Arch Sci (2007) 7:73–87
DOI 10.1007/s10502-006-9037-z

ORIGINAL PAPER

# ICPSR meets OAIS: applying the OAIS reference model to the social science archive context

**Mary Vardigan · Cole Whiteman**

Published online: 23 August 2007


**Abstract** This paper reviews the archival process at the Inter-university Consortium for Political and Social Research (ICPSR), a repository of digital social science data, and maps ICPSR's Ingest and Access operations to the Open Archival Information System (OAIS) Reference Model. The paper also assesses ICPSR's conformance with the archival responsibilities of "trusted" OAIS repositories, with the proviso that audit criteria for archival certification are still under development. The ICPSR to OAIS mapping exercise has benefits for the larger social science archiving community because it provides an interpretation of the reference model in the quantitative social science environment and points to preservation-related issues that may be salient for other social science archives. Building on the archives' long tradition of shared norms and cooperation, we may ultimately be able to design a federated system of trusted social science repositories that provides access to the global heritage.

**Keywords** Data archives · Trusted digital repositories · Social science research data · Digital preservation · Designated user community

## Introduction

The social science research community was among the first to recognize the benefits of archiving digital data for use by others. Indeed, since the advent of survey research in the 1930s, many data archives, most of them nationally funded, have been established around the world to preserve social science data resources. There are active archives in Central, Eastern, and Western Europe, the Americas, Australia, South Africa, the Middle East, and

M. Vardigan (✉) · C. Whiteman
Inter-university Consortium for Political and Social Research (ICPSR),
University of Michigan, Ann Arbor, MI, USA
e-mail: vardigan@umich.edu

C. Whiteman
e-mail: colew@umich.edu

Asia, with many more under development. In the U.S. the National Archives and Records Administration (NARA) archives records of the federal government, but there is no "official" archive or repository that preserves social science data generated from federally funded projects and other research and scholarship. Rather, a set of archives with ties to major research universities has emerged, including the Roper Center for Public Opinion Research at the University of Connecticut; the Howard W. Odum Institute for Research in Social Science at the University of North Carolina, Chapel Hill; the Henry A. Murray Research Archive and the Harvard-MIT Data Center at the Institute for Quantitative Social Science, Harvard University; and the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan.

This paper investigates how some of the core archival procedures at ICPSR map to the Open Archival Information System (OAIS) Reference Model (Consultative Committee for Space Data Systems 2002). One of the oldest and largest social science data repositories, ICPSR was established in 1962 when digital archiving was in its infancy and data collections were stored and distributed on punched cards. It has evolved to become a major quantitative social science data archive, distributing over 25 terabytes of data through its Web site in fiscal year 2006 and maintaining over 500,000 discrete files in its data repository. While social science archives around the world may differ somewhat in size and services offered, we share a basic mission: to provide continuing access to data for research and instruction. Because of our common goals and our tradition of cooperation, exploring ICPSR's conformance to the OAIS Reference Model has potential benefits for the entire community.

We begin with some basic questions about preservation and what it means in the OAIS context to be a "trusted" digital repository providing "reliable, long-term[1] access to managed digital resources for a designated community, now and in the future" (Research Libraries Group 2002). We then look at OAIS concepts in the ICPSR context, with an emphasis on preservation metadata issues and the types of information about digital assets that are required for an archive like ICPSR to manage its collection effectively across time. Next, we present an overview of the "data pipeline" process at ICPSR, again with reference to OAIS terminology, specifically in the area of Ingest. We examine the responsibilities of an OAIS-compliant repository with a focus on how ICPSR is fulfilling those responsibilities and then finally turn to how ICPSR might participate with other trusted social science digital archives in a distributed archiving model.

## Basic questions and definitions

Recent research on digital archiving has emphasized the importance of clarifying fundamental concepts related to preserving information. Below we highlight three central concepts.

### Preservation

What does it mean to preserve a digital entity? According to Kenneth Thibodeau, "In order to preserve a digital object, we must be able to identify and retrieve all its digital

[1] In the OAIS model, the authors point out that "long term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community" (Consultative Committee for Space Data Systems 2002, p. 1-1). Margaret Hedstrom observes, "'Long term' does not necessarily mean generations or centuries. It may simply mean long enough to be concerned about the obsolescence of technology" (Hedstrom 2002).

components. The digital components of an object are the logical and physical objects that are necessary to reconstitute the conceptual object...The process of digital preservation, then, is inseparable from accessing the object. You cannot prove that you have preserved the object until you have re-created it in some form that is appropriate for human use or for computer system applications" (Thibodeau 2002). Thus, preservation implies access, and clearly the OAIS framework has taken account of this fact. The social science archives noted earlier all operate on this model.

