The Center for Research Libraries

10/14/2005

Archive Profile

Inter-university Consortium for Political and Social Research (ICPSR)

by Robin Dale, Project Director, Certification of Digital Archives Project and Program Officer, Research Libraries Group, US

Organization

Mission and Governance

The Inter-university Consortium for Political and Social Research (ICPSR), established in 1962, is an integral part of the infrastructure of social science research. ICPSR maintains and provides access to a vast archive of social science data for research and instruction, and offers training in quantitative methods to facilitate effective data use. To ensure that data resources are available to future generations of scholars, ICPSR preserves data, migrating the collection to new storage media as changes in technology warrant. In addition, ICPSR provides user support to assist researchers in identifying relevant data for analysis and in conducting their research projects.

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. A Council of leading scholars and data professionals guides and oversees the activities of ICPSR. See Appendix 1 for a complete 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 2.

Funding

As a member-based organization, ICPSR charges annual membership fees to generate revenue. Payment of the membership fee provides each member institution with campus-wide access to the ICPSR data holdings. The membership costs are based on categories that are tied to the Carnegie Classification of Institutions of Higher Education. Nonmember access to data is possible and is managed on a fee basis. Together, membership and access fees generated more than $3.4 million in income in Fiscal Year 2004-2005. ICPSR also generates revenue from archiving and servicing for contract clients, many of whom are federal government departments. (An example is the crime and justice archives of the Department of Justice.) Additional revenue in the amount of approximately $5.8 million was gained through gifts and grants in FY05.

Preservation of the “data archive” is the responsibility of the Data Preservation Unit. The 2006 direct budget for that department is slightly more than $341,000 in “designated costs”; more than $232,000 of that represents staff costs. These designated costs are hard funding and are recovered primarily through membership dues. Overall, the funding for the Data Preservation department is 13.8 percent of the membership-funded expenses of the FY06 budget.

Agreements with Data Depositors

ICPSR works with and encourages social scientists in all fields to preserve their research data. By agreeing to deposit data with ICPSR, researchers are able to use ICPSR not only for redistribution and reuse of their data, but also for “long-term safekeeping of data, protecting it from obsolescence, loss, deterioration, or irreversible damage” (*“Why Should I Archive Data?”* 2005). While depositors voluntarily enter into deposit agreements with the ICPSR, submitting data to ICPSR is also seen as a method for fulfilling granting agency obligations to archive data once a funded project is completed.

Depositor agreements must be submitted and approved before ICPSR will accept any data. The *ICPSR Data Deposit Form* (http://www.icpsr.org/access/deposit/data-deposit-form.pdf) is the mechanism used by ICPSR to gather administrative and descriptive metadata about the data. Depositors receive assistance in providing the correct and complete information through the supplemental deposit information found in the ICPSR publication, *Guide to Social Science Data Preparation and Archiving: Best Practices Throughout the Data Life Cycle* (2005).

The deposit agreement is also the vehicle for transfer of both distribution and preservation rights from the depositor to the data archive. A depositor can stipulate “preservation-only” rights for a period of time (no more than one year), but all data are accepted with the intention of making them accessible to ICPSR members and users. In return, the archive assumes responsibilities to prepare the content for access (including enhancing and reformatting of data) and preserve it for the long term (including periodic migrations). Multiple publications and fact sheets available through the ICPSR Web site reiterate the ICPSR’s acceptance of this preservation commitment.

ICPSR Archive System

Content Characteristics

Content characteristics of the data archive are readily known because of the data archive’s policy on acceptable data formats for deposit and preservation. Data 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. Figure 1 illustrates the ingest and conversion of a typical ICPSR dataset:

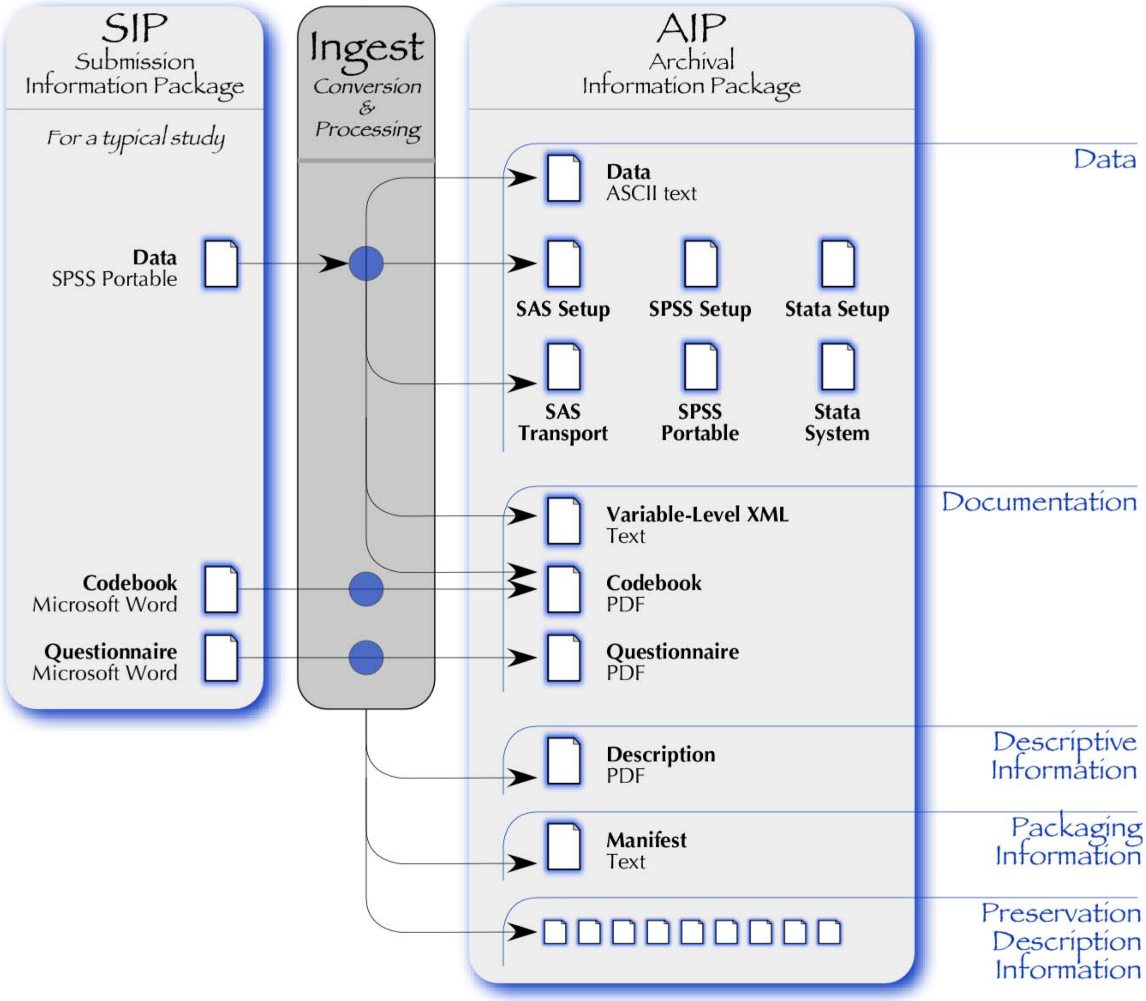

Figure 1: The Ingest Process for a Typical ICPSR Study

Once received, the dataset is run through the "ICPSR Data Pipeline," the process designed for the preparation of access and preservation copies of the data. During processing, ICPSR assesses the dataset, creates new metadata (descriptive, structural, administrative), and enhances existing metadata. Proprietary data formats like SPSS are converted to a more archival, normalized format like ASCII, and "setup files" are generated to allow SPSS and other data compilation programs to read the data and recreate proprietary formats. ICPSR considers the combination of ASCII data plus setup files to be the optimal archival format for long-term preservation. Through another processing step, ICPSR further transforms the processed data into "point-and-click" dissemination formats for three popular statistical analysis packages (SAS, SPSS, and Stata).

During processing, staff also evaluate all data for disclosure risk and alter the data if necessary to protect confidentiality of human subjects. If a public-use version of the data cannot be prepared without compromising the analytic utility of the content, the data may be designated as "restricted use," requiring potential users to sign a legal agreement with ICPSR to gain access.

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 formats, the data are virtually packaged for storage. One copy is sent to the off-site archival system and another is sent to the active access and distribution system. The archives believes that these format conversions, combined with separate archival and distribution formats, preserve the data's significant properties while at the same time, rendering the data in formats preferred by the user community.

Technical Architecture

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

ICPSR maintains 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, and the archival collection and its associated database are accessible only to the Data 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 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 should any one of the components fail. Incremental backups are performed every evening, and a full backup once a week.

In addition to standard Solaris operating system utilities a number of other software systems are considered critical components of ICPSR's services, including:

- An Oracle database engine
- Perl programs (generally ICPSR-produced)
- Apache Web server
- Tomcat servlet container
- A number of access-related software tools such as data analysis packages

ICPSR has a contract with Sun 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 contract 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.

Scale

The ICPSR manages two separate copies of the data archive: one for service use (distribution) and one that solely supports the preservation mission.

As of 23 August 2005, the statistics for the distribution archive were as follows:

Size in TB:	.182 TB (compressed)
Number of files:	490,538
Number of studies:	5,791

As of 24 August 2005, the statistics for the full archive of preserved files were as follows:

Size in TB:	2.2 TB
Number of files:	1,188,557 (± .01%)

Users (Designated Community)

ICPSR users are predominantly researchers from the ICPSR member community (Member list: http://www.icpsr.org/membership/ors.html). Previously, use of the data archives generally involved shipping data files on removable media to the local contact person on the member campus. Over the ten years, ICPSR access has changed dramatically as it moved to a Web services model. In 2001, ICPSR began offering access through a new option, ICPSR Direct, a service offering immediate download of files to authorized users on member campuses. With ICPSR Direct, users at member institutions no longer have to request access to data through their local representatives. As a result, in FY03 users downloaded over twice as much data as the year before: over 7.9 million MB of data.

As data usage by traditional users now empowered with advanced data access increases, ICPSR strategic planning documents also indicate a need to serve a broader user base than simply social scientists. In a recent *Archival Science* article Mary Vardigan, Assistant Director of the ICPSR, described ICPSR users this way:

> "The Designated Community for ICPSR data and other social science archives has traditionally been social science researchers and graduate students who use the data for secondary analysis. Increasingly, however, and particularly with the advent of new dissemination strategies that permit wider access to data, the information held in social science data archives is of interest to other constituencies such as undergraduates, policymakers, practitioners, and journalists, who may not have the expert knowledge base of the traditional constituency. Thus, ICPSR and other social science archives may need to provide more support in the form of assistance to users, tutorials on data use, user guides explaining unique data concepts, online analysis systems, etc., to help users understand the disseminated data."

As implied by the final sentence of that quote, the ICPSR envisions taking on additional responsibilities of mediating between a previously knowledgeable user base and one that will need much more intervention and assistance than that of the past. The 2005 *Data Deposit*

Agreement reflects an invigorated attempt to capture as much descriptive information as possible in order to adequately describe the data collection to potential users.

Appendix 1: ICPSR Council Members, 2004-2006

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