysis ^[12], Harvard's Virtual Data Center ^[13], which will drive the U.S. DataPASS partnership's common catalog, and the European Nesstar ^[14], which drives the European common catalog. These tools make it possible for a student or professional researcher to find data, examine them, and do varying kinds of analysis without having to download the data and load them into a software package like SPSS, Stata, or SAS.

Unlike social science digital repositories, most institution-based repositories are not oriented to deal with research data, even though data files can be deposited in them. Some have been set up to accept and actively pursue diverse collections, but are not charged with providing discipline-focused environments or statistical services for preparing or using the data. (Lagoze, et al 2005).

The practice of "self-archiving" at many institutional repositories has the potential to pose difficulties for effective long-term sharing and preservation of social science research data. (Humphrey, 2005) In many cases, the depositing mechanisms have been made so user friendly and generic that they are inadequate for the demands of preparing and processing data for secondary use. Producing adequate data documentation demands significant amounts of time and labor that most researchers do not have. While work is under way to develop standards for archive-ready datasets, and while ICPSR and some of the other large data archives have publications and services designed to assist researchers in data preparation, in most cases dissemination-ready data are created through partnerships among data producers, data analysts, data archivists, and technicians. Moreover, these processes take place over several stages in the research life cycle, and not merely at the end.

Institutional repositories are also limited in their ability to preserve and manage social science data over time. The level of long-term support for different kinds of content is an important issue for potential depositors and users of social science data. The forty-plus year experience of social science data archiving reveals that, while some core formats for datasets have persisted over time, many formats have not. (Green, Dionne, Dennis 1999) Even where formats have persisted, standards of metadata preparation and file organization have changed dramatically. As part of its commitment to keep data formats current, ICPSR has revised and reissued files from 250 older studies in the past two years alone; this is more than half the number of new studies it acquired during this period.

Most institutional repositories do not and cannot offer support for managing dataset formats over time. Few have policies or commitments in place that would allow them to claim they have extensive data archiving capabilities, even though repository software could potentially support their preservation management of datasets.^[15] Policies for long-term stewardship vary among institutions, but many have developed a sliding scale of preservation promises. For example, the Duke University Digital Archive has developed three basic tiers of custodianship based upon the assessed value of the material.^[16] A similar tiered strategy from the Massachusetts Institute of Technology is based upon file formats.^[17]

The creation of institutional repositories is an important and valuable development for university-based researchers, who now increasingly have mechanisms for preserving and sharing the results of their work, both inside their institutions and with the larger scientific world. For now, repositories are more interested in collecting materials near the end of the research life cycle, employing an acquisition process that is simple and a preservation process that is not designed to support complex content. These limitations may decrease over time as the institution-based repository movement matures, but now is the time to begin developing successful partnerships with domain-specific depositories.

Partnerships among social science researchers, institutional repositories, and social science data archives

The complex landscape of institution-based repositories and domain-specific digital repositories can lead to lost data, missed opportunities, and competition for scarce resources. This network of repositories is confusing to most social science researchers and often can only be navigated by the most experienced professionals. We need to clarify intersecting nodes of interest, activity, and mission. Different and shared roles need to be developed to build partnerships between repositories that will lead to consistent support of data throughout the research process. We are not alone in making these assertions. The NSB report on Long-lived Data Collections (2005, p 18-21) urges all those involved in the world of data to “act collectively to pursue some of the higher level goals important to the entire field.” The 2006 report by the National Science Foundation’s Cyberinfrastructure Council (p. 20) also encourages “the establishment of strong, reciprocal, international, interagency and public-private partnerships’ to ensure the stewardship of valuable data assets.

A series of questions arise: How can data archives, social science researchers, and institutional repositories best work together to improve the digital landscape supporting social science research? What tools and standards, policies, support, hand-offs are needed from one role to the next? What is being passed along from researcher to repository to archive, and what tools are needed to enhance those activities and improve data quality and access? What specific partnership arrangements could help channel academic resources into long-term access and preservation environments?

In short, we advocate the creation of partnerships that support digital life-cycle management, policies, tools, and use of best practices and standards, and that include all

potential stakeholders. We propose a layered approach with ongoing conversations between researchers, institutional repositories, and the domain-specific data archives.