## Significant properties

What do we preserve in a digital archive? It is often not an original physical object but a conceptual or virtual object that contains, replicates, or embodies the original object's "significant properties" or "essential qualities." The National Archive of Australia calls this "essence," that is, "the characteristics that must be preserved for the record to maintain its meaning over time" (Heslop et al. 2002). Defining an object's significant properties is a critical decision that digital archives must make. For example, an archive may decide that the visual presentation of a word-processed document is one of its significant properties and that an exact replica of the document must be preserved. The archive may instead decide that it is only the text or content of the document that is significant and that a conversion to ASCII text is sufficient. Archives have traditionally made these kinds of decisions on behalf of their communities of users, based on input from the intended audience.

Related to this, Thibodeau notes, "To preserve a digital object, is it necessary to preserve its physical and logical components and their interrelationship, without any alteration? The answer, perhaps surprisingly, is no. It is possible to change the way a conceptual object is encoded in one or more logical objects and stored in one or more physical objects without having any negative impact on its preservation" (Thibodeau 2002). While a data migration may change the way a data object is encoded (e.g., EBCDIC to ASCII), the basic intellectual content may not be compromised during such a conversion.

## Trust

What does it mean for a digital archive to be considered "trusted"? Clifford Lynch (2000) observes that "virtually all determination of authenticity or integrity in the digital environment ultimately depends on trust...Trust plays a central role, yet it is elusive." To clarify this issue of trust, an RLG-OCLC Report (Research Libraries Group 2002) laid out the attributes and responsibilities of trusted repositories and set forth a series of recommendations to encourage more rigor in best practices and their codification. Closely aligned with the OAIS Reference Model in content and spirit, the report among other things called for a certification process for digital repositories and more specific definition of the metadata needed for preservation.

Both of those recommendations have spurred additional research, notably the OCLC/RLG project on preservation metadata (OCLC/RLG Working Group on Preservation Metadata 2002) and more recently a Mellon-funded project led by the Center for Research Libraries to develop repository certification metrics. For the latter project, a joint task force made up of individuals from the Research Libraries Group (RLG) and the National

Archives and Records Administration (NARA) will define certification requirements, delineate a process for certification, and identify a certifying body (or bodies) that can implement the process. The project, which ran through October 2006, has released a draft of the audit criteria and has conducted test audits to refine the audit measures with three subject archives: Inter-university Consortium for Political and Social Research (ICPSR); Koninklijke Bibliotheek National Library of the Netherlands, which maintains the digital archive for Elsevier Science Direct Journals; and Portico, an archive for electronic journals incubated within Ithaka Harbors, Inc.

## OAIS concepts in the ICPSR context

The OAIS Reference Model (see Fig. 1) employs a very general nomenclature made up of "terms that are not already overloaded with meaning so as to reduce conveying unintended meanings" (Consultative Committee for Space Data Systems 2002, pp. 1–7). This is useful because many of us rely on jargon unique to our own archival settings, which can make communication across archives difficult. The result of these general terms, however, is that the various archival communities need to translate OAIS terminology to the relevant concepts that apply in their specific contexts. In the next section we map OAIS terms to the social science data archive environment, acknowledging that our interpretations need to be reviewed by other social science archives.

### Information objects and packages

In the OAIS framework, "Information is defined as any type of knowledge that can be exchanged, and this information is always expressed (i.e., represented) by some type of data" (Consultative Committee for Space Data Systems 2002, pp. 1–10). The Content Information, which is the original target of preservation by the OAIS, is the Content Data Object together with the Representation Information that allows for the full interpretation of the data into meaning that can be understood by a Designated Community. In OAIS terms, a Designated Community is "An identified group of potential Consumers who should be able to understand a particular set of information" (Consultative Committee for Space Data Systems 2002, pp. 1–10).

The Content Information and its associated Preservation Description Information (PDI) and Packaging Information make up an Information Package, of which there are three types in the OAIS model: a Submission Information Package supplied by the data depositor, an Archival Information Package suitable for long-term preservation, and a Dissemination Information Package delivered to a user (see Fig. 1).

In the social science archive context, a typical example of a Data Object to be preserved would be a numeric survey data file; its associated technical documentation (sometimes called a "codebook"), which is used to understand and interpret the numeric codes in the data file, would comprise the Representation Information. A data file is ultimately just a string of numbers and not understandable on its own; it can only be interpreted and comprehended intellectually through use of the technical documentation, which indicates a variable's location in the numeric data file, the question it was based on, all possible responses to the question, how the population of interest was sampled (for surveys), and so forth. Together, the data file and its documentation make up the Content Information, sometimes called a data collection or a study.

