Center for Research Libraries
Auditing and Certification of Digital Archives Project

ICPSR Audit Report
For the period ending 24 October 2006

Report prepared by Robin Dale, with contributions from G. Sayeed Choudhury, Tim DiLauro and Marie Waltz.

Note: This report was produced as part of a test of the RLG/NARA Draft Audit Checklist for the Certification of Trustworthy Digital Repositories and other metrics developed by the Center for Research Libraries under a grant from the Andrew W. Mellon Foundation. Because the metrics and methodologies applied were still in development at the time of the audit, this report should not be considered a definitive assessment of the repository described.

TABLE OF CONTENTS

1. Summary Statement

The audit of the Inter-university Consortium for Political and Social Research (ICPSR) took place between August 2005 and March 2006. The goal of the test audit was to evaluate the ICPSR to form an overall risk analysis as it relates to long-term scholarly access to data acquired and managed by ICPSR, and to test a methodology for conducting such assessments developed by the Center for Research Libraries. The methodology was based largely on the August 2005 draft of the *RLG-NARA Checklist for the Certification of Trusted Digital Repositories*, which provided a set of metrics and controls, and on other instruments prepared specifically for this project. ICPSR was evaluated on three aspects of its archiving operations:

a. characteristics of the ICPSR organization that might affect performance, accountability, and business continuity;
b. technologies and technical infrastructure employed in its archiving activities; and
c. preservation processes and procedures adopted by the archive.

Auditors found an organization with a mature, fully operational archive. ICPSR as an organization has a 44-year history of growing and managing a large (at the time of the audit 2.3 Terabytes) data archive of valuable content with a virtually unblemished record in data management and access. The future prospects of both the organization and the data appear favorable, due to a combination of sound financial management and planning, multiple sources of revenue, robust reporting and accountability mechanisms, and sound technical decisions related to processes, procedures, and formats.

Concerns raised include the absence of a succession plan for the organization, and the potential for future increases in the membership cost of ICPSR that may be needed to support management of the growing amount of content being managed and the increasing complexity of that content. Such increases might be avoided if the ICPSR subscriber base continues to grow. Moreover, the organization's activities rely heavily (50%) on grant funding, which can be volatile.

In addition, the acquisition of preservation rights from depositors of data, and the consistent maintenance of documentation of changes made to ICPSR systems and digital objects should be made priorities as well.

While some issues requiring resolution were identified in the audit, when taken as a whole ICPSR appears to provide good stewardship of the valuable research resources in its custody. Since the time of the audit ICPSR has made significant progress in addressing several of the concerns raised in this report. Contributors of data to the ICPSR archives and users of those data should feel confident about the state of the organization, as well as the processes, procedures, technologies, and technical infrastructure it has in place.

2 Executive Summary

In March 2005, the Andrew W. Mellon Foundation funded the Center for Research Libraries (CRL) *Auditing & Certification of Digital Archives* project, an endeavor to develop an audit and certification process for digital repositories and archives. Rigorous auditing and certification are necessary to determine the level of assurance that particular archiving arrangements provide to publishers/depositors and users, and to ensure that valuable digital resources will continue to be available and functional over time. As part of the CRL project, a team of four auditors performed an audit and assessment of the digital archives (the Data Archive) at the Inter-university Consortium for Political and Social Research (ICPSR). ICPSR agreed to participate in the project as a "subject archive," one of three such archives to undergo a test audit as a part of the process development.

The audit of the Inter-university Consortium for Political and Social Research (ICPSR) took place between August 2005 and March 2006, with an on-site visit by auditors on 30 January – 1 February, 2006. Among the goals of the test audit was to evaluate the ICPSR to form an overall risk analysis as it relates to long-term scholarly access to data acquired and managed by ICPSR. The methodology and metrics for the audit were based largely on the August 2005 draft of the *RLG-NARA Checklist for the Certification of Trusted Digital Repositories,* which provided a set of controls and criteria for assessing the preservation capabilities of systems for digital archiving, as well as other instruments developed specifically for this project. ICPSR was evaluated on three aspects of its archive operation:

a. characteristics of the archiving organization that might affect performance, accountability, and business continuity;
b. technologies and technical infrastructure employed by the archive; and
c. processes and procedures adopted by the archive.

On the basis of the 2.5-day site visit and subsequent analysis, auditors found ICPSR to be an organization with a mature, fully operational archive. In fact, ICPSR has a 44-year history of growing and managing a large (2.3 Tb) data archive of valuable content with a virtually unblemished record in data management. The future prospects of both the organization and the data archived appear to be good, owing to a combination of sound financial management and planning, diverse sources of revenue, robust reporting and accountability practices, and solid technical decisions related to processes, procedures, and formats.

As expected, some vulnerabilities were identified in the course of the audit. Most of these related to several high-level classes of technical issues, including documentation, automation of key critical processes, and security. Many of these vulnerabilities had already been identified by ICPSR through a self evaluation using the *Audit Checklist*. At the time of the audit the lack of process and procedural documentation (including metadata) was scheduled to be addressed by a new Digital Preservation Officer (start date - September 2006), while other issues will take a concerted effort by the organization to introduce new processes and some new system infrastructure.

Of greater concern was the absence of a succession plan for ICPSR. This calls into question the long-term viability of the data in the event that the Institute for Social Research and the University of Michigan, ICPSR's parent organizations, no longer choose to support its work. While ICPSR management has had numerous conversations with organizations that might serve as trusted inheritors of the data archives in the event of some sort of economic or technical failure, at the time of the audit they were as yet unsuccessful in identifying such a successor. At the time of the audit no other institution capable of managing the range of ICPSR data types, content, and access requirements was interested in taking on the ICPSR data archives. Though ICPSR is unlikely to fail, it should continue to seek to identify appropriate inheritors of individual *collections* if not a sole, trusted inheritor of the entire archives.

Other concerns are the heavy reliance of ICPSR on grant funds (50% of ICPSR FY 2006 revenues) and the potential that significant subscription price increases may be needed in the future to accommodate the growing amount of content being managed and the increasing complexity of that content. The repository has little control over the adoption of new types of data sets and software by the social science community, developments that are sure to create new costs for ICPSR. The burden of these costs on ICPSR members, however, might be reduced if the ICPSR subscriber base continues to grow.

In addition, acquisition of preservation rights from depositors of data is not part of the ICPSR routine, potentially limiting the repository's ability to make the changes or modifications to data sets and digital objects that might be necessary to migrate them to new platforms and adapt them to new user needs. Auditors also noted inconsistencies in the maintenance of documentation of changes made to ICPSR policies, systems and digital objects over time. Remediation of these items should be made a priority as well.

Outweighing the problems however, were the series of good decisions underpinning the entire process cycle of the archive, from acquisition of material to ingest to archival storage and finally to provisions for access to the material. ICPSR decisions reflect the management's experience as well as its commitment to integrity of data and long-term user access. ICPSR procedures reflect the needs of stakeholders (both data depositors and users) and generally adhere to standards and community best practices. The effectiveness of ICPSR in meeting the immediate needs of its designated community is evinced by its history of satisfying the needs of a large population of researchers for archived social science data sets.

Finally, it is important to stress the quality of ICPSR management. Under Myron Gutmann, the current Director of the ICPSR, the organization has strengthened itself fiscally and has undergone an organizational talent evaluation and reclassification to make sure key staff members have appropriate skills, roles, and responsibilities to fulfill ICPSR's mission critical operations. Moreover, organizational planning & budgeting over the past few years have allowed ICPSR to develop a fiscal contingency fund for use in the event of a funding gap (although at the time of the audit, this fund would merely support salaries and other expenditures for several months rather than being applied to data transfer or other archive succession plans). In fiscal 2004-05, the ICPSR fund balance reached its healthiest level to date with an overall reserve of almost $2.2 million. This reserve has continued to grow since the time of the audit.

Taken as a whole, ICPSR appears to provide responsible stewardship of the valuable research resources in its custody. Depositors of data to the ICPSR data archives and users of those archives can be confident about the state of its operations, and the processes, procedures, technologies, and technical infrastructure employed by the organization.

3 Full Report

3.1 Introduction

As a part of the Center for Research Libraries Auditing & Certification of Digital Archives project, a team of four auditors (Robin L. Dale, Sayeed Choudhury, Tim DiLauro, and Marie Waltz) performed an on-site assessment of the Data Archive at the Inter-university Consortium for Political and Social Research. At the time of the audit the Andrew W. Mellon Foundation-funded project was engaged in developing a complete audit and certification process for digital repositories and archives. Rigorous auditing and certification are necessary to determine the level of assurance that particular archiving arrangements provide to publishers, depositors and users of digital resources, and to ensure that the valuable digital resources archived will continue to be available and functional over time. ICPSR agreed to participate as a "subject archive," one of three such archives to undergo a test audit as a part of the project. The results of that test audit are the subject of this report.

3.1.1 Repository type & background

Established in 1962, the Inter-university Consortium for Political and Social Research (ICPSR) maintains and provides access to a vast archive of social science data sets for research and instruction. It is the world's largest digital social science data archive, comprising over 6,200 collections and more than 1.2 million discrete files. It acquires and ingests approximately 300-400 collections per year and manages some of the largest and most widely-used social science data sets available.

To ensure that data resources within the Data Archive are available to future generations of scholars, ICPSR migrates the collections to new storage media as changes in technology warrant. This has occurred several times over the course of the ICPSR's existence. In addition, ICPSR provides user support to assist researchers in identifying relevant data for analysis and in conducting their research projects.