Partnership roles and activities during the research life cycle

Social science research programs in the U.S. that produce digital output rarely receive funding for sharing or preserving data, descriptive research materials, or laboratory results, which are the “raw materials” of research projects. Most institutional repositories do not offer social science metadata production or management tools, nor do they have the resources to address issues of confidentiality or assist with preparing datasets that need to be anonymized and prepared for use by researchers outside the initial data production and research team. On the other hand, domain repositories can be physically remote from the research enterprise, and their standards and practices may be difficult for the individual researcher to learn about. The proximity of institutional repositories to principle investigators could put partners on the ground to provide essential help in moving social science data into safe repository environments.

Partnerships among digital repositories will establish communication flows to make domain-specific support and expertise available during the full data life cycle. Early partnerships that match the skills and knowledge of the data producer (data production) with those of the repository (data life cycle management expertise and long-term curation planning) can have significant impacts: efforts made in the data production stages will reap long-term benefits in the publishing, reuse, and archiving stages. Informed selection of file formats and metadata standards at the creation of digital resources can increase both short- and long-term benefits. It is necessary to provide tools and processes that make best practices attractive and cost effective at the design and production phases of the data life cycle.

The next section examines how such partnerships might operate during the stages of the research life cycle. The figures for each section illustrate potential ‘partnering moments’ among stakeholders. The Researcher (or research group) is shown as part of the larger research community. Solid lines designate original flows of information; dashed lines designate secondary flows of information.

Discovery and Planning

Initial conversations regarding the agenda, outcomes, and support requirements of data intensive research should take place at the earliest stages of a research project. Information from those conversations can then be passed from principal investigators to the repository infrastructure layers. Research descriptions will serve to alert the research community about the project and also put a placeholder for the emerging research output in institution-based and domain-specific information systems. (See Figure 2.) We suggest that institutional repositories might explore the possibility of capturing contextual metadata (the documents and core materials for research projects) at this very early stage in the life cycle.

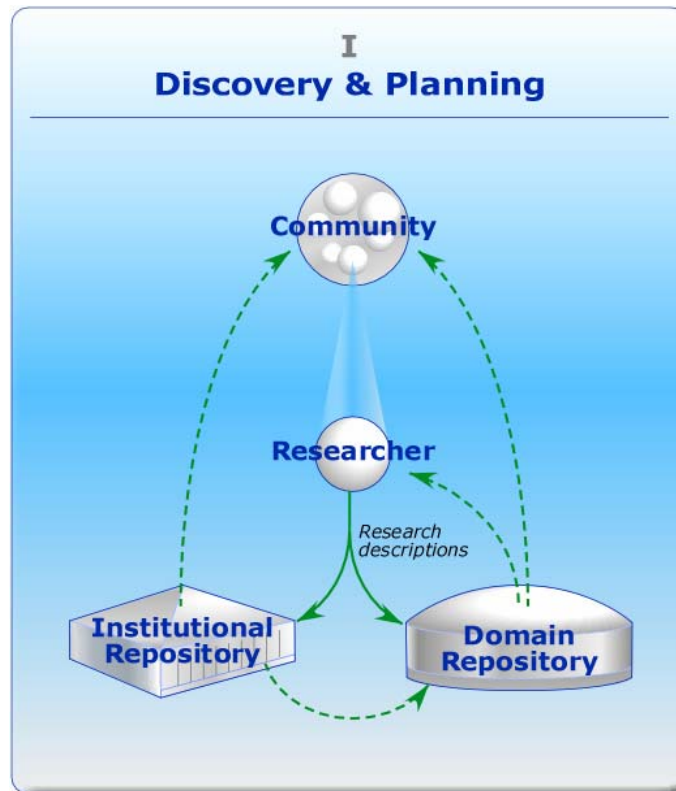


Figure 2. Research Descriptions are passed from Researcher to Repositories and on to the Research Community

Even at these early stages, research projects can benefit from conversations with repository experts about intellectual property issues, long-term digital preservation planning, access controls, confidentiality considerations, file format options, and metadata standards.^[18] Grant proposals and subsequent fiscal planning ought to include resources in their budgets to cover the costs of data preparation and archiving. Articulating those requirements and strategies can trigger connections among researchers and repositories and demonstrate the utility of channels of support for principal investigators. (See Figure 3.) Researchers need to know that support, standards, advice, and tools are available for content management and metadata production at this first stage of the research life cycle.