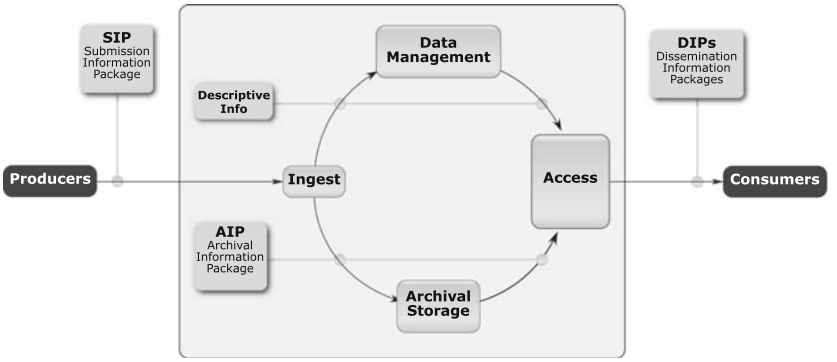


**Fig. 1** Overview of the OAIS Functional Model

## Representation information

In the OAIS model, Representation Information carries both semantic information and structural information. Thus, in addition to the semantic information mentioned above that describes the variables in the data file and what they mean, we also need to understand the encoding structure, or the actual sequence of bits that make up the file type—e.g., ASCII—so that the file can be rendered into the future. In other words, "Structure Information is oriented toward making the Content Data Object understandable to computer systems, while Semantic Information is oriented toward making the Object understandable to humans" (OCLC/RLG Working Group on Preservation Metadata 2002).

It is important to note the layered and recursive nature of Representation Information. In our example above, we mentioned the basic information used to understand the numeric data, but we also need to understand the interpretive information itself. Thus, we might need another document such as the original questionnaire used in administering the survey so that data users can understand the question flow of the interviews and determine how questions relate to variables in the resulting data file. "In principle, the Representation Information even extends to the inclusion of definitions (e.g., dictionary and grammar) of any natural language used in expressing the Content Information" (Consultative Committee for Space Data Systems 2002, pp. 4–24).

## Designated community and its knowledge base

According to the Reference Model, an archive may "make a decision between maintaining the minimum Representation Information needed for its Designated Community, or maintaining a larger amount of Representation Information that may allow understanding by a larger Consumer community with a less specialized Knowledge Base" (Consultative Committee for Space Data Systems 2002, pp. 2–4). This concept of a Designated Community merits further elaboration in the ICPSR and social science context. The Designated Community for ICPSR data and other social science archives has traditionally been social science researchers and graduate students who use these 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.

## Preservation description information

We referred above to the Preservation Description Information, the purpose of which is to ensure that information stored is described well enough that it can be accurately retrieved for use by future generations. The PDI is "specifically focused on describing the past and present states of the content information, ensuring it is uniquely identifiable, and ensuring it has not been unknowingly altered" (Consultative Committee for Space Data Systems 2002, pp. 4–27). Preservation Description Information (PDI) is made up of four types of information, all of which must be present in an Archival Information Package:

1. *Provenance* describes the source of the Content Information. According to the OCLC-RLG (2002) Report, "In addition to recording the 'chronology' of the archived Content Data Object, Provenance Information also can be considered 'event-based' metadata, describing the Object as a dynamic entity." Clifford Lynch, in his article on authenticity and integrity of digital information, called for provenance information to be standardized. His view was that we do not yet "have a clear understanding of (and surely not consensus about) where provenance data should be maintained in the digital environment, or by what agencies. Indeed, it is not clear to what extent the record of provenance exists independently and permanently, as opposed to being assembled when needed from various pools of metadata that may be maintained by various systems in association with the digital objects that they manage. We also lack well-developed metadata element sets and interchange structures for documenting provenance" (Lynch 2000).

   The Reference Model helps to set standards for provenance information, and some metadata schemes have been created for this purpose. One example is the Metadata Encoding and Transmission Standard (METS 2006), which was designed to encode descriptive, administrative, and structural metadata in XML. Indeed, METS emerged as a best practice in the PREMIS report on *Implementing preservation repositories for digital materials*, with 54% of the repositories polled using the METS scheme (PREMIS Working Group 2004).

   The Data Documentation Initiative (DDI) is another XML metadata model specifically designed for social science metadata. While it has traditionally focused on what the OAIS would consider Descriptive Information and Representation Information, the specification is evolving to document the broader life cycle of a data collection. Version 3.0, to be published in 2007, will make explicit the relationship between the DDI and other metadata standards like METS. Many of the social science data archives including ICPSR currently use the DDI, and some are now exploring METS as well.

   In the ICPSR context, some of the provenance information is described in the associated metadata record (Descriptive Information); this includes bibliographic information and collection changes and history. Additional information about data

deposit and data migrations is maintained in the archival storage database. Processing history—i.e., the various procedures that data undergo during the data processing phase—is also captured and preserved along with the dataset.