ICPSR manages two separate copies of the Data Archive: one for service use (distribution) and one that solely supports the preservation mission. Statistics for late 2005 provide a brief view of size and complexity of data managed by the ICPSR:

> Distribution archive:
> Size in Tb: 1.82 Tb (compressed)
> Number of files: 490,538
> Number of studies available: 5,791
>
> Preservation Archive:
> Size in Tb: 2.3 Tb
> Number of files: 1,189,000 (±0.01 %)

ICPSR prides itself on its solid reputation for preserving data – at the time of the audit it had a virtually unblemished 44-year history of preserving electronic data on behalf of the research community.

3.1.2 ICPSR Philosophy

ICPSR works with and encourages social scientists in all fields to preserve their research data. Their mission statement clearly specifies their philosophy and goals:

> The Inter-university Consortium for Political and Social Research is an organization of member institutions working together to:
> - Acquire and preserve social science data
> - Provide open and equitable access to these data
> - Promote effective data use
>
> ICPSR encourages and facilitates research and instruction in the social sciences and related areas by acquiring, developing, archiving, and disseminating data and documentation relevant to a wide spectrum of disciplines, and by conducting related instructional programs.
>
> A strategic undertaking of ICPSR is the acquisition and long-term preservation of social science data, recognizing and taking into consideration increases in the volume of data and changes in technology for archiving, processing, documenting, and distributing data.[1]

By depositing data with ICPSR, researchers are able to use ICPSR not only for redistribution and reuse of their data, but also for "long-term safekeeping . . . , protecting it from obsolescence, loss, deterioration, or irreversible damage."[2]

ICPSR accomplishes its twin missions of access and preservation by operating separate service and preservation archives. See Section 3.3.2, *Technical Analysis*, for more information about the differences in archives, data ingested, and philosophy for preserving the data.

3.1.3 Organizational Structure (Brief Overview)

ICPSR is an organizational unit within the Institute for Social Research at the University of Michigan. Its operations are distributed among eight functional divisions, in addition to the Director's Office. The divisions each play a role in the functions of the Data Archive and include: Data Security and Preservation, Computing and Network Services, Central Administration, Collection Development, Research Staff, Faculty Associates, Educational Activities, and Collection Delivery. A detailed view of the internal organizational structure can be seen in Appendix 4.3, *Organizational Chart* and in section 3.3.1.2 *Staff*, below.

ICPSR is also governed by the ICPSR Council, a group of leading scholars and data professionals who guide and oversee its activities.

[1] http://www.icpsr.org/org/mission.html
[2] *Archiving Data at ICPSR*, http://www.icpsr.umich.edu/org/publications/archivingdata.pdf

3.1.4 Technical Architecture (Brief Overview)

The technical architecture of the ICPSR system has evolved over time and supports not only the archival mission, but also the Web services and business continuity of the enterprise. It is managed by the ICPSR Computer and Network Services Group in concert with the Preservation Group.

ICPSR maintains two separate collections of data, one for servicing clients and an archival collection for ensuring the archival integrity of the collection. Each collection is managed by an independent database, and the archival collection and its associated database are accessible only to the Data Security & Preservation staff. The archival collection is routinely migrated from medium to medium but is used only when a dataset in the servicing collection fails. Two copies of each file from this collection are stored off-site on DLTs.

The ICPSR system is comprised of multiple Sun Enterprise servers (model E3500). Each server has four processors, four internal hard drives, and one or more external disk arrays. The redundancy here allows for continual functioning of both archive and distribution systems should any one of the components fail. Incremental backups are performed every evening, and a full backup once a week.

Detailed information about the technical architecture can be found in Section 3.3.2, *Technical Analysis*.

3.2 Audit Objectives, Scope, & Methodology

3.2.1 Scope

This audit evaluated and provides information on the following topics:

- Organizational Infrastructure
- Technical Analysis (Digital Object Management, Technologies and Technical Infrastructure)
- Content
- Vulnerabilities
- Observations & Recommendations

In all areas, the focus was on identifying and describing issues that could affect the viability and stability of the repository and the digital objects stored within it.

3.2.2 Method of Work

The work performed in this audit consisted of a review of documentation (both publicly available information and answers to sets of questions posed in advance of the onsite visit); interviews conducted with key staff during the onsite visit; completion of the *RLG-NARA Checklist for the Certification of Digital Repositories*; and observations. (See Appendix 4.3 for a list of ICPSR staff interviewed.)

The onsite visit included inspection of server rooms, network utilities, HVAC provisions, and servers associated with providing information to users. Security arrangements for the server room were also noted and examined.

No detailed technical testing was conducted. Rather the technical auditors created several scenarios related to ingest, data security, archival storage, and access provisions to which technical staff had to verbally address and respond. These scenarios were designed to detect potential vulnerabilities in policies, functionality (ingest, processing, archival package creation, data loss detection & resolution, access, etc), staff knowledge, and facilities. Such scenario testing does not test the validity of the digital records within the digital archive, but can provide insight into threat detection and risk management capabilities of the archive. Finally, an analysis of content was made to investigate discovery and delivery options as well as test them against ICPSR access policies.

3.2.3 Standards against Which the Audit Was Completed

The *RLG-NARA Checklist for the Certification of Digital Repositories* (August 2005) provided the metrics for this audit. The checklist was developed by an international task force of experts on digital preservation, digital repositories, and data archives. While the Checklist is not an international standard, its draws upon and incorporates a number of international standards and

best practices such as the *Reference Model for an Open Archival Information Syste*m (ISO 14721:2004), *Control Objectives for Information and Related Technologies* (COBIT) 4.0, *Information Technology—Security techniques—Code Of Practice for Information Security Management* (BS ISO/IEC 17799:2005), *PREMIS* Preservation Metadata (2005), and *Trusted Digital Repositories: Attributes and Responsibilities* (2002).

3.3 Findings

3.3.1 Organizational Analysis

3.3.1.1 Governance

A unit within the Institute for Social Research (http://www.isr.umich.edu/) at the University of Michigan, ICPSR is a membership-based organization, with over 500 member colleges and universities around the world. It is not an independent legal entity or corporation. A council of leading scholars and data professionals guides and oversees the activities of ICPSR. (See Appendix 4.2 for a listing of Council members.) ICPSR's day-to-day activities are managed by the Director of the ICPSR, Myron Gutmann. This position operates on a five-year rotating term (renewable one time), and Gutmann is in the fifth year of his first term. A complete organizational chart can be found in Appendix 4.1.

3.3.1.2 Staffing

The ICPSR maintains a staff of approximately 90 people with an additional five faculty associates advising and contributing as required. Of the 90, approximately 55 are associated with the Director's Office and the following divisions: Data Security and Preservation, Computing and Network Services, Central Administration, Collection Development, Research, Educational Activities, and Collection Delivery (including metadata, as well as user and web support). The remaining staff comprise units dedicated to processing and loading of specific archives such as the Child Care Archive, the Health and Medical Care Archive, and the National Archive of Criminal Justice Data. In September, 2006, the ICPSR began to employ a full-time staff member dedicated to digital preservation.

Current staffing levels were deemed to be adequate for the size and scale of the collections under the care of ICPSR. ICPSR has experienced significant staff growth in the last 3-5 years although staff acknowledge a level of backlogged work within Collection Development and Collections Delivery (including processing) departments. This is a level of "normal" backlog with which the ICPSR is comfortable.

Staff skills and expertise appear to be exemplary. Due to the fluid nature of grant funded activities in some departments, existing staff are often transitioned to new projects, minimizing the number of new hires who would require a period of training and time to acquire the appropriate skills. Job descriptions clearly articulate required job skills and all job descriptions were recently updated to reflect current organizational needs. Individual job performance is measured with reference to staff workplans. Performance reviews take place biannually and new skill needs are identified and addressed through professional development opportunities. A number of ICPSR staff members are SANS-certified in systems security.

3.3.1.3 Policies and Procedures

Policy and procedural documentation is treated at different levels within the ICPSR. High level policies related to digital collection acquisition for the repository, the organization, and other general enterprise-wide matters are formalized, reviewed and approved by the ICPSR Council. The review and renewal period for these kinds of polices is generally five years. The review and

renewal procedures are transparent and results are conveyed in public documentation. A long history of repository policy development is available through the official ICPSR organizational archive, which is maintained by the University of Michigan's Bentley Library.

Of more concern are the policies and procedures related to changes made to digital objects, the repository technical architecture, and ICPSR systems. In response to questions regarding a documented history of the changes to its operations, procedures, software, and hardware, traceable to its preservation strategies ICPSR asserted, "Looking back, we can point to Council meeting minutes and briefing books, which contain a history of such changes and decisions. In addition, since early 2005 we will have this in the form of internal and external Web announcements of such changes." This reply indicates that the kind of detailed, technical information on such changes is not regularly maintained in a way that would allow for easy auditing and analysis of the ICPSR archiving system over time. That is not to say that the information is not gathered, but that the current, diffuse manner in which the information is gathered and logged makes it almost impossible to get a longitudinal view of all changes made that could potentially affect the digital objects managed by ICPSR. ICPSR management acknowledged that more standardized, comprehensive documentation facilitates this perspective and that improvement in this area is needed.

ICPSR performed well in the areas related to transparency and accountability. In general, information that affects users and depositors is kept up-to-date via the ICPSR website. This is the chosen, "one stop" vehicle for communicating changes to the designated community. For policies and procedures under development, ICPSR uses an intranet to manage the draft and discussion process so that affected staff and Council members have access to and may comment on proposed documentation changes. ICPSR has committed to formal, periodic review (including certification when offered) and the organization's voluntary participation in this project is evidence of such a commitment.

3.3.1.4 Financial Analysis

As one of the few long-term data archives, ICPSR has had the advantage of time to establish a fairly firm financial base and a well-formed business planning process.