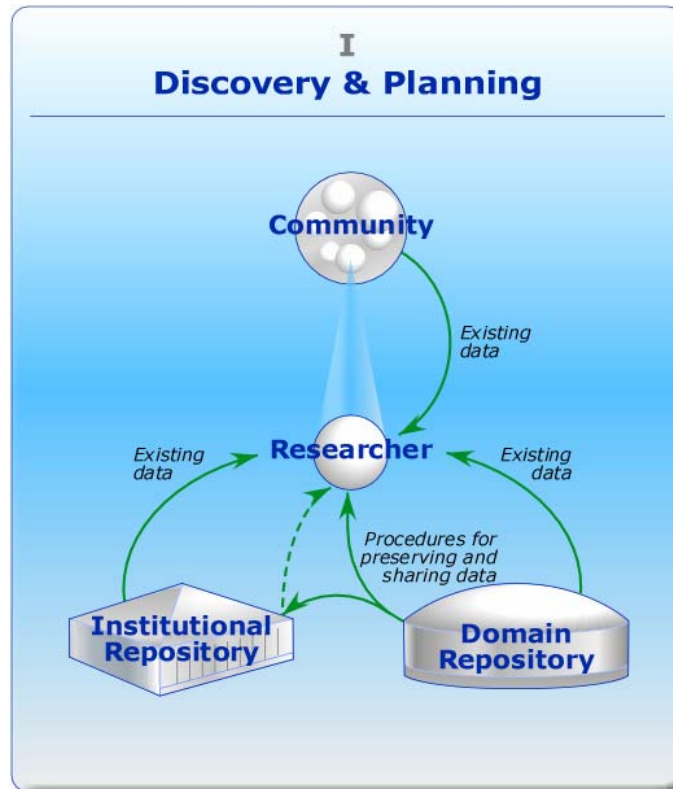


Figure 3. Existing data are made available to Researchers from the Research Community and from Digital Repositories. Procedures for preserving and sharing data are passed from the Domain Repository to the Institutional Repository and Researcher.

Initial Data Collection

During the initial data collection and processing phase of the research cycle, researchers may choose to make a preliminary transfer of data and documentation into institution-based repositories for safe keeping and early data sharing among collaborators. Local institution-based repositories could function as research workspaces with backup services, data management, and processing support.

High level metadata (descriptions of the project goals, funding, methodologies, etc.) can be produced and passed from researchers to local repositories when initial data collection efforts get underway. Also, institutional repositories can expose information about the research project and its data collection activity to a federated repositories structure in standard formats, for example OAI-PMH.^[19]

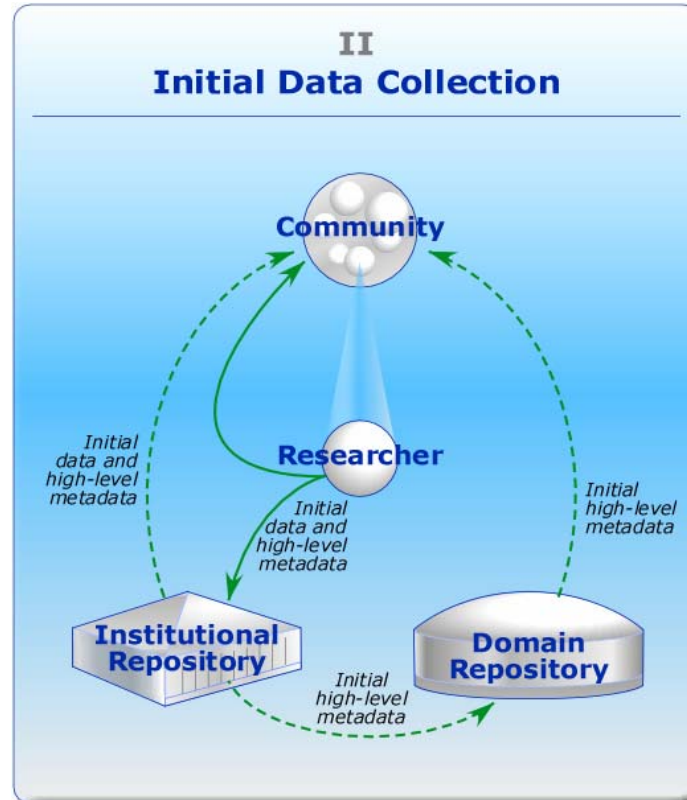


Figure 4. Initial data and high-level metadata are passed from the Researcher to the Institutional Repository and Research Community. Initial high-level metadata are passed from the Researcher to the Institutional Repository, and on to the Domain Repository.