2. *Context* describes how the content relates to other information outside the information package, including why it was produced. "It is important to note that Context Information is directed at informational requirements associated with managing the preservation process, not those aimed at facilitating understanding and interpretation of the Content Data Object's intellectual content. The latter is addressed by metadata elements within the Object's Representation Information" (OCLC/RLG Working Group on Preservation Metadata 2002). The Context type of Preservation Description Information was the most difficult for us to understand and to interpret in the ICPSR environment, and it would be useful to have more examples of this concept. Our sense is that context information may refer to the relationships among the different formats for a given data file and the original documents submitted. In the ICPSR context, much of this information is embedded in the archival storage database structure, that is, all of the successive iterations of a data object are archived together under a reference number. Processing history also encompasses the changes that are made to data and the rationale for creating new versions (e.g., recodes for confidentiality reasons, optimization of file size, etc.).
3. *Reference* provides one or more identifiers, or systems of identifiers, by which the content information may be uniquely identified. At ICPSR, the unique identifier at the highest level is the five-digit ICPSR study number. Other study-level identifying information is provided in the bibliographic metadata record and the citation. Studies consist of numbered datasets, which in turn consist of files. Each file name includes the study number, the dataset number, the ICPSR file type code that categorizes the file according to type of content, and a Windows-style suffix. ICPSR is exploring a mechanism for assigning more formal unique identifiers and presenting them in the form of a machine-readable data citation, following the model proposed by the Harvard-MIT Virtual Data Center (Altman and King 2006). This would include a unique global identifier of some type, a Persistent Uniform Resource Locator (PURL),[2] and a Universal Numeric Fingerprint (UNF)[3] for each ICPSR dataset in the holdings.
4. *Fixity* provides a wrapper or protective shield that protects the content information from undocumented alteration. ICPSR has traditionally provided its data users with file measurements such as byte count, record count, and record length with which to gauge that the requested file transferred accurately. In addition, a system to automatically calculate and associate checksums for each data file has recently been deployed. ICPSR is currently researching the best way to provide more detailed fixity information, possibly through the UNF mechanism indicated above.

Our assessment of the Preservation Description Information that ICPSR maintains is that it needs to be more robust given its centrality to the preservation enterprise. We plan to consult with experts in the digital preservation field about which fields we should add to our existing set of preservation metadata to ensure optimal coverage.

[2] Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client (OCLC Research, nd).

[3] A universal numeric fingerprint is used to guarantee that a defined subset of data is substantively identical to a comparison subset (The UNF package 2006).

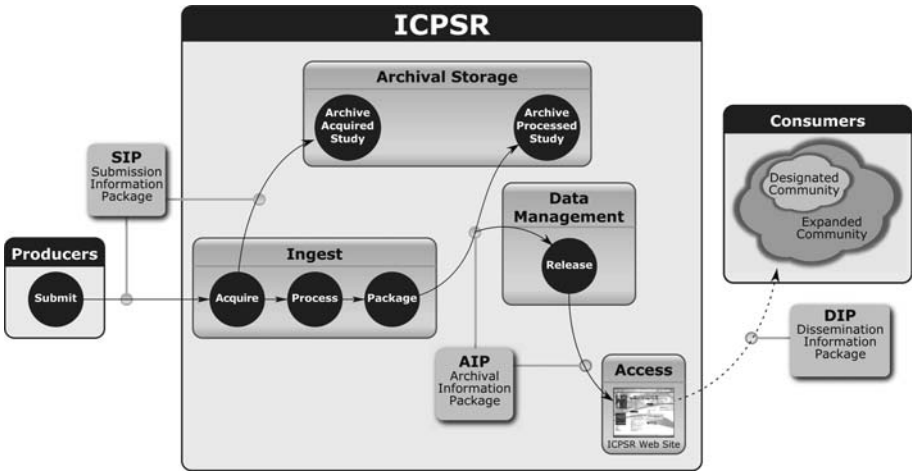


**Fig. 2** The ICPSR process from submission to access

## ICPSR information flow

We now turn to an overview of ICPSR's "data pipeline"[4] and how parts of it map to the OAIS framework (see Fig. 2).

### Submission

Data submissions at ICPSR are initiated in various ways. As noted in the Reference Model, the data deposit may be voluntary and unsolicited, i.e., a researcher understands the importance of long-term preservation of digital data and decides to deposit his or her data for future generations of scholars to use. In other cases, data are submitted as a requirement of a grant or sponsoring agency agreement with ICPSR. Increasingly, however, ICPSR takes a proactive approach in identifying data of interest and negotiating for their deposit into the ICPSR Archive.

The submission process begins when a data submitter prepares a Submission Information Package. (ICPSR makes available a *Guide to Social Science Data Preparation and Archiving (ICPSR 2005a)* to assist the data submitter in preparing and depositing data into an archive like ICPSR.) The SIP should consist of data and related technical documentation, i.e., the Content Information. For example, the SIP might contain a data file in the

[4] The OAIS mapping activity was possible because of a year-long effort that ICPSR undertook in 2003 to assess its "data pipeline" process—that is, the path of data flow from data ingest, through data processing, to data preservation and dissemination. ICPSR sought to identify ways to streamline the pipeline process, to incorporate standards, and to improve study tracking. To realize those objectives, a process improvement committee at ICPSR charted the existing process and designed a new, "ideal" pipeline process based on current technology. The committee also looked to the outside world and to external initiatives, like the OAIS, for guidance on what was happening in the broader archival community in terms of best practice and emerging standards in the field. An External Review Committee later endorsed the process recommendations.

proprietary format of a commercial statistical package such as SPSS[5] plus related technical documentation consisting of a codebook and questionnaire deposited in Microsoft Word format (Microsoft Office Online 2006).