Revenue & Expenses

Financial information on ICPSR was provided in the form of annual budget statements for the organization. Audited financial statements were not available for ICPSR itself, owing to the program's status as a part of the University of Michigan. Funding for operations is generated by a variety of activities and services including grants and gifts to process data collections, tuition from the ICPSR Summer Program, membership dues, and contracts with the US government. For 2006, the projected budget was $14.36 million. ICPSR estimated that 2006 revenues (sources noted below) would exceed expenses by approximately $181,000. Added to the previous year fund balance, the positive ICPSR fund balance for fiscal 2006 was expected to be over $2 million.

<u>2006 Revenue Sources:</u>

Gifts & Grants: 50.3% (direct & indirect)

<u>2006 Membership-funded Expenses by Functional Area</u>

General Income: 23.6%
Indirect Costs: 15.4%
Recovery from Rebill: 6.5% (computer recharge)
General Fund Transfers: 4.0%
Investment transactions: 0.1%

Collection Delivery: 21.1%
Collection Development: 23.7%
Other Central Expenses: 24.5%
Data Security and Preservation: 13.8%
Administration: 11.7%
Educational Activities: 5.2%

In addition to a positive fund balance, the ICPSR maintains a Director's Contingency Fund to use in times when expenses exceed revenue. The ICPSR also maintains mandated reserve funds to cover termination expenses should the organization ever fail. This fund however, is intended to be applied to staff-associated costs and would not necessarily be adequate to subsidize transfer of the data to another repository or organization for the long-term.

One significant vulnerability or risk incurred by ICPSR is its dependence upon the University of Michigan for some of its critical infrastructure (campus network, network security, electrical power, Oracle license) and facilities. The relationship is governed by a Memorandum of Agreement between ICPSR and the Regents of the University, the text of which is published on ICPSR's Web site. While the University is obviously a stable, longstanding organization, its contribution to the ICPSR is largely pro bono, and may thus be regarded as discretionary, since most of ICPSR's user population resides outside the University community. Similarly, ICPSR's reliance upon gifts and grants for a significant percentage (50.3%) of its annual income presents something of a risk as well.

Funding Dedicated to Data Preservation

Balance sheets divide ICSPR budgets into two areas: Membership activities, and Indirect Cost Recovery activities. Membership-funded expenses – those funded by the anticipated $2,769,141 in 2006 membership dues – include the Data Security and Preservation unit of ICPSR. Normally, approximately 14% of the membership revenues are used to support Data Security and Preservation, yielding a budget for that unit of approximately $341,000 for fiscal 2006.

Funding designated specifically for data preservation is slowly beginning to grow because of two changes at ICPSR. First, ICPSR previously sought funding (grants, contracts, etc.) specifically to fund data processing and distribution. As existing contracts expire and as new grants are sought, ICPSR has been adding the cost of preservation activities to its overall requests for funding. It is anticipated that newer grants will include this maintenance funding at least to a minimal degree and that future funding requests and contracts will continue to grow those funds. Second, ICPSR created a new preservation endowment in 2005 to generate and manage funds for data preservation. At the time of the audit the endowment balance was only $5,000 but based on annual increases in other ICPSR endowments, they envision annual increases of $2,000 for the preservation endowment.

Budget Direction & Cost Controls

The ICPSR budget is created under the auspices of the Director's Office and Central Administration. The budget is subject to review and approval by the ICPSR Council. The

Budget Committee within the Council identifies and addresses issues of consequence and has the power to direct changes. The full Council also addresses budgetary matters during the mandated annual review. (ICPSR staff noted that the meeting for the annual budget review is one of the more lively ones.) Council members have the ability to question plans – in fact, the Council is required to be "hands-on," and is so, according to ICPSR administrative staff.

3.3.1.5 Contracts (Submission Agreements) & Licenses

Content ingested into the ICPSR Data Archive falls into two general categories: content provided by producers, generally researchers, survey organizations, and others; and public information data sets created by the U.S. government. Content related to the latter is covered under the public information protections statutes and thus no contracts are needed between ICPSR and the publisher to govern uses of this content. In acquiring researcher-produced data, however, ICPSR executes a variety of contracts, from formal deposit agreements with depositors to agreements with funding agencies. Interestingly, contracts with the data depositors do not usually transfer copyright of the data to ICPSR or grant broad rights, but instead allow ICPSR only to "archive and distribute" the data. Some contracts require the ICPSR to protect privacy (restrict access to data) for specified time periods or impose other restrictions on access (e.g., onsite access in the Data Enclave only). Section 3.3.2.4, *Content Accessibility*, provides more information on contractual privacy obligations and ICPSR's mechanisms for honoring them. Copies of the ICPSR standard deposit forms (contracts) are available to the public on the ICPSR website.

While general access and copyright issues are clearly articulated in deposit agreements and contracts with funding agencies, at the time of the audit preservation rights – more specifically the explicit right allowing the ICPSR to manipulate or reformat the data for purposes of preservation – were not explicitly addressed. ICPSR's preservation rights then are implicit at best, which could potentially pose legal problems for ICPSR or restrict its ability to carry out necessary migrations or reformatting of data.[3]

ICPSR does not license data from other institutions and is otherwise very careful not to ingest material with unclear ownership rights. The use of data deposit forms for each new collection facilitates this process.

3.3.1.6 Succession Planning

According to the Director, Myron Gutmann, ICPSR has held several conversations with a select few data centers with the goal of establishing a clear plan for succession. In each case, the other data center declined to agree to serve as successor to the ICPSR data archive. In every instance, the negative response was attributed to the sheer size of the existing ICPSR data archive as well as the costs and responsibilities that would come with such an "inheritance."

Auditors encouraged ICPSR to consider succession plans for key discrete collections of data, promoting the sustainability of and reliability of access to its unique digital collections (not all collections in the ICPSR Data Archive are uniquely held). Since the auditors' visit, ICPSR has

[3] ICPSR management concurred with this assessment and has since added specific language to its deposit agreement and collection development policy to specifically allow ICPSR the right to appropriately modify digital data and content.

begun to explore this possibility with a variety of institutions, including the San Diego Supercomputing Center and other organizations building infrastructure as a part of the National Digital Information Infrastructure and Preservation Program. It is assumed that agreements to provide backup for at least the unique digital collections (if not for ICPSR access versions of more common data sets) will be established in the near future. Because the establishment of institutional agreements will occur over time, ICPSR should establish some sort of registry or tracking mechanism for collections which are being replicated at and preserved by other institutions. This is information that should be made publicly available.

3.3.2 Content Analysis

3.3.2.1 Logical and Physical Content

The ICPSR repository is principally a data archive – that is, the archive holds the results of social research studies, the results of which come in a variety of formats. At the time of the audit ICPSR contained approximately 7,000 collections (representing over 1 million discrete files) with approximately 5,800 studies available for use through their online system. A complete listing of holdings at the study level is available through the ICPSR search engine and additional access is provided through a variety of alternative access points such as bibliographies and special topic archives. In addition, the ICPSR publishes a list of data collections processed each year as part of its annual report.

Content characteristics of studies or within special topic archives vary greatly, depending on the research outputs of the researcher or producer, although these can be quantified by ICPSR in its data deposit policies. ICPSR data deposit forms specify the data formats acceptable for deposit and preservation. Data files submitted to the archive must be submitted in one of the preferred formats: SPSS portable files, SAS transport files, Stata data files, or ASCII data files. In addition, depositors must provide an electronic format codebook that describes the contents of each variable, and identifies the range of possible codes, and their meanings for each variable. Again, the submission of codebooks must be done using an ICPSR preferred format: MS Word, ASCII, or in the newly-developed Data Documentation Initiative (DDI) compliant XML. Finally, any supporting documentation used to generate the data (data collection instrument, bibliography of publications based on the data, etc.) must be submitted along with the data.

Supporting documentation is also processed for both access and preservation. Documentation is generally received in Microsoft Word and then converted to PDF (soon PDF-A and TIFF). ICPSR also creates DDI-compliant, variable-level XML files. (ICPSR is actively involved in the creation of this community standard for supporting documentation.)

Once documentation and data have been converted to a variety of user formats, the data are virtually packaged for storage. One copy is sent to the archival system (referred to as “the Data Library”) and another is sent to the active access and distribution archive.

3.3.2.2 Extent of Content (Titles, Date Ranges, etc)

A full listing of the current studies is available through an ICPSR website search. Hundreds of new studies are added annually. Notifications of new collections are announced through the website and are made accessible through the catalog before releasing the study.

3.3.2.3 Usability of the Information

ICPSR prides itself on the usability of its collections and studies. Users are not simply provided with the data sets, codebooks, and documentation contributed by depositors, but instead are provided the data in usable data format and statistical package options, as well as some "prepackaged, already interpreted data reports." These latter types of files are a reflection of the changing nature of the archives users – from mostly experienced social scientists with statistical data file experience to an increasing number of undergraduates requiring data, but unable (or unwilling) to use commonly available statistical packages to interpret the data. Previous practice had been to make studies available as SAS and SPSS setup files, but as of March 2005, ICPSR began to release each new study with Stata setup files and "ready to go" files (SAS transport, SPSS portable, Stata system).[4] Studies released before that date are gradually being updated to include these expanded usability options, and collections that are emerging as popular among lay users are being considered for re-processing into the "prepackaged" data reports.

Users may request and download studies from ICPSR's Web-based authentication system, MyData. Since MyData went into production in fall of 2004, ICPSR has been better able to identify and understand its users. MyData enables ICPSR to monitor user preferences and expectations, data set use, file format choice and other features which thus allows them to identify usability issues such as the need for "interpreted data reports" for inexperienced users. Use of "Footprints," a help desk software tracking tool provides ICPSR a mechanism to monitor user questions and responses given, and spurs the creation of new FAQs (Frequently Asked Questions) on the website. These feedback mechanisms both attest to and strengthen the ICPSR's responsiveness to its designated community.