The transfers of data from researchers to local repositories make it possible to trigger a local discussion about data processing requirements and what the research team needs to know in order to deposit data in institutional repositories and with the later possibility of passing archive-ready versions of data to domain-specific repositories. These discussions could include relevant information about processing standards and archive-ready submission requirements. Domain-specific repositories could design and provide local repositories with guidelines and strategies regarding an initial data submission package and templates for “archive-ready” dataset production.^[20] This is also a moment at which work on confidentiality issues should begin, with the support of domain-specific repositories.

Final Data Preparation and Analysis

As a research project moves into the final stages of data preparation and analysis, it is critical to provide tools and support for long-term sharing and archiving. This is the moment that researchers would benefit from mechanisms that:

- 1) Inform the social science community about data sharing opportunities and assistance with preparing and releasing an announcement of the characteristics of the initial dataset and metadata (instruments, methodologies, working papers, etc). This information could be pushed from institutional repositories into domain-specific repositories for integration into the domain-specific knowledge base.
- 2) Provide researchers with guidelines for data processing and metadata production, confidentiality review, and other requirements for later stages of metadata exposure and data sharing. (See Figure 5.)

Tools and services could become part of the support system to which institution-based repositories refer their constituents, thus providing a link between the developing domain-based services and on-the-ground researchers. According to a study at the University of California at Santa Barbara, faculty consistently reported having difficulties in managing research information. Since tools are regarded as extremely task specific and not terribly relevant outside a specific research field, faculty stressed the importance of getting the right tool to the right users at the right time. (Pritchard, Carver, Anand, 2004)

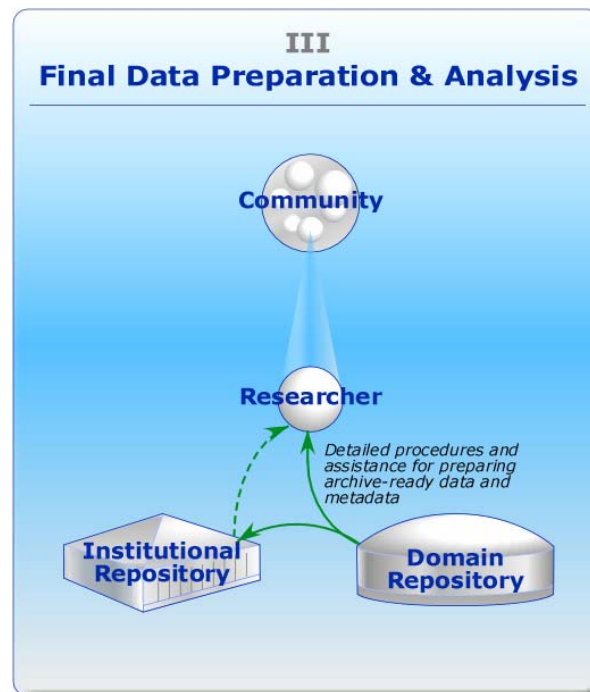


Figure 5. Detailed procedures and assistance are provided by the Domain Repository to Researchers and Institutional Repositories.

Publication and Sharing

At the publication and sharing stage, the results of the research project have reached the point of full-scale sharing of “archive-ready” datasets. (See Figure 6.) Confidentiality review, metadata production, and integration into a domain-specific digital repository are fully under way. Of course, not all research projects will contribute their digital output beyond publications to the social science commons, and not all datasets are destined to be archived in perpetuity. It is clear, however, that in order to increase the participation rates by researchers and to fulfill the mandates of some funding agencies, new partnerships, support, and channels of communication can dramatically influence the processing of research data. Since these efforts can be labor-intensive and expensive, we propose that support, tools, expertise, and the use of standards be employed during the entire research life cycle so that the process of archiving data for reuse and long-term accessibility will be greatly enhanced.