The submitter completes an electronic Deposit Form, which captures some types of Preservation Description Information and other metadata describing the SIP and provides for secure upload of the digital materials. By signing off on the Deposit Form, the submitter asserts that he or she holds copyright and is giving ICPSR authority to redistribute and preserve the data. Because social science data may contain identifying information about human subjects, ICPSR also requires the depositor to certify that the data have been anonymized and that any individual identifiers have been removed.

Note that a complete deposit may require several Submission Information Packages, as when a depositor has not sent a fully documented dataset or has not otherwise complied fully with deposit standards.

## Ingest

Once received, the Submission Information Package resides in the Ingest space, which is separate from ICPSR's archived data for security purposes. For unsolicited data submissions, ICPSR assesses their relevance to ICPSR to ensure that they accord with ICPSR's Collection Development Policy. Here, ICPSR assigns a study number and other bibliographic reference information to the data collection. ICPSR also evaluates the SIP and determines whether confidentiality issues exist. Even though a data producer may have certified that data contain no personal identifiers, ICPSR considers it a key responsibility to conduct an independent disclosure analysis. ICPSR also determines the level of data processing that the collection will undergo to make it usable by the community.

Next, ICPSR staff process and enhance the submitted data collection to make it easier to use. This may involve file format conversion, Representation Information conversions or enhancement, and reorganization of the Content Information in the SIPs. Data processing is a major part of the ICPSR pipeline, but the OAIS model discusses such value-added activities only briefly, acknowledging that "The complexity of this Ingest process can vary greatly from OAIS to OAIS" (Consultative Committee for Space Data Systems 2002, pp. 4–50).

Most of the social science data archives routinely perform value-added processing, which can be quite resource-intensive. ICPSR, with the University of Michigan's School of Information, is currently taking part in a research project sponsored by the National Science Foundation and the Library of Congress. This project is examining incentives for data depositors to submit "archive-ready" datasets to repositories like ICPSR. ICPSR has a set of guidelines (*Requirements for Rapid Release of Data*) which, when followed, ensure that data processing is expedited so that data can be released to the Web quickly (ICPSR 2006). Currently, few depositors comply fully with these guidelines.

During processing, the submitted data may be reviewed for internal consistency issues such as undocumented codes. The data processor may contact the depositor if the data are to be significantly revised, especially if confidential data are to be recoded or purged. ICPSR and the data producer may decide that both a public-use and a restricted-use version of the data will be released. (At ICPSR special policies and procedures exist for access to restricted-use data.)

[5] SPSS, or Statistical Package for the Social Sciences, was developed in 1968 and was the first statistical analysis product to be used extensively by social scientists. It continues to be used widely today, although there are now several statistical analysis packages, including SAS and Stata.

At this point in the process, ICPSR transforms a proprietary data format like SPSS into a more appropriate software-independent archival format, specifically raw ASCII text data with SPSS "setup" files that enable a user to read in the raw data to recreate the proprietary SPSS format. ICPSR considers the combination of raw data plus setup files to be the optimal archival format for long-term preservation because this package has the best chance of being readable into the future. Through automation, ICPSR further transforms the processed data into "point-and-click" dissemination formats for three popular statistical analysis packages (SAS, SPSS, and Stata). ICPSR also creates DDI-compliant variable-level XML files, which are also ASCII text at their core. XML-tagged documents contain content-specific markup (e.g., <title>American National Election Study, 2000 </title>), rendering them more useful and more robust for purposes of preservation, fielded searches, and import into existing systems such as online analysis and subsetting. Note that the format conversions preserve the data object's significant properties while at the same time rendering the data easier to use by the Designated Community.

During the Ingest phase, the Representation Information is also transformed—Word files are generally converted to PDF (moving toward the PDF/A ISO standard, intended for long-term preservation). A metadata record—i.e., Descriptive Information in OAIS terminology—is prepared during this phase as well. The Ingest phase concludes with the assembly of the AIP (see Fig. 3). Both the original data (that is, the Submission Information Package as received from the submitter) and a processed version are maintained for