Some very old studies under the ICPSR's care had not yet been converted to electronic format at the time of the audit. These studies are converted on demand or as time allotted for retro-conversion projects permits. Meanwhile, the study data and associated codebooks may be accessed on-site should a user need to, although they are stored remotely under the management of the Data Security and Preservation Division. Other studies – especially those with sensitive or private data – are processed and ingested into the general digital archives, but are not made available for download through MyData.

Depending upon the nature of the collection and the restrictions placed on the study by the depositor, there are a variety of access points, from onsite use to onsite use in the Data Enclave. The Data Enclave is a physical space in the basement of ICPSR that houses a non-networked version of ICPSR data. For access to data in the Data Enclave users must sign a registry, and that record of use is maintained for the long-term by the ICPSR. All data use is monitored and the computers in the Data Enclave have no physical provisions for saving the data on external

[4] *Building Partnerships & Leadership: ICPSR 2004-2005 Annual Report*. Ann Arbor, MI: ICPSR, p.10.

devices. Thus, all use of materials in the Data Enclave is onsite, monitored by an ICPSR staff member, and information can only be recorded via pen and paper.

In general, it appears that ICPSR is appropriately tracking, managing, and responding to user needs, especially those related to the production of data options. New data output options are constantly being created to meet the community's needs for usability of the data. Simultaneously, the Archival Information Packages are updated (new snapshots) to include all data formats or digital object options available for each study.

3.3.2.4 Content Accessibility

Open Content

Access to content is dependent upon ICPSR data access policies. In general, all users with a web browser can navigate to www.icpsr.umich.edu and navigate to the Data Access & Analysis tab. Users can then search for content in a variety of ways. Access controls are utilized to determine who can download the studies, datasets, and documentation.

As a member-based organization, ICPSR is partially supported by member institution dues. Members receive benefits in return, including ICPSR Direct, a service that enables download of most data via the MyData system (mentioned above). Potential users from nonmember institutions can contact ICPSR for access options. This generally entails a pay-per-download transaction, but makes the datasets available to all potential users. In 2005, institutions downloaded over 321,000 datasets.[5]

Controlled Content

Because of the private or confidential nature of information captured in some social science studies, ICPSR has put in place policies and procedures to address the protection of the identities of survey respondents and the confidentiality of the content of their responses. To explain how ICPSR addresses such issues the guide *ICPSR Practices for Preserving Respondent Confidentiality* was created. The following information is excerpted from that document:

> To deal with issues of respondent confidentiality, ICPSR has established three levels of access to data files. Our control of the use made by researchers of datasets provided at each of the three levels (in respect to monitoring adherence to our norms for protecting respondent confidentiality) varies from very little to "a lot." The vast majority of ICPSR's datasets are made available as *"public-use files."* The staff has performed a confidentiality review on all these files, has redacted variables as necessary, and believes that the resulting files pose minimal disclosure risk. They are posted on our Web site and are available for members to download. The organization has virtually no control over how public-use datasets are used once the online Responsible Use Statement (containing the invocation to use the data in an ethical manner) has been presented to the researcher.
>
> A medium level of control is exercised over datasets containing original or unredacted items that the staff feels could lead to disclosure of the identity of respondents. These datasets have been designated as *"restricted-use datasets."* Most often, these datasets have a public-use dataset counterpart that has been "treated" so that it can be more freely

[5] Ibid.

used with slight risk of endangering respondent confidentiality. Restricted-use datasets are not posted on the Web site, and are not stored on ICPSR's main file server system. (At the present time, these are stored on CD-ROMs in the "secure data enclave.") Access to restricted-use datasets can be had for legitimate research purposes by individuals who apply for and complete a "Restricted-Use Dataset Agreement." These agreements act as a type of contract or license between ICPSR and the applying researcher. They spell out conditions of use pertaining to respondent confidentiality, as well as measures required for the physical safekeeping of the restricted-use dataset while in the possession of the researcher. When countersigned by both ICPSR and a responsible official at the requesting researcher's institution, a copy of the restricted-use dataset is sent on CD-ROM via registered mail directly to the researcher. The typical Agreement requires the applicant to destroy the data after a set period of time, and to provide ICPSR with proof that the data have been destroyed at the end of the period of use.

The greatest level of control over use of confidential data has been necessitated by the deposit of data that have severe confidentiality problems with them, and for which there is a heightened sensitivity to disclosure expressed by either the depositor or the ICPSR staff. Export of data such as these for research conducted elsewhere has been deemed undesirable or impossible, even with use of restricted-use agreements. The only form of access to data such as these is by on-site analysis by a researcher in ICPSR's new *secure data enclave.* A researcher applies for access to such data, and if approved, conducts the analysis in the enclave under very controlled conditions, including the constant presence of a monitor in the enclave. Analytic output and notes taken during the research process are reviewed by archive staff before being released to the researcher. The policies and procedures for the operation of ICPSR's enclave have been reviewed by Council, and are currently being examined by the University of Michigan's Office of the General Counsel. While the enclave is equipped and now functional, no researcher use had yet been made of it as of June 2005.

3.3.2.5 Link Management Solutions

As noted in Section 3.3.3.3, *Data Deposit/Ingest*, each study is assigned a unique five-digit identifier, which is treated as an integer. There are possible dependencies on this identifier that might affect the processing software and the Data Library (e.g., database fields that are numerical). At the time of the audit ICPSR was considering the Harvard-MIT VDC "fingerprints" as a next step for identifiers.

3.3.3 Technical Analysis

3.3.3.1 Architecture

At the highest level, ICPSR's business architecture is a pipeline (See Appendix 4.5, *ICPSR's Current Pipeline Process: the Insider's View, Detailed*.]. New materials come into ICPSR at one end of the pipe, and ICPSR takes a "before" snapshot of the materials submitted (an "as is" capture of the files as submitted, no processing or normalization has occurred at this point) for archival purposes. The "before" snapshot contains all submitted content. As the materials move

through the pipeline, ICPSR refines the materials to correct errors, completes gaps, and creates additional value by producing derivative products from the materials. At the end of the pipeline, ICPSR takes an "after" snapshot for archival purposes, and releases the processed materials on its website. The "after" snapshot constitutes only processed files. These snapshots, which may represent the closest match to the OAIS concept of AIPs, are stored in what the ICPSR calls the Data Library. [Note, ICPSR staff members also refer to the department with Data Library management responsibility as "the Data Library." To ease this confusion, further references to the department will be noted as "the Data Library department."]

The technical architecture of ICPSR has evolved over time and supports not only the archival mission, but also ICPSR Web services and business continuity. The Computer and Network Services Group, in concert with the Preservation Group, manage the architecture.

ICPSR has a contract with Sun Microsystems to provide hardware maintenance on the servers. Much of the software employed is commonly used open source and tends to be fairly robust. The Oracle software is covered by a general license between the University of Michigan and Oracle. These services are provided to ICPSR at no cost. The University of Michigan provides additional infrastructure support for network services and the security-controlled server room.

3.3.3.2 Data Security

The ICPSR data archives are conceived of and managed as two separate collections of data: one for servicing clients and one for ensuring the archival integrity of the collection. Each collection is managed by an independent database. The archival collection, its associated database, and backup copies are accessible only to the Data Security and Preservation staff.

A copy of the archival collection is stored in the Data Library department. A remote location holds another copy of archival backup tapes and paper logs from the Data Library. A set of tape backups of the archival collection, paper logs , and a small number of older collections (and their associated analog documentation) requiring retroconversion are stored offsite in a warehouse approximately 2 miles from the ICPSR.

Physical Security

The physical archival media stored onsite are never allowed out of the Data Library department. There is a request program available for processors to request digital objects to be downloaded to their ICPSR internal workspace (no data is downloaded to removable media). An approval memo must accompany data requests for materials outside a processor's topical archive from the Director of the requestor's topical group.

At present, there is no method to effectively protect archival analog materials from damage or authorized access without denying access to the original materials. All material checked out of the Data Library department's warehouse for on-site use is inventoried against both pre- and post checkout to ensure that no object was removed or altered, and that all identifiers are intact.

There is a remote location (warehouse as seen in Pipeline Process documentation) for backup tapes and paper logs from the Data Library. A CNS staff person decides which backup tapes are

moved during each trip to the remote site. The warehouse uses key access. There are 4 sets of keys that are held by:

- ISR Inventory Coordinator and Service Supervisor
- ICPSR facilities manager
- Data Library department, which keeps the key in a locked box in a locked cabinet

File folders from older collections awaiting processing, as well as original materials submitted with studies may be checked out of the warehouse through a request program used by processing staff, but the materials must remain in-house. Information regarding these analog archival items checked out and the processor in question are kept in a database and monitored weekly by library staff. At present, ICPSR has no specific policy that materials should not leave the building, and no mechanisms to ensure that such action has not taken place.

Authentication

As a mature archive supporting changes by a number of staff members at ICPSR, authentication exists in many different contexts at ICPSR. A typical authentication scenario involves a login/password pair. In this scenario, a human proves his/her identity to some system component at ICPSR using the password as the credential. Examples include: UNIX login and password via ssh; Windows login and password via NTLMv2; samba login and password via NTLMv2; LDAP login and password via SSL channel; MyData (home-grown authentication system used by the ICPSR web site) login and password via SSL channel; magnetic swipe card for access to certain rooms and areas; and, etc. Some MyData logins have special roles (e.g., intranet access) and all MyData credentials are transported via https. The main use of ftp is anonymous ftp. Anonymous ftp was the only troubling aspect of ICPSR authentication and at the time of the audit ICPSR was planning to eliminate this kind of access in the near future.

For better and worse, there is relatively little cross-component authentication in use at ICPSR, but one important instance occurs between software systems and the ICPSR's Oracle database. In this case a generic database user and password are used (which could be possibly "buried in a perl library"). Additionally, communication channels between CGIs and the staging server may not be encrypted.