Making archiving choices responsive to particular research projects is key to data sharing and preservation success. For example, ICPSR is developing flexible archiving options for active data collection projects in which ongoing data management and dissemination are important components of the research activity that make the transfer of data to ICPSR not immediately appropriate. They provide producers with five primary alternatives to the traditional archive model:

- Data preservation only
- Data preservation with delayed dissemination
- Restricted-use data
- Enclave release of data
- “Virtual” data archiving

This “virtual” data archiving option “offers [data] producers a selection of virtual options currently in use at ICPSR that can improve the visibility and usability of producer-disseminated data to users. The package of services includes union catalog listing, full-text linked bibliography, user registration and monitoring, and linked Web access. Thus, a data producer [or institution-based digital repository] can retain control of data dissemination but DSDR will provide access through the ICPSR search interface. Transitional Web pages are being developed that will standardize the linkage between ICPSR and the producers’ data dissemination Web pages.” (LeClere, 2006)

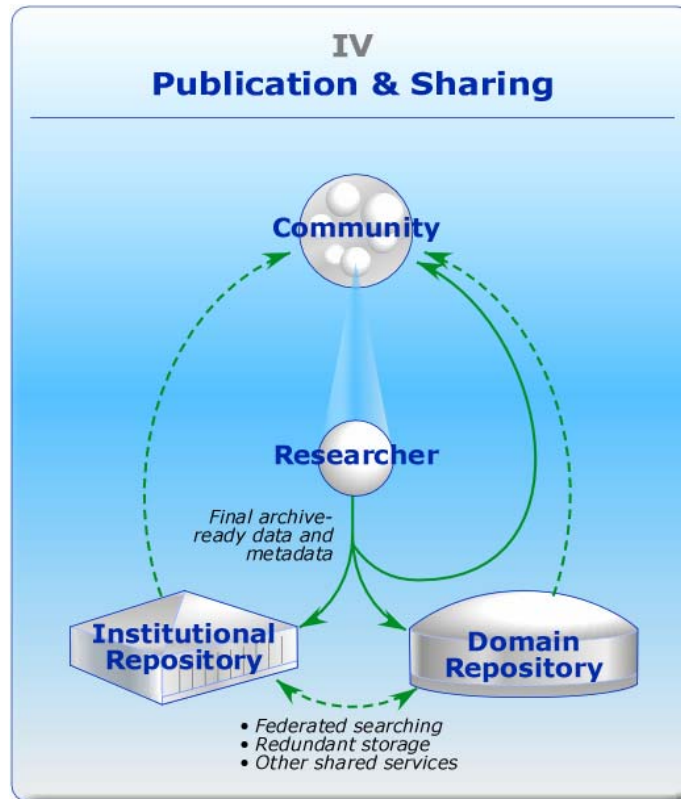


Figure 6. Final archive-ready data and metadata are moved to the Institutional Repository and/or the Domain Repository. Shared services are developed and supported among repositories.

Long-Term Management

Because domain-specific repositories have as core missions the long-term preservation of social science data, partnerships between those repositories and single institution-based repositories have begun to emerge. For example, in the United Kingdom, SHERPA-DP^[21] is developing a model whereby a network of institution-based repositories will outsource preservation to the Arts and Humanities Data Service^[22].

Domain-specific repositories also have a mission to monitor and perform ongoing analysis of domain-focused curatorial and access needs of institutional repositories and social science researchers. In order to take on these responsibilities, these institutions seek support and resources for making improvements and developing tools and standards for the curation and preservation of digital resources. This includes research and development of proprietary file format migration issues, managing access changes over time, creating and maintaining bibliographic linkages from publications to datasets, and succession planning at institution levels.

