

Ingest Process: Submission and ‘Pre-Ingest’ Activities

Jared Lyle

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

Overview

Archives are based on trust. Records deposited into an archive have long term value and are expected to live on for decades, if not forever. Depositors trust that an archive will accept responsibility for and safeguard their digital objects. Users trust that the objects they access at the archive are accurate and true to their original form. Trust is not built on self-selection or self-aggrandizement; rather, it is based on transparent adherence to and certification against community standards.

Archival trust is initiated with the ingest process, which serves as the gatekeeper of all other archival functions. Existing standards define the ingest process, which encompasses acquiring content and then creating an archival package that is the basis for preservation and access.ⁱ This chapter will focus on the portion of ingest dealing with acquisition of content -- also referred to as “submission and ‘pre-ingest’ activities”.ⁱⁱ These activities include:ⁱⁱⁱ

- Checking for viruses and validating the integrity of the digital object.
- Assigning objects unique identifiers.
- Ensuring that everything expected upon submission has been received.
- Ensuring that all necessary metadata for long-term maintenance and continuing access accompanies the object.
- Assessing the significant properties of the digital object, such as its look and feel, or functionality.
- Selecting content based on a collection development policy.

Specifically, this chapter will detail how submission and ‘pre-ingest’ activities are implemented at the Inter-university Consortium for Political and Social Research (ICPSR), a data repository of social and behavioral science research. While some aspects of the ICPSR repository system are specific to data-intensive scientific workflows, the overall design and implementation are still applicable to any repository looking to safeguard and provide access to digital materials in a trustworthy manner.

ICPSR, which is based at the University of Michigan, has been archiving social and behavioral science research data for over 50 years. It is in the business of providing long-term access to content. Media and formats have changed over time, as have staff and infrastructure; in the early years, for instance, data were preserved on punched cards and 9-track tapes, while files are now managed across a replicated preservation system on servers and cloud storage.

ICPSR needs a trustworthy repository system to insure long-term access to its valuable research data. Why? There are several key reasons. First, science is based on data, to both validate past research and generate new ideas.^{iv} Without trustworthily archived data, accurate replication and validation would not be possible. Second, ICPSR provides data to specific communities of practice -- e.g., political scientists, economists, criminologists. These communities look to professional repositories as trustworthy sources of information.^v With the explosion of the Web, finding data is easy; finding data from trustworthy and reliable sources, however, is not as easy or straightforward. Third, governments are increasingly requiring data from funded research to be preserved in trustworthy repositories rather than through any of the countless data storage options available.^{vi} Fourth, repositories themselves are looking for

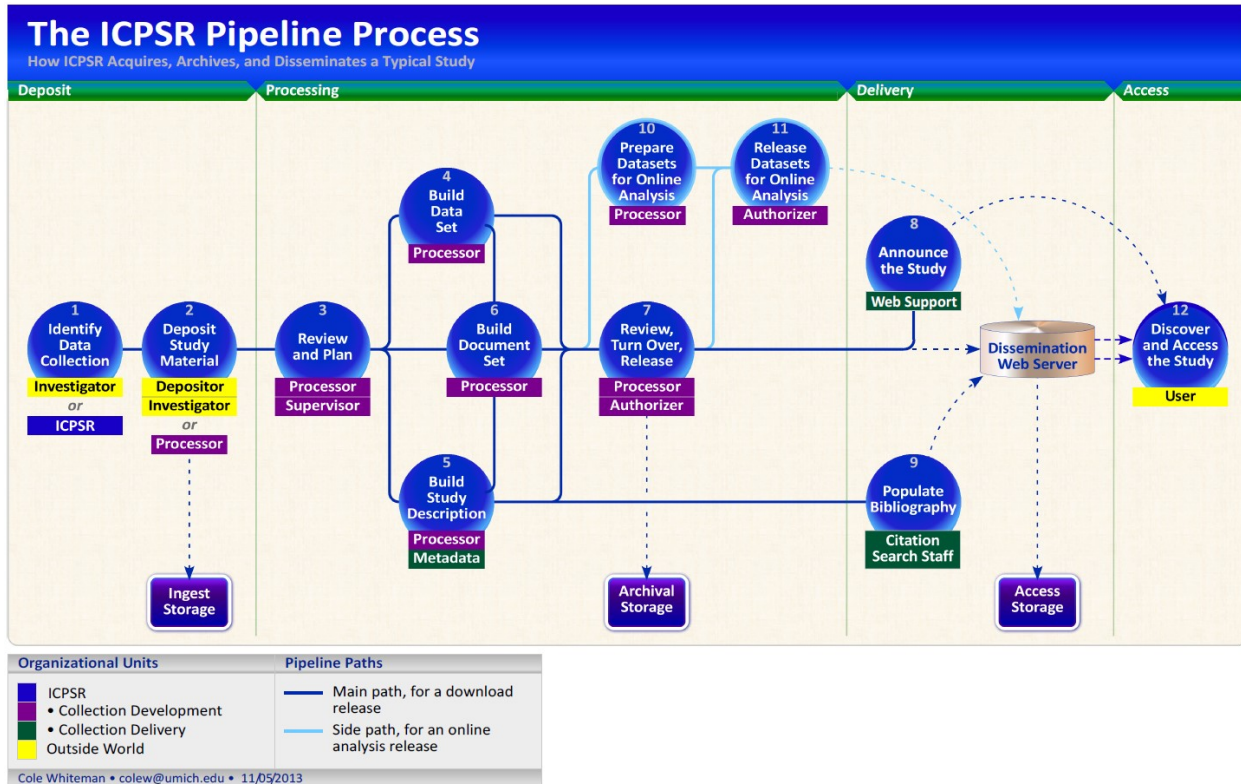
operating standards for trustworthiness against which they may self-assess management, operations, and technologies.^{vii}^{viii}