Authorization

Because of the varying types of users, as well as privacy restrictions on data, authorization also exists in many different contexts at ICPSR. A typical authorization scenario at ICPSR involves an identity (or role) and a resource in some technology context. Examples include: UNIX file permissions dictating what files and directories may be accessed by that user (UID) or that role (or group, GID); samba share permissions dictating what identities may access a shared drive; MyData roles dictating what portions of the Intranet may be accessed by a member of the ICPSR staff; etc. One notable exception to this is that ICPSR uses IPv4 address as the authorization mechanism for user access to the bulk of its website. Stated briefly, there is a privileged collection of IPv4 address blocks that are allowed to access ICPSR's entire service archive holdings; IPv4 addresses from outside that collection are allowed to access only a subset of the ICPSR holdings. This collection is maintained by ICPSR in cooperation with its members. Access control policies are specified within the Data Library in free-form text, which is

structured in that processors choose from a fixed set of choices. Some aspects of access control are specified within software, rather than within administrative metadata.

Anomalous network activity is reported by two different systems. The University of Michigan Network Operations Center monitors the campus network (including the portion used by ICPSR) for a handful of unusual network traffic signatures (e.g., large bursts of outbound SMTP), and reports incidents in real-time. If initial investigation uncovers something interesting, the IT team will open a trouble ticket to track the issue. The ICPSR also uses a product called "Peakflow X" to monitor network activity at the "flow" level, and this product also reports anomalous activity, but using a much broader set of tests than the above system.

Only depositors may gain access to both physical (original data components deposited) and digital data. Such database permissions are parsed out to Computing and Network Services and the Data Library through layered permissions. Only one person can make changes to this structure.

3.3.3.3 Data Deposit & Ingest

Formats

To the extent possible, the ICPSR controls the data and file formats they will accept. Much of this is predicated on the fact that research data and documentation tend to be generated with common, known statistical data packages. Grant proposals often dictate the kinds of information (including codebooks) that must be generated as a part of funded projects. These requirements benefit the ICPSR because in general, they provide a level of uniformity to the types of files likely to be produced and therefore submitted for archiving. At the same time, not all commonly used formats are acceptable, durable formats and so the ICPSR performs some format normalization as one of their digital preservation strategies.

The ICPSR accepts MS Word documents, although it does not consider this a canonical format. They generate PDF files from MS Word, and they are considering PDF/A when it is ready for production use. The Digital Preservation Officer addresses the issue of canonical formats, including multimedia, which represents a growing presence in data submissions. ICPSR expressed its intention to use format registries, such as the Global Digital Format Registry (GDFR) when they become available.

Data Deposit

At the time of submission, depositors are asked about sensitive information (e.g., an Excel file with social security numbers). Processors report the suspected presence of sensitive information to a supervisor, who then works with the depositor to remove such information. If removal is not possible, then a restricted version of the data set is created. Data is copied from processors' PCs onto servers, and shipped to the Data Library for DLT via software called Acquire; there was some discussion about whether these transfers were conducted in the clear (i.e., without encryption). The Data Library and the Web server are not connected, but there is NFS cross mounting between the staging server and the public server. There is no known vulnerability or security problem with this arrangement. Sensitive files are not on cross-mounted file systems between the two servers.

ICPSR uses an acquisition form, similar to a manifest, which captures relevant information from a data depositor. All physical documents, such as codebooks, are scanned as a part of the ingest process. Metadata about information is logged, so if there is a loss, there would be some record of what documentation was missing. Items are marked and tagged for proper folders, which correspond to data study folders.

Metadata Capture
Processors fill out a template that captures relevant study-level metadata, and submit these completed forms to an editor, who verifies the metadata on the basis of the deposit form and returns it to the processor. [For more information about all metadata captured throughout the pipeline process, see Section 3.3.3.6, *Data Management (metadata, logs, etc.)*] There is no direct handoff of data between processors—there is mandatory centralized review and verification. Metadata records are captured in "SPIRES" format and placed within an Oracle database. An XML record is derived from these metadata records. The Data Library uses the Oracle database and flat files stored in a single Solaris 8 UFS file system, both of which reside on the web server. There is documentation for the Oracle database schema and data dictionary.

While ICPSR does not earmark metadata specifically as "preservation metadata" there are several elements from the acquisition form and deposit procedure that could support preservation. For example, information on the submission package, down to the file level (e.g., name, format, version of software), types of files, specifications of files (e.g., length, block size, number of lines), issues during ingest, processing history, etc. is captured. ICPSR uses a different form when it purchases data, and there are some variations within topical archives, but the overall process is generally consistent.

Validation
ICPSR validates content at two stages. In the beginning of the pipeline, after materials are acquired but before they are processed, ICPSR adds a pair of snapshots (basically encapsulated, authentic, archival copies of the files as submitted) to its Data Library file. At the end of processing, ICPSR adds a second pair of snapshots to its Data Library. Invalid objects (e.g., a PDF file that will not render properly) are routinely ingested and archived. Fixity information is retained via the MD5 checksum, but there is room for improvement related to retrospective generation of checksums. The pair of snapshots (multiple copies) is routinely evaluated for incongruities; if any are found (e.g., bad tape), a new set of snapshots are made. Compression is used for the "circulation copy", but not for the copies within the Data Library.

There is relatively recent usage of MD5 as a checksum, especially with new data submissions. There is discussion about how to generate checksums for older studies, but there was yet no formal plan at the time of the audit.

Identifiers
Each study is assigned a unique five-digit identifier, which is treated as an integer. There are possible dependencies on this identifier that might affect the processing software and the Data Library (e.g., database fields that are numerical). At the time of the audit ICPSR was considering the Harvard-MIT VDC "fingerprints" as a next step for identifiers.

3.3.3.4 Archival Storage

The archival collection is routinely migrated from medium to medium (with requisite file format migrations required by the media migrations) but otherwise is used only when a dataset in the servicing collection fails. Two copies of each file in the archival collection are stored off-site on DLTs. (See 3.3.3.2, *Data Security*, for more information on the backup copies of the archival files.)

3.3.3.5 Preservation Planning (Strategies)

General

ICPSR has discussed preservation planning and strategy, and has extensive experience with data management and file migration. A draft version of the Preservation Policy is available on the ICPSR Intranet. ICPSR also contracted with Council Member Nancy McGovern (Cornell University) – since hired as the ICPSR Preservation Officer – to conduct an analysis of the ICPSR systems and identify areas of risk. That report was made available to the auditors while onsite and is appended to this report as Appendix 4.4, *ICPSR Digital Preservation Review*. Further policies and more comprehensive, coherent preservation strategies were expected to be developed by the new Preservation Officer.

Specific: Migration

At the time of the audit, there was no policy for a timeline of migration, but media is migrated every ten years. For prior migrations, which have typically revolved around specific tools or media, ICPSR has followed NARA guidelines. The policy calls for 100% compliance and transfer. All previous migrations have been successful with the exception of the one instance of unrecoverable data loss mentioned earlier. (This loss did not involve unique data and therefore ICPSR was able to obtain a new copy of the collection from another data archive to recover from this data loss.) If there is a major technological change, that time period can be reduced to migrate to a new type of media (for instance, ICPSR is currently moving from 9-track tape to DLT). ICPSR estimates that with 3 FTEs, all tapes could be migrated in three months.

There is some question about migration of data loader or application software. For example, while it would be possible to determine when a specific version of software (e.g., SPSS) was installed, it would be time consuming or onerous to determine which studies, or files within studies, were processed using that particular version. This kind of preservation metadata is important to capture and maintain because if an instance of corruption caused during migration were detected after the fact, there would be no automated way to quickly identify affected files so that the migration processed could be rerun on only those files. The ICPSR staff acknowledged this and recognized the need for a process change to enable the automatic capture and recording of this migration data.

3.3.3.6 Data Management (Metadata, Logs, etc.)

Management of Archival and Service Files

Backup tapes from the Data Library are loaded and tested periodically, often prompted by user requests or migration needs. These tapes are generated at the time of processing and a diff check is conducted (and log is examined) to ensure consistency across tapes. The Data Library database is exported from Oracle into a file system and stored on tapes similar to disaster recovery tapes. This is standard, good practice for the management of off-line archival copies.

Data migration is conducted using a program with built-in error reporting capabilities. ICPSR also performs manual checks by looking into file specifications and content lists. Media errors are detected during use and corrected with its partner media by downloading the good tape and copying its content to a fresh partner tape. From DLT to DLT, the ICPSR has a 100% correctable rate. From tape to tape, ICPSR has a 99.5% correctable rate and from tape to DLT, less than 0.4% failed due to tape error. According to the ICPSR, there has been only one exceptional case when an external entity (at MIT) was contacted for a backup copy.

In the context of the delivery function, ICPSR sets an implicit "next business day" recovery goal for its web site, and this is reflected in the service level agreements and maintenance contracts the ICPSR has established with key vendors. (ICPSR intends to modify this goal to "same business day" in FY07.) Recovery time would also be measured like availability. In order to bring the ICPSR website back online from a disaster recovery, the backup tapes and Web presence would have to be loaded on another University of Michigan server(s). To the auditors' knowledge, this service recovery plan has never been tested and testing of service recovery was encouraged by the auditors to discover any possible procedural issues when no crisis exists.

Metadata:
At ICPSR, they distinguish four types of metadata that the repository creates, preserves, and distributes:

- *Study-level metadata*, also known as abstracts or study descriptions or metadata records. These constitute the highest level of metadata, describing the study or collection as a whole; it is primarily intended for resource discovery purposes.
- *File-level metadata*. These describe the properties of individual files in a data collection.
- *Variable-level metadata*. These are detailed in technical documentation such as codebooks and data definition statements (DDS) and are essential to effective and accurate interpretation and use of the numeric and character data.
- *Administrative and structural metadata*. These are critical to ongoing maintenance and preservation of the electronic data collections. They must be complete enough to permit future archivists to discern how the files were produced and how they might be migrated or emulated in a new technological environment.