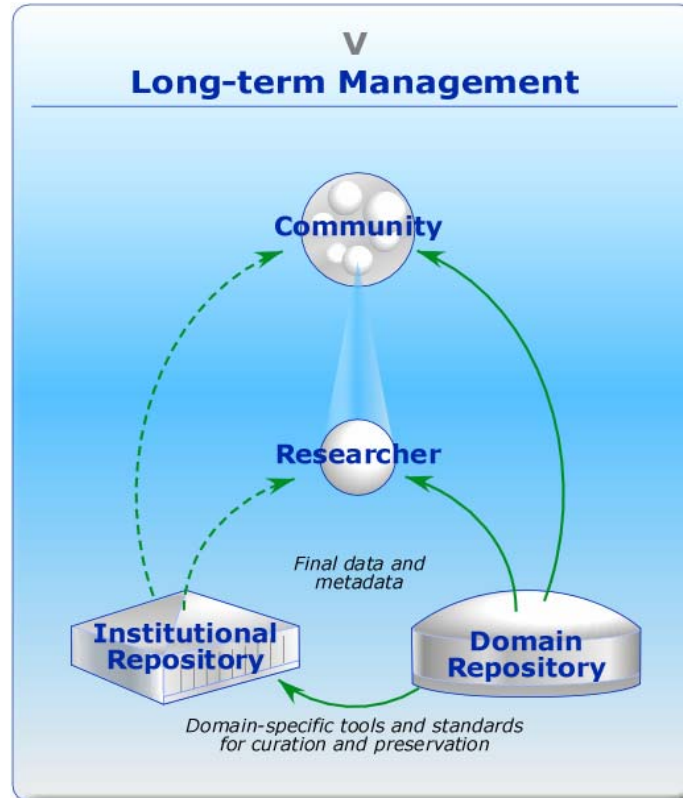


Figure 7. Access to and stewardship of data and metadata over the long-term are commitments made by the Domain Repository. Ongoing contributions of research and development of domain-specific tools and standards for creation and preservation are made available by the Domain Repository to Institutional Repositories and the Research Community.

Conclusions and Recommendations

We have asserted here that the next step in the evolution of digital repository strategies should be an explicit development of partnerships between researchers, institutional repositories, and domain-specific repositories. In order to articulate these ideas as clearly as possible, we have emphasized the scientific domain we know best, the quantitative social sciences. Many of its key features are transferable to other domains. By assigning these roles to these two types of repositories, and by understanding how they work within the research community, we believe that it will be much easier to succeed at the sharing and preserving of all forms of digital research materials.

Our key message is that by visualizing the role of repositories explicitly in the life cycle of the social science research enterprise, the pathways to collaboration will be clear. These workings can be seen as a sequence of reciprocal information flows between parties to the process, triggers that signal that one party or another has a task to perform, and hand-offs of information from one party to another that take place at crucial moments. Providing an illustration of one such partnership, we show that when data

collection is completed and a preliminary version of data is ready for transfer to a repository, high-level metadata are prepared and move from the researcher to the institutional repository (with the data), and then on to the domain repository (on their own). This transfer of metadata to the community becomes a signal that an important research project is underway, and it triggers a discussion between the researcher and the domain repository about the format in which the eventual, final metadata and data deposit should be made.

This approach envisions both cooperation and specialization. The researcher produces the scientific product, both data and publications; the institutional repository has specialized knowledge of campus conditions and the opportunity to interact frequently with the researcher; and the domain-specific repository has specialized knowledge of data management approaches to data in a specific scientific field, for example, domain-specific metadata standards (the DDI in the case of the social sciences), as well as the ability to expose the research products to the field in a way that will have the greatest impact. Put another way, the researcher is the essential element that sets the whole process in action, with the domain repository facilitating all the elements of data-oriented scientific collaboration, and the institution-based repository facilitating transactions between the researcher and the domain-specific repository while gathering digital research outputs deemed of local value.