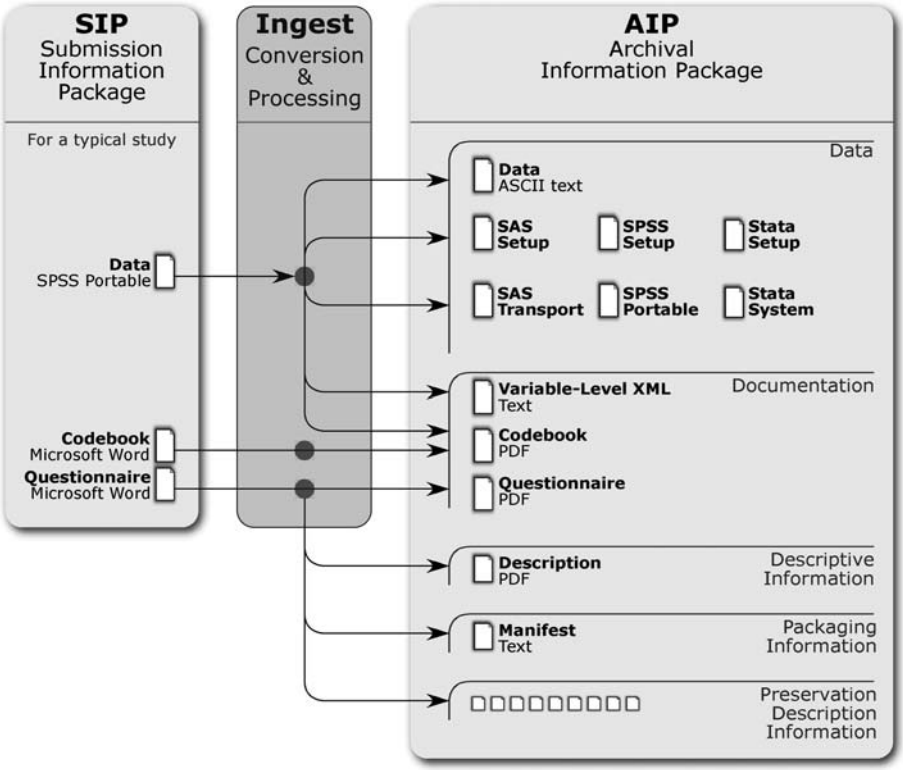


**Fig. 3** The ingest process for a typical ICPSR study

the long term, and quality assurance happens at every step in the process. ICPSR has built several quality control mechanisms into its data pipeline, including both automated checks and checks by humans.

### Data release and archival storage

Once the study AIP is complete, a copy is moved to the ICPSR Web server, and ICPSR Data Management announces to its Designated Community that the study is available on the ICPSR Web site. Additional archival copies are made and sent off-site on removable media, currently Digital Linear Tapes (DLTs). The AIP on the Web server will form the basis of one or several Dissemination Information Packages, or DIPs, requested by users from the ICPSR Web site. The DIP, which is assembled on the fly and transferred to the user as a Windows-style zipped file, has associated Packaging Information in the form of a file manifest that lists the files comprising the download.

### Access

In the case of public-use data, accessing data from ICPSR is fairly straightforward. Generally, as the OAIS specifies, the Consumer will use the OAIS Finding Aids (ICPSR search engine) that operate on Descriptive Information (ICPSR catalog of metadata records) to identify a data collection of interest. For ICPSR, institutional membership confers access privileges, so the typical "orders" for data are what the OAIS would consider Adhoc Orders and involve data downloads from the ICPSR Web site. Members are granted access based on their IP addresses, which have been captured in a database, and there is also an authentication process called MyData, which requires a user ID and password. Some ICPSR data are freely available and may be downloaded without IP verification and authentication. ICPSR also provides support to Consumers in using the data.

### OAIS entities and the pipeline

Providing a detailed mapping of the ICPSR data pipeline to all six functional entities of an OAIS—Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access—and the tasks involved in each is beyond the scope of this paper. However, we find that in the ICPSR model some of these functions and tasks span several departments. For example, Preservation Planning in the OAIS context contains a mix of tasks that at ICPSR would be performed by the Preservation Group along with the Computer and Network Services Group. These groups work closely together to manage large-scale data migrations, such as the migration of the mid-1990s when data were migrated from a central mainframe to local disk for storage and dissemination, and data distribution via magnetic tapes gave way to file transfers over the Internet.

## Responsibilities of an OAIS-compliant repository

A critical component of the OAIS Reference Model is the set of responsibilities that an OAIS-conformant archive must fulfill consistently in order to be considered trustworthy.

Here we describe how ICPSR fulfills these responsibilities currently, and we suggest future improvements.

*Negotiate for and accept appropriate information from Information Producers*. In the ICPSR environment, data depositors complete and sign a Data Deposit Form, which elicits information about the salient intellectual and physical characteristics of the collection being deposited. The form also elicits assurance that confidential information has been removed and that the proper approvals have been sought and granted.

*Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation*. The Data Deposit Form asks the depositor to attest to the fact that he or she has copyright to the data collection and thus has the authority to grant approval for ICPSR to redistribute the data. The form also obtains permission for ICPSR to migrate or transform content for preservation purposes.

*Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided*. As indicated above, the user community is evolving to include more novice users, and ICPSR is undertaking a new initiative to understand this broad community and the types of products it might need. ICPSR does have access to the "other parties" indicated above to assess its current and potential user communities. First, ICPSR has a network of local representatives on member campuses who provide information and feedback on their experiences with users. Second, ICPSR's governing Council also provides guidance, and would be consulted if a dramatic change in the Designated Community were contemplated. And finally, ICPSR is allied with other social science data archives around the world. The result of all of these networks is that ICPSR receives abundant feedback and data to monitor its Designated Community and the changing needs over time.