At given points throughout the processing pipeline, metadata is re-keyed or copied and pasted multiple times in varying formats. This creates numerous opportunities to introduce error. It is also a challenge to ensure that all instances of the metadata remain synchronized when even a minute change occurs (often requiring repeated manual editing). At the time of the audit ICPSR was implementing a Tracking System to centrally capture, track, and reuse metadata throughout the process. This would assist ICPSR staff and users, and was a recommendation from the *Process Improvement Committee – Pipeline Redesign* (October 2003). At the same time, the committee recommended that ISCPR continue to utilize the Data Documentation Initiative (DDI)

XML specification for social science metadata as the desired format for catalog records. All study descriptions are now fully compliant with the specification and the ICPSR continues to play a leadership role in the development of the DDI.

3.3.3.7 Access Management

Access Policy

The ICSPR User Guide (http://www.icpsr.org/org/publications/ICPSR_User_Guide.pdf) is a publicly available document which provides a brief overview of the products and services ICPSR offers. ICPSR access policies are based on several factors: basic data access, but also access based upon issues of data confidentiality. See Section 3.3.2.4, *Content Accessibility*.

Basic access to descriptive information (the catalog or website searches) is open to any visitor to the ICPSR website. Access to the actual datasets and collections -- especially through ICPSR Direct and MyData – invokes established authorization & authentication controls.

Individuals affiliated with member institutions have the greatest privileges and most direct access to almost all unrestricted processed studies. Non-members must negotiate with ICPSR staff and generally pay a fee to obtain studies. Almost all ICPSR datasets are accessible to all groups in this manner. Access to restricted data requires further authorizations and is accomplished off the network. For some sensitive data, access is provided only via removable media (i.e., CD-ROMs) which must be destroyed at the end of use. For the most secure/sensitive data, onsite use in the Data Enclave is required.

The sensitivity level of the data along with the guidance of the principal investigator determines the mode of access for restricted collections. Weighing these factors, ICPSR makes some restricted data available on CD only and other restricted data available only on-site through the Data Enclave. The study descriptions note data restrictions for a study, if any, and provide access details.

Service Levels & Performance Expectations

In the context of the delivery function, ICPSR does not set specific response time goals for web access. In the context of the retention function, ICPSR's Data Library maintains a goal of two-business days response time for requests for data not available through MyData or ICPSR Direct.

In the context of the delivery function, ICPSR sets a 99% availability goal for its Web site. The University of Michigan Network Operations Center monitors availability on a 24 x 7 basis, and reports unavailability in real-time. In the context of the retention function, ICPSR's Data Library maintains a goal of "normal business hours" for availability.

3.3.3.8 Business Continuity, Environmental Management, and Disaster Planning

Environmental Management

The ICPSR machine room – which houses ICPSR's web server and a staging area for Data Library content – is onsite in an environmentally controlled space. Two dedicated air handlers provide cooling and humidity control. Two APC Matrix UPS systems, which were located

directly on the floor, provide power conditioning and approximately one hour of stand-by battery power. There is no generator, and during an extended power outage, the ICPSR machine room would simply be without power. Since occupying the building in Dec 2002, this happened once, in August 2003. Computing & Network Services staff indicated that although they did lose power and were unable to provide access to their services during this multi-state power outage, the temporary lack of environmental controls in this space during this power outage had no lasting adverse effect on the machines and systems.

The temperature of the Data Library's off-site storage facility (the site of backup tapes and some paper files) is 65° F and is monitored by a standard thermostat. The humidity is monitored but at the time of the audit there was no way to control it. The temperature is checked during warehouse visits. The Storage Warehouse has neither secondary power supply nor backup generator in case of power loss. There were no communication devices there at all. Nor were there any equipment or appliances present there that required electricity.

There are water pipes that could affect the machine room, especially above a tape reading station that is used (infrequently) for older tape formats. There have not been any water-related reported problems.

Access to the server room, network and backup tapes is limited to Computing and Network Services (CNS) and Data Library staff with card swipe access. Anyone with a master key has *unmonitored access*.

For purposes of security the location of the ICPSR warehouse is not widely known to most staff. The building has no distinguishing logos to call attention to it as an ICPSR location, and all windows are covered. Doors to each section are locked and without the proper keys one is denied access to other areas. All file cabinets and media cabinets are locked, and all file location information is maintained in Oracle tables with access only granted to Data Library staff.

Business Continuity

ICPSR's *Business Continuity Plan for ICPSR Services* (August 2005) serves as a reference document for ICPSR partners and other organizations that contract with ICPSR for the delivery of online services. Clearly, this document applies to the service/dissemination

In the context of the delivery function, ICPSR sets an implicit "next business day" recovery goal for its web site, and this is reflected in the service level agreements and maintenance contracts ICPSR has established with key vendors. (ICPSR intends to modify this goal to "same business day" in FY07.) Recovery time would also be measured like availability.

Disaster Planning

A noteworthy aspect was the lack of a formal disaster plan for ICPSR. Several elements of a disaster plan can be found in existing documentation such as the *Business Continuity Plan for ICPSR Services*, but no single disaster plan or disaster response manual exists. The creation of a comprehensive and cohesive plan would highlight the need for changes identified during other areas of the audit (e.g. both backup copies of the archives exist within a 2 mile radius of each

other – should a regional disaster hit, both copies of the Archival collection as well as the Service Library would be damaged or lost).

It is recommended that a disaster plan be created as soon as possible. This will highlight other risks and vulnerabilities, some of which can be resolved with additional documentation. Others will require new policies and procedures to be developed. This should be an immediate priority for the incoming Digital Preservation Officer.

3.3.4 Vulnerabilities

3.3.4.1 Significant Repository Events

In the past, there was one instance of data loss from which recovery was not possible.[6] This sole instance involved a collection that was not unique to ICPSR. Because of this, the ICPSR was able to obtain a copy from another data archive and reload it into the system. No other instances of nonrecoverable data loss have occurred.

3.3.4.2 Liabilities

ICPSR is dependent upon the University of Michigan for key infrastructure and capital facilities and a large percentage of its budget derives from grant sources.

3.3.5 Final Observations and Recommendations

It is easy to understand why ICPSR has such an impressive track record as a repository – as a long-standing organization and a trustworthy digital archive.

ICPSR has a long history of reliable and stable service to the user community. Current leadership, under the direction of Dr. Myron Gutmann, has brought about an era of membership growth, as well as fiscal growth. These leadership contributions have led to an impressive degree of economic sustainability for an organization with such a specialized constituency. ICPSR brings a tremendous amount of expertise and thought to bear on its processes for data deposit, ingestion and preservation. Moreover, much work has been done toward articulating and documenting these processes and the specific roles that support them.

Our first recommendation is to build upon that work, to develop and explicitly articulate policies and documentation. For example, while there have been no problems with data backup or recovery, at the time of the audit there was is no policy for notification of problems or for identification and prioritization of tapes for transfer to the offsite facility. Additionally, there was no policy regarding the location of the Data Library metadata. The backups of these metadata do not appear to be stored in the same manner as the content components of the Data

[6] Interview with Sheila Grindatti, Manager of the Data Library, and Cole Whiteman, Acting Director of Data Security and Preservation Division, 31 January, 2006.

Library. By creating this additional documentation, ICPSR staff will achieve a greater shared understanding of the ICPSR operations and environment, and greater potential for obtaining diverse, comprehensive feedback. Another area where greater documentation is needed is the procedure for the physical rotation of the tapes and the corresponding log files.

A second recommendation pertains to uniformity in the terminology used to describe critical ICPSR infrastructure. It was apparent to auditors that staff members at ICPSR refer to the same technical entities using different terminology, even though ICPSR manages an internal glossary on the organizational intranet. For auditors, it was both perplexing to not only sort out particular discussion topics, but also see staff similarly struggle through the same exercise. In several cases, one entity under discussion was referred to by three different internal designations. This is true of the designation of the master archive of ICPSR data as the Data Library (which is also a department within the ICPSR) and the use system as "The Archive." While this could be dismissed as variations based on staff longevity at ICPSR or a simple semantics issue, in reality it is poses risks that could potentially affect the system and the data it manages. To manage potential risks and vulnerabilities as well as allow more transparent, easy understanding of the ICPSR systems, it is recommended that the ICPSR reconsider the terminology it uses to adjust certain descriptors as necessary.

A third recommendation focuses on automation. There are instances of double keying of metadata, which introduces the possibility (albeit remote) of inconsistent data and inefficiency. Greater automation in the pipeline would make it easier to reprocess data if there is a problem in the pipeline.

Additional recommendations of a specific nature (grouped by functional area) include:

- Introduce precautions against water damage in the server room, such as raising UPS above floor level, and introducing environmental monitoring sensors, especially to detect the presence of water.
- Develop a disaster plan and investigate the impact required changes will have on other processes and procedures.
- Periodically audit the entry to secure areas (e.g., server room, enclave)
- Examine the PREMIS preservation metadata standard for comparison (and possible augmentation) to existing, captured metadata to allow for better management of the digital objects in the long-term
- Develop a mechanism and procedure for checking between the Data Library database and files on tape backups to ensure that metadata is still consistent between copies.
- Continue to develop funding where possible to reduce the reliance on grant funding.
- Build preservation costs into contracts with large data providers and government contracts.

Finally, ICPSR mentioned that the soon to be hired Digital Preservation Officer will consider:

- Canonical file formats
- Unique identifiers
- Use of checksums

These recommendations are meant to ensure that ICPSR continues its excellent tradition of providing excellent, trusted service to its community in the future.