References

- Burnhill, P., Geraci, D. and Rice, R. (2005), The Social Science of Data Sharing: Distilling Past Efforts. Available <http://www.ukoln.ac.uk/events/pv-2005/posters/burnhill-geraci-rice.pdf>
- Cervone, H.F. (2004), "The Repository Adventure", Library Journal, Vol 129 No10. Available <http://www.libraryjournal.com/article/CA421033.html>
- Crow, R. (2002), The Case for Institutional Repositories: A SPARC Position Paper. The Scholarly Publishing & Academic Resources Coalition. Available <http://www.arl.org/sparc/IR/ir.html>.
- DDI Alliance. (2004), DDI Version 3 Conceptual Model. Available <http://www.icpsr.umich.edu/DDI/committee-info/Concept-Model-WD.pdf>.
- Economic and Social Data Services (2005), Guide to Good Practice: Data Management. Available <http://www.esds.ac.uk/support/guides/A4.pdf>
- Economic and Social Data Services (2005), Guide to Good Practice: Micro Data Handling and Security. Available <http://www.esds.ac.uk/news/microDataHandlingandSecurity.pdf>
- Green, A. (2005), Review of Digital Repositories. Report to the Integrated Access Council, Yale University Library. Available http://www.library.yale.edu/iac/documents/DR_Review_final_27Sept05.pdf
- Green, A., Dionne, J. and Dennis, M. (1999), Preserving the Whole: A Two-Track Approach to Rescuing Social Science Data and Metadata. Washington: Council on Library and Information Resources. (CLIR Publications 83). Available <http://www.clir.org/pubs/reports/pub83/contents.html>
- Gutmann, M., Schurer, K., Donakowski, D. and Beedham, H. (2004), "The Selection, Appraisal and Retention of Social Science Data." CODATA Data Science Journal 3, pp. 209-221. Available http://www.jstage.jst.go.jp/article/dsj/3/0/209/_pdf
- Harnad, S. (2006), Generic Rationale for University Open Access Self-Archiving Policy. Available <http://eprints.ecs.soton.ac.uk/12078/01/genericSApolicy-linked.html>
- Heery, R. and Powell, A (2006), Digital Repositories Roadmap: Looking Forward. Joint Information Systems Committee (JISC), Digital Repositories Programme. Available www.jisc.ac.uk/uploaded_documents/rep-roadmap-v15.doc
- Humphrey, C. (2006), e-Science and the Life Cycle of Research. Available <http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>

Humphrey, C. (2005), "The Preservation of Research Data in a Postmodern Culture." IASSIST Quarterly, Vol 29 No 1, pp. 24-25. Available <http://iassistdata.org/publications/iq/iq29/iqvol291humphrey.pdf>

Humphrey, C.K., Estabrooks, C.A., Norris, J.R., Smith, J.E. & Hesketh, K.L. (2000), "Archivist on board: Contributions to the research team." Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol 1 No 3. Available <http://www.qualitative-research.net/fqs-texte/3-00/3-00humphreyetal-e.htm>

Inter-university Consortium for Political and Social Research (2005), Guide to Social Science Data Preparation and Archiving. Third Edition. Ann Arbor: Inter-university Consortium for Political and Social Research. Available <http://www.icpsr.umich.edu/access/dataprep.pdf>

Jacobs, J., and Humphrey, C. (2004), "Preserving Research Data." Communications of the ACM, Vol 47 No 9, pp. 27-29.

Johnson, R.K. (2002), "Institutional Repositories: Partnering with Faculty," DLib Magazine, Vol 8 No 11. Available <http://www.dlib.org/dlib/november02/johnson/11johnson.html>

Joint Information Systems Committee (JISC), Digital Repositories Programme. (2005), Digital Repositories Review. Available http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf
Digital Repositories Review: Annex 1: Focus Group Report. Available http://www.jisc.ac.uk/uploaded_documents/rep-review-Annex1-fg.pdf.
Digital Repositories Review: Annex 2: JISC Digital Repositories Review Software Developers Survey. Available http://www.jisc.ac.uk/uploaded_documents/rep-review-Annex2-software.pdf.
Digital Repositories Review: Annex 3: Repository Issues...from a Teaching and Learning Perspective. Available http://www.jisc.ac.uk/uploaded_documents/repos_issues_cetis_feb05.pdf

Lagoze, C., Krafft, D.B., Payette, S. and Jesuroga, S. (2005), "What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL." DLib Magazine, Vol 11 No 11. Available <http://www.dlib.org/dlib/november05/lagoze/11lagoze.html>

LeClere, F. B. (2006), "Data Sharing for Demographic Research at ICPSR." ICPSR Bulletin, Vol 26 No 2, pp. 3-8. Available <http://www.icpsr.umich.edu/org/publications/bulletin/2006-Q1.pdf>

Lynch, C. (2003), "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." ARL Bimonthly Report 226, pp. 1-7. Available <http://www.arl.org/newsltr/226/ir.html>.