*Ensure that the information to be preserved is Independently Understandable to the Designated Community*. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information. Data processors at ICPSR evaluate each data collection to determine what kinds of explanatory or descriptive information need to be provided in order to use the data most effectively. ICPSR also provides technical support to users of the data, and the local campus representatives perform this service as well. The experts who produced the information—in this case principal investigators, other researchers, and government agencies—are not considered contacts unless they request to play that role.

*Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original*. While ICPSR has prepared a Technology Usage Plan for disaster recovery and has a strategy for preservation, it has not yet articulated that strategy in a comprehensive way for the Designated Community. It is clear that codifying and publishing our preservation policies and practice must be high-level priorities as we move forward in this area.

In terms of verifying copies, ICPSR provides detailed file-level specifications that enable users to determine that they have received the proper version of the file, and is also investigating best practice in the areas of provenance, fixity, and formal unique identifiers. ICPSR's processing history details the chain of custody of a file in its successive versions.

*Make the preserved information available to the Designated Community*. ICPSR provides the Designated Community with access to its data holdings via its public Web site (ICPSR 2005b), which receives over one million Web site hits each month. Restricted data are made available through separate means with legal contracts that stipulate that these data must be certified as destroyed at the end of the contract period so that ICPSR has

verification that these sensitive data are no longer available. Other data with even tighter access constraints may only be analysed on-site at ICPSR in the Secure Data Enclave. The goal is to provide access, even when there are stringent constraints on data use.

### Areas for improvement

Our initial assessment reveals that ICPSR is fulfilling many of the key responsibilities of an OAIS-modelled archive but that two challenges stand out: the need for a published preservation policy, and the fact that Preservation Description Information is incomplete and not always clearly labelled in the ICPSR system. ICPSR hopes to make significant progress on these challenges in the next fiscal year and to continue assessing conformance with the model across other OAIS functional entities not discussed in this paper.

## Interoperability among archives

As mentioned earlier, the social science data archives have formed a variety of partnerships and federated relationships. Early on in their history, they established an arrangement whereby more than one social science data archive stored copies of some of the same datasets for convenience to users and to guard against loss. As technology evolves and a more stable and trusted distributed architecture comes into being, this type of redundancy may be accomplished in different ways.

Increasingly, the social science archives are moving in the direction of distributed architectures. For example, member archives of the Council for European Social Science Data Archives (CESSDA) have developed a common data portal as part of the MADIERA Project[6] funded by the European Union (MADIERA 2005). This portal enables a multilingual search across the holdings of the partners and also provides data for online analysis. Thus, CESSDA qualifies as a Federated Archive structure in OAIS terms: a group of "archives with both a Local Community (i.e., the original Designated Community served by the archive) and a Global community (i.e., an extended Designated Community) which has interests in the holdings of several OAIS archives and has influenced those archives to provide access to their holdings via one or more common finding aids" (Consultative Committee for Space Data Systems 2002, p. 6-2).

In the United States, the Library of Congress recently recognized the importance of preserving the heritage of the social sciences by making an award to a group of data archives through its National Digital Information Infrastructure and Preservation Program (NDIIPP, nd). The NDIIPP has funded a partnership among ICPSR, the Roper Center, the Odum Institute, the Murray Center, the Harvard-MIT Data Center, and the Electronic and Special Media Records Service Division of the National Archives and Records Administration (NARA). In the aggregate, the institutions in the partnership (not counting NARA's holdings) hold over a million files.

These institutions have come together in a collaboration called the Digital Preservation Alliance for the Social Sciences (Data-PASS, nd), designed to facilitate the acquisition and

[6] The MADIERA project (Multilingual Access to Data Infrastructures of the European Research Area) was designed to develop an effective infrastructure for the European social science community by providing a common integrated interface to the resources of the majority of the existing 20+ social science data archives in Europe including several newly established archives in the candidate countries.

preservation of classic at-risk studies in the social sciences. The partnership may be viewed as a federation in OAIS terms since it provides for a shared catalog and has an extended Designated Community. The project establishes a clear division of responsibilities in terms of which institution will preserve which type of information. Thus, under this agreement the Odum Institute will acquire and archive private social science research, Harris polls, and state polls while ICPSR will acquire and archive surveys of political life and university-based research, and so forth. The implication is that each of the institutions is trusted to maintain its designated AIPs in perpetuity for the benefit of the larger social science research community.

Institutional repositories at universities and other research institutes can also play an important role in preserving data that might otherwise be lost. The challenge for the community is to create a federation mechanism to facilitate searches across these repositories.

## Conclusion

ICPSR's self-assessment of OAIS compliance with respect to a subset of the model's components has helped us to identify areas for improvement and to set a clear agenda for preservation-related work. We note that the UK Data Archive and the National Archives in Great Britain recently conducted a more formal assessment and have released a comprehensive report (Beedham et al. 2005) evaluating their level of compliance with the OAIS standards. We urge other archives to apply the framework to their own settings so that a fruitful dialog on the Reference Model can begin.