Appendices

4.1 ICPSR Organizational Chart

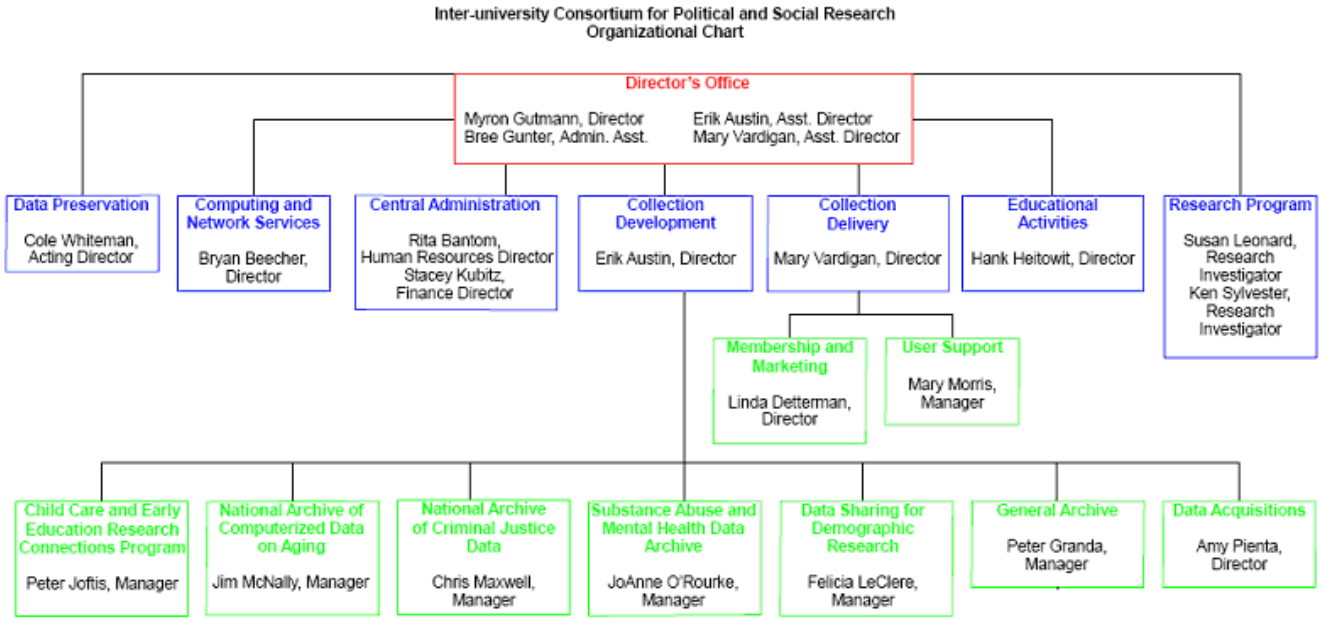
4.2 ICPSR Council (at time of audit)

4.3 ICPSR Staff Members Interviewed

4.4 ICPSR Digital Preservation Review (N. McGovern report)

4.5 ICPSR's Current Pipeline Process: the Insider's View, Detailed

Appendix 4.1
ICPSR Organizational Chart



Appendix 4.2

ICPSR Council Membership (at time of audit)

Name	Institution	Term
Mark Hayward, Chair	University of Texas	3/2002-2/2006
Darren W. Davis	Michigan State University	3/2004-2/2008
Ilona Einowski	University of California, Berkeley	3/2002-2/2006
Charles H. Franklin	University of Wisconsin	3/2004-2/2008
John Handy	Morehouse College	3/2002-2/2006
Paula Lackie	Carleton College	3/2004-2/2008
Nancy Y. McGovern	Cornell University	6/2004-2/2006
Samuel L. Myers Jr.	University of Minnesota	7/2004-2/2006
James Oberly	University of Wisconsin, Eau Claire	3/2004-2/2008
Ruth Peterson	Ohio State University	3/2004-2/2008
Walter Piovesan	Simon Fraser University	3/2004-2/2008
Ronald Rindfuss	University of North Carolina, Chapel Hill	3/2002-2/2006

Appendix 4.3

ICPSR Staff Interviewed During the Audit

Erik Austin, Assistant Director, ICPSR

Rita Bantom, Human Resources Director

Bryan Beecher, Director, Computing and Network Services

Peter Granda, Assistant Director, Collection Development

Sheila Grindatti, Assistant Manager, Data Preservation (Data Archive)

Myron Gutmann, Director, ICPSR

Stacey Kubitz, Finance Director

Asmat Noori, Assistant Director, Computing and Network Services

Mary Vardigan, Assistant Director, ICPSR

Cole Whiteman, Acting Director, Data Security and Preservat

Appendix 4.4

ICPSR Digital Preservation Review (N. McGovern report)

ICPSR Digital Preservation Review
Prepared by Nancy Y. McGovern
Based on discussions during a visit to ICPSR April 17-18, 2005

Overview

This report is the first step in a formal review of ICPSR's digital preservation approach. The objective of the review is to complete, at minimum, the self-assessment portion of the emerging Trusted Digital Repository certification process.[1] Ideally, ICPSR will seek full certification when the process is finalized.

These are some observations about ICPSR and digital preservation that serve as the basis for this report:

- ICPSR has a good track record retaining their published datasets over time and has been an innovator in receiving, processing, and delivering data (including legacy data) to users.
- As a data archives, ICPSR's operational program is appropriately focused on providing access to data. However, that focus creates a potentially problematic perspective and circumstance for digital preservation at ICPSR. The original materials are received and processed; the original versions are generally considered to be the archival masters in archival programs. The processed versions are brought into the distribution process; the processed versions are generally considered to be the access copies in archival programs. The access copies at ICPSR have archival value and need to be preserved, as do the archival masters.
- Processing of the data is required (and desirable) to remove personal identifiers, convert data and documentation to usable formats, ensure consistency and accuracy, etc. Yet, the originals contain unique information that is often not available elsewhere and may be needed at some point in the future to replace or revise the access copy. In some ways, the archival masters are set aside once processing proceeds.
- The access copies are systematically managed and this management process would score well on most digital preservation criteria, though there are some recommendations for improvement. The archival masters are retained, but not as comprehensively managed. These files (in both paper and digital formats) have potential risks that should be addressed. Some of the formats and the media these files are stored on are or are becoming obsolete; not all of the original documentation has been able to be made digital, for various reasons; there is not adequate redundancy for the archival masters, which are stored offline. If the scope of the current NDIIPP project can include these at risk files, additional funding should be sought to address these concerns.
- Like most institutions, ICPSR does not yet have a fully-implemented digital preservation program. The current situation for digital preservation could be described as project-based; that does not form the basis for a sustainable program. One objective should be to develop a program that is appropriate to the requirements of both archival masters and access copies. ICPSR does engage in good practice and some staff at ICPSR are vigilant about both the preservation of the archival masters and the access copies.

This report includes[2]:

- A gap analysis of ICPSR's digital preservation approach based on the Trusted Digital Repository attributes. This section consists of observations, recommendations, and questions.
- A summary of the results using the three aspects of a digital preservation program: organizational infrastructure, technological infrastructure, and resources framework.

[1] See the certification website at: http://www.rlg.org/en/page.php?Page_ID=580.
[2] This set of metrics (the TDR gap analysis, the three aspects of digital preservation, and the maturity model) was devised at Cornell University Library for the Digital Preservation Management Workshop (see the workshop website at: http://www.library.cornell.edu/iris/dpworkshop/.

- A concluding analysis based on the Five Organizational Stages of Digital Preservation maturity model.

Trusted Digital Repository (TDR) Attributes

OAIS Conformance: This is the underlying principle of the TDR framework. The organization commits to implementing a repository that conforms to the OAIS reference model.

Status: ICPSR explicitly made this commitment to OAIS in a recent article entitled: "OAIS Meets ICPSR: Applying the OAIS Reference Model to the Social Science Context" written by Mary Vardigan and Cole Whiteman. ICPSR should use that article as a starting point to conduct a deeper audit of their digital preservation progress using the OAIS functional entities. The draft of the digital repository certification document will be released at the beginning of July 2005. That document could provide the basis for the audit.

1. ***Administrative Responsibility***: The organization makes a fundamental commitment to digital preservation. The criteria for this attribute demonstrate that the organization:
   - Is able to provide evidence of a fundamental commitment to implementing community-agreed standards, best practices
     Status: ICPSR demonstrates an awareness of and intent to comply with community standards. Determining the implications of the recent release of PREMIS[3] for ICPSR will become a priority as the digital preservation community responds to this important document. Are there plans in place to address the implications of PREMIS?
   - Meets national/international standards on environment
     Status: These standards apply to the environment in which equipment and storage media are maintained. If this is not already in place, ICPSR should designate staff to confirm that the current environment complies, document that compliance, and establish a regular review.
   - Meets or exceeds community standards and share measurements with depositors
     Status: The ICPSR website makes information like this available. Perhaps a designated section for digital preservation could document activities and progress. How is this information currently documented and made available?
   - Involves external community experts in regularly validating/certifying processes and procedures
     Status: ICPSR's initiation of this review demonstrates their intention to seek external review. That intention must extend to addressing concerns and recommendations as well.
   - Has written agreements with depositors that address all appropriate aspects of acquisition, maintenance, access, and withdrawal.
     Status: ICPSR's policies for depositors are explicit. Scheduled review of policies and agreements should ensure compliance with evolving requirements. Are additional policies and procedures needed?
   - Ensures that ongoing risk management and contingency planning play a routine part of the organization's annual strategic planning activities.
     Status: ICPSR actively engages in both risk management and contingency planning for data and the technology infrastructure. They need to ensure that their ongoing risk management and contingency planning extends to the full range of digital preservation requirements.
   - Commits to transparency and accountability in all actions
     Status: The ICPSR operation attests to this commitment. Care should be taken to ensure that this commitment continues and that current policies and practice are readily accessible.