Lyon, L. (2003), "eBank UK: Building the links between research data, scholarly communication and learning", Ariadne Issue 36. Available <http://www.ariadne.ac.uk/issue36/lyon/intro.html>

Online Computer Library Center (OCLC) (2004), Research and Learning Landscape: Institutional Repositories, Scholarly Communication and Open Access, The 2003 OCLC Environmental Scan: Pattern Recognition: A Report to the OCLC Membership. Dublin, OH. Available <http://www.oclc.org/reports/escan/research/repositories.htm>

Open Society Institute (2004), A Guide to Institutional Repository Software v 3.0. New York: Open Society Institute. Available <http://www.soros.org/openaccess/software/>.

Pritchard, S. M., Carver, L. and Anand S. (2004), Collaboration for Knowledge Management and Campus Informatics. Available http://www.library.ucsb.edu/informatics/informatics/documents/UCSB_Campus_Informatics_Project_Report.pdf

Royal Statistical Society and the UK Data Archive. (2002), Preserving and Sharing Statistical Material. Colchester: UK Data Archive. Available <http://www.data-archive.ac.uk/news/publications/PreservingSharing.pdf>

Rusbridge, C. (2005), Information Life Cycle and Curation. Available www.dcc.ac.uk/docs/dcc-life-cycle.ppt

Swan, A. (2005), Open access self-archiving: An Introduction. JISC/ Key Perspectives Technical Report. Available <http://eprints.ecs.soton.ac.uk/11006>

United States. National Science Board. (2005), Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. Pre-publication draft. Arlington, VA: National Science Foundation. Available http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf.

United States. National Science Foundation (2006), NSF's Cyberinfrastructure Vision for 21st Century Discovery. Version 5.0. Available http://www.nsf.gov/od/oci/ci_v5.pdf

¹ Ann Green is a consultant with Digital Life Cycle Research & Consulting, New Haven, CT (www.dlifecycle.net). Myron Gutmann is the Director of the Inter-university Consortium for Political and Social Research (ICPSR) and Professor of History at the University of Michigan.

² <http://www.icpsr.umich.edu>

³ Other digital life cycle diagrams of interest can be found in ICPSR (2005), DDI Alliance (2004), UKDA/ESDS (2005, p. 8), Humphrey, C. (2006), Rusbridge, C. (2005), and Lyon, L. (2003).

⁴ <http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf>, article 36.

⁵ http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

⁶ <http://www.ropercenter.uconn.edu/>, <http://www.hmdc.harvard.edu/jsp/supportedgroups.jsp?id=6>, <http://www.cpanda.org/>

⁷ <http://library.physics.harvard.edu/dspace/index.jsp>

⁸ <http://www.icpsr.umich.edu/DataPASS>, <http://www.irss.unc.edu/odum/>, <http://www.hmdc.harvard.edu/jsp/index.jsp>, <http://www.archives.gov/>

⁹ <http://odwin.ucsd.edu/idata/>

¹⁰ <http://www.cessda.org>

¹¹ <http://www.ddi.org>

¹² <http://sda.berkeley.edu>

¹³ <http://www.thedata.org>

¹⁴ <http://www.nesstar.com>

¹⁵ Two projects presented at the 2005 FEDORA User's Conference are investigating the use of FEDORA for preservation management. See: "Researching FEDORA's ability to Serve as a Preservation System for Electronic University Records – Dockins, R., Glick, K., Wilczek, E. and "Digital Preservation Using the FEDORA Framework" – Jantz, R. http://www.scc.rutgers.edu/fedora_conf_2005/program.html

¹⁶ <http://www.lib.duke.edu/its/diglib/digarchive/custodianship.html>

¹⁷ <http://libraries.mit.edu/dspace-mit/build/policies/format.html>

¹⁸ See Preserving and Sharing Statistical Material (Royal Statistical Society and the UK Data Archive) as an example of a document produced by a data archive to promote the need for preserving and disseminating statistical research output.

¹⁹ <http://www.openarchives.org/>

²⁰ ICPSR's *Guide to Social Science Data Preparation and Archiving* and the ESDS publication *Guide to Good Practice: Data Management* provide guidance to depositors, and potential depositors, of data to ICPSR and to ESDS.

²¹ <http://ahds.ac.uk/about/projects/sherpa-dp/>

²² <http://ahds.ac.uk/>