As we look to the future, we see the potential for even greater cooperation among the world's social science data archives, with shared authentication protocols that enable users to access the accumulated data holdings of the archives. Conforming to the OAIS standard will permit the archives to communicate more effectively and to provide access to a network of trusted digital repositories that together will preserve the legacy of the social sciences.

## References

Altman M, King G (2006) A proposed standard for the scholarly citation of quantitative data, http://gking.harvard.edu/files/cite.pdf (Consulted 05 Nov 2006)

Beedham H, Missen J, Palmer M, Ruusalepp R (2005) Assessment of UKDA and TNA compliance with OAIS and METS standards. Joint Information Systems Committee (JISC), United Kingdom, http://www.jisc.ac.uk/uploaded_documents/oaismets.pdf (Consulted 05 Nov 2006)

Consultative Committee for Space Data Systems (2002) Reference model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1 Blue Book, January 2002, http://public.ccsds.org/publications/archive/650x0b1.pdf (Consulted 05 Nov 2006)

Data-PASS (nd) Data preservation alliance for the social sciences, http://www.icpsr.umich.edu/DATAPASS/ (Consulted 05 Nov 2006)

DDI (2005) Data documentation initiative, http://www.icpsr.umich.edu/DDI (Cited 05 Nov 2006)

Hedstrom M (2002) The digital preservation research agenda. In: The state of digital preservation: an international perspective – conference proceedings, July 2002, http://www.clir.org/pubs/reports/pub107/hedstrom.html (Consulted 05 Nov 2006)

Heslop H, Davis S, Wilson A (2002) An approach to the preservation of digital records. National Archives of Australia, Canberra, http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf (Consulted 05 Nov 2006)

ICPSR (2005a) Guide to social science data preparation and archiving: Best practice throughout the data life cycle, 3rd edn. ICPSR, Ann Arbor, MI, http://www.icpsr.umich.edu/access/dataprep.pdf (Consulted 05 Nov 2006)

ICPSR (2005b) Inter-university consortium for political and social research, http://www.icpsr.umich.edu (Consulted 05 Nov 2006)
ICPSR (2006) ICPSR requirements for rapid data release, http://www.icpsr.umich.edu/access/deposit/dissemination-ready.html (Cited 05 Nov 2006)
ISO (2005) International organization for standardization. Document management – electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1), http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=38920&scopelist=PROGRAMME (Consulted 08 Nov 2006)
Lynch C (2000) Authenticity and integrity in the digital environment: an exploratory analysis of the central role of trust, http://www.clir.org/pubs/reports/pub92/lynch.html (Consulted 05 Nov 2006)
MADIERA (2005) The MADIERA project, http://www.madiera.net/ (Consulted 05 Nov 2006)
METS (2006) Metadata encoding and transmission standard, http://www.loc.gov/standards/mets/ (Consulted 05 Nov 2006)
Microsoft Office Online (2006) Microsoft Office Word 2007, http://www.office.microsoft.com/en-us/word/FX100487981033.aspx (Consulted 05 Nov 2006)
NDIIPP (nd) National digital information infrastructure and preservation program. Library of Congress, http://www.digitalpreservation.gov/ (Consulted 05 Nov 2006)
OCLC Research (nd) PURLS, http://www.purl.oclc.org/ (Consulted 05 Nov 2006)
OCLC/RLG Working Group on Preservation Metadata (2002) Preservation metadata and the OAIS Information Model. A metadata framework to support the preservation of digital objects, http://www.oclc.org/research/projects/pmwg/pm_framework.pdf (Consulted 05 Nov 2006)
PREMIS Working Group (2004) Implementing preservation repositories for digital materials: current practice and emerging trends in the cultural heritage community, http://www.oclc.org/research/projects/pmwg/surveyreport.pdf (Consulted 05 Nov 2006)
Research Libraries Group (2002) Trusted digital repositories: attributes and responsibilities – an RLG-OCLC report, http://www.rlg.org/longterm/repositories.pdf (Consulted 05 Nov 2006)
Research Libraries Group and National Archives and Records Administration (August 2005) An audit checklist for the certification of trusted digital repositories: draft for public comment (August 2005) Research Libraries Group, Mountain View, CA, http://www.rlg.org/en/pdfs/rlgnara-repositorieschecklist.pdf (Consulted 05 Nov 2006)
SAS http://www.sas.com/ (Consulted 05 Nov 2006)
SPSS http://www.spss.com/ (Consulted 05 Nov 2006)
Stata http://www.stata.com/ (Consulted 05 Nov 2006)
Thibodeau K (2002) Overview of technological approaches to digital preservation and challenges in coming years. In: The state of digital preservation: an international perspective – conference proceedings, July 2002, http://www.clir.org/pubs/reports/pub107/thibodeau.html (Consulted 05 Nov 2006)
The UNF package (2006) http://www.cran.r-project.org/doc/packages/UNF.pdf (Consulted 05 Nov 2006)