[3] PREMIS is the acronym for Preservation Metadata Implementation Strategies, an international working group convened by RLG and OCLC that just released a data dictionary and final report. See http://www.oclc.org/research/projects/pmwg/.

2. ***Organizational Viability***: The organization has the authority, mandate, resources, and organizational infrastructure for digital preservation. The criteria for this attribute demonstrate that the organization:

   - Demonstrate viability and trustworthiness
     Status: ICPSR has a long and respected track record for sustaining its operation. Stakeholders clearly trust ICPSR to be the steward of essential data.
   - Reflect commitment to long-term retention/management in mission statements
     Status: ICPSR's mission statement does include an explicit commitment to preserving and archiving data.
   - Have appropriate legal status, staff, and professional development for responsibilities
     Status: ICPSR has the legal status to preserve its holdings. The staff are very capable, but specific digital preservation training and development is not institutionalized. This is especially important with the recent retirement of the preservation manager. ICPSR should have a digital preservation officer. Development plans should be in place for key staff.
   - Establish transparent business practices, effective management policies
     Status: ICPSR is open in its practices and has an established and effective management structure. Digital preservation should be elevated to a core function not an ad hoc activity and care taken to ensure that a full program is implemented.
   - Define comprehensive written agreements with depositors
     Status: The deposit process provides an explicit agreement with depositors.
   - Review and maintain policies and procedures
     Status: While ICPSR has comprehensive procedures for depositors, for processing, and other lifecycle stages, higher level policies are not as comprehensive or explicit as needed. This is an immediate concern to address. An explicit, high-level digital preservation policy is needed. A scheduled review process is also required.
   - Undertake risk management, contingency and succession (trusted inheritors) planning
     Status: ICPSR meets this criteria for its technology infrastructure. ICPSR should actively extend that compliance to all aspects of digital preservation requirements: redundancy, disaster planning, etc.

3. ***Financial Sustainability***: This attribute attests to the fundamental financial commitment by the organization to digital preservation. The criteria for this attribute demonstrate that the organization:
   - Establish and maintain good business practices and an auditable business plan
     Status: ICPSR meets this criteria for its overall operation; specific practices and the business plan for digital preservation are not as clear.
   - Demonstrate financial fitness and ongoing financial commitment
     Status: ICPSR has a sound financial approach for its overall operation. Financial support for digital preservation should include designated support to establish a full digital preservation program and maintain that program over time.
   - Balance risk, benefit, investment, expenditure
     Status: ICPSR takes a holistic approach to investments. This approach needs to incorporate preservation planning and technology monitoring to be able to anticipate and respond as technology evolves.
   - Maintain adequate budget and reserves and actively seek potential funding sources
     Status: The current ICPSR budget does not clearly enough identify the nature of expenditures that are specific to digital preservation. The costs currently seem to address only salaries. Combining preservation with data security possibly makes the financial commitment to digital preservation less clear.

4. ***Technological Suitability***: This attribute ensures that the organization is able to identify and implement appropriate digital preservation solutions. The criteria for this attribute demonstrate that the organization:

   - Consider and adopt appropriate preservation strategies
     Status: ICPSR routinely plans for data conversions and migrations. As more complex digital objects are acquired (e.g., audio, video), they are aware that preserving these materials will be more difficult. This is an area for growth and development for ICPSR.
   - Ensure appropriate infrastructure (hardware, software, facilities) for acquisition, storage, access
     Status: ICPSR has a reliable infrastructure in place. Keeping up with changing technology will be a challenge, as it is everywhere. Considering storage in light of the archival masters and access copies discussion at the beginning of this report. Sharing redundancy responsibilities with other data archives is an option to consider.
   - Establish a technology management policy for the repository (replacement, enhancement, funding)
     Status: ICPSR does technology planning well. ICPSR needs to ensure that this planning addresses the full range of digital preservation requirements as well as access requirements.
   - Comply with relevant standards and best practices (supported by adequate expertise)
     Status: ICPSR is actively addressing this criteria in technology areas with recent certification training, for example, and is seeking to do so for digital preservation.
   - Undergo regular external audits on system components and performance
     Status: Regular external audits should be established if they have not been already, and the results, while not releasing security essentials, should be readily available.

5. ***System Security***: This attribute addresses the ability of the organization to provide a secure environment for the digital preservation program throughout the lifecycle of the digital content. The criteria for this attribute demonstrate that the organization:
   - Assure security of systems for digital assets
     Status: This is a high priority for ICPSR, especially for access copies. Archival masters are maintained offline. ICPSR should ensure that security is optimal for archival masters as well.
   - Establish policies and procedures to meet requirements (copying, authentication, firewalls, backups, disaster preparedness, response, recovery, training)
     Status: ICPSR is expanding and updating its procedural manual. Many policies and procedures are in place. An audit should be conducted on the results and regularly scheduled to ensure that all of these areas are comprehensively and correctly covered.
   - Stress processes that will detect, avoid and repair loss, document and notify about changes and resulting actions
     Status: What are ICPSR's current arrangements in this area?

6. ***Procedural Accountability***: This attribute confirms that the organization comprehensively documents its policies, procedures, practices, and actions, and provides the basis for the organization to seek certification. The criteria for this attribute demonstrate that the organization:
   - Enact all relevant policies and procedures for specified tasks and functions, document all practices
     Status: ICPSR has a very explicit and well-documented process for acquiring and processing data. This process should be reviewed to ensure that digital preservation practices are also comprehensively included.
   - Establish monitoring mechanisms to ensure continued operation of systems and procedures
     Status: What is ICPSR's current capability in this area?
   - Record and justify preservation strategies
     Status: In what ways and to what extent does ICPSR meet this criteria?

- Set up feedback mechanisms to support problem resolution and negotiate evolving requirements between providers and consumers
  Status: ICPSR has effective mechanisms for this process. Establishing the digital preservation program may lead to additional interactions with producers and consumers to identify and address evolving expectations and requirements.

Summary Review of Digital Preservation at ICPSR

Organizational Infrastructure is best defined by the Trusted Digital Repositories framework and best reflected in policy development, implementation, and preservation planning.

The ICPSR operation has developed and maintained policies and procedures. Ensuring that these policies and procedures extend to digital preservation should be a high priority for ICPSR. This is especially true for the development of a specific digital preservation policy to set out the purpose, mandate, objectives, scope, operating principles, roles and responsibilities, selection and acquisition, preservation activities, access and use, challenges and incentives, and collaboration efforts. There are some good examples from other institutions that might serve as a starting point for the development of this key document.

Preservation planning is the one functional entity of the Open Archival Information System model that is not specific to an OAIS implementation, but shared by the digital preservation community. The preservation planning functions via the administration functions are the bridge between the OAIS (or digital preservation repository) and the organizational infrastructure. ICPSR needs to ensure this portion of the operation is as vital as the access operation.

Technological Infrastructure is best defined by the OAIS reference model and includes the combination of hardware, software, network protocols, and technical skills the organization brings to the development of its digital preservation program.

ICPSR has a skilled technical team and has ample skills to support the pipeline process – study processing and preparation. More attention may be needed to provide continual development of digital preservation skills.

ICPSR has a strong technological infrastructure to sustain its access operation. ICPSR should make it a high priority to audit the current infrastructure in place for the archival masters, identify at risk materials, formulate a plan to address these materials, and incorporate these materials more cohesively into operational planning. This is a potential current gap in the capability of ICPSR to be certified as a trusted digital repository and in demonstrating conformance to OAIS.

Six copies are recommended for effective preservation redundancy. ICPSR could more easily achieve this objective by collaborating with other data archives. Redundancy is a common objective. One possibility would be to consider the applicability of LCOKSS for such a consortial approach. Could LOCKSS work for both archival masters and access copies? That is a key question to address.

Resources Framework includes the funding, staff, technology, space, and other resources needed to sustain an organization's digital preservation program.

As previously stated and like most other institutions, ICPSR does not have a fully-implemented digital preservation program. The current status could be characterized as project-based. ICPSR needs to devise a plan for developing a complete program that fully conforms to OAIS, specifically designate sustainable funding, prepare for contingency costs of migrations and other preservation activities, and seek additional funding as needed. The NDIIPP project is a good model, but those activities may not be sustainable beyond the life of that project.

ICPSR and the Five Organizational Stages of Digital Preservation

The Five Stages of organizational development for digital preservation are:

1. *Acknowledge*: understanding that digital preservation is a local concern
2. *Act*: initiating digital preservation projects
3. *Consolidate*: segueing from projects to programs
4. *Institutionalize*: incorporating the larger environment and rationalizing programs
5. *Externalize*: embracing inter-institutional collaboration and dependency

ICPSR is beyond the first stage, acknowledge. They are well aware of the issues, the requirements, and the fundamental need to undertake digital preservation. ICPSR has attained the second stage, act. There are increasing examples of ICSPR initiating digital preservation projects. ICPSR is moving toward the third stage, consolidate. A digital preservation policy, a proactive approach to technology planning for preservation, and designated sustainable funding for digital preservation are necessary to achieve this stage. At stage three, an organization has a baseline digital preservation program. ICPSR should seek to fully achieve stage three development and conduct a gap analysis to develop a plan to achieve stage four, institutionalize. Not many institutions have achieved this stage. Like many other institutions that are active and involved, ICPSR exhibits some stage 5 indicators, externalize. ICPSR has effective partnerships in place. Organizations must achieve a full digital preservation program at stage four before stage five can be completely realized. Collaboration on digital preservation is difficult if not impossible at earlier stages, though several stage five organizations can effectively incorporate the efforts of lower stage organizations if the roles and responsibilities are well-defined.

Appendix 4.5

ICPSR's Current Pipeline Process: the Insider's View, Detailed