While ICPSR staff have provided thoughtful data stewardship over a half century, steps taken within the past decade, in particular, have proven crucial towards solidifying the repository practices and procedures. In 2006, ICPSR participated in a formal external test audit of the Trusted Repositories Audit & Certification (TRAC) criteria and checklist.^{ix} “While some issues requiring resolution were identified in the audit,” the final report noted, “when taken as a whole ICPSR appears to provide good stewardship of the valuable research resources in its custody...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.” ICPSR made internal corrections to improve those issues identified by the external test audit. Additionally, in an attempt to increase transparency, repository practices and procedures, unless confidential, were posted on the public Web site. More recently, ICPSR has opened itself to further external audits -- the Data Seal of Approval^x and the ICSU World Data System^{xi}. Like previous analyses, these audits allowed the organization to further improve procedures and policies. Becoming a trustworthy repository is an ongoing process of refinement and revision.

Submission and ‘Pre-Ingest’ Workflow at ICPSR

What follows are implementation details of the submission and ‘pre-ingest’ workflow at ICPSR. Figure 1 provides context to where submission and ‘pre-ingest’ activities (found within the section labeled ‘Deposit’) fit into the overall ICPSR workflow.

Figure 1. The ICPSR Processing Pipeline



Submission through a Deposit Form

All electronic content is submitted to ICPSR via an online deposit form.^{xii} The form serves multiple functions. It enables the depositor to: transfer the content, describe the content, and provide legal permission for ICPSR to reformat, archive, preserve, and disseminate deposited materials.

For the upload process, a user simply uploads files via a Web browser. Preferred file types and formats are suggested within the upload section.

For the description (otherwise known as metadata), basic information is asked: title, principal investigator(s), and description or abstract (see Figure 2). Additionally, the depositor

may provide more detailed metadata, including methodological details such as response rates, sampling selection, and mode of data collection.

Figure 2. ICPSR Deposit Form -- Describe the Collection^{xiii}

Describe the Collection

This section collects basic details about the data collection. On the next page of the deposit form, you will be able to describe this data collection in more detail.

Title of the Data Collection
Descriptive titles typically include the time period(s) and geographic location(s) that the data cover.

Principal Investigator(s)
Please list the Principal Investigators in order of importance to the study. If the Principal Investigator is an organization, complete just the affiliation field.

PI First Name	PI Last Name	PI Affiliation
<input type="text"/>	<input type="text"/>	<input type="text"/>

ADD ANOTHER ROW

Description or Abstract
Please give the user a clear sense of what the study is about. The focus should be on questions such as the purpose of the study, the major topics covered, and what questions the PIs attempted to answer when they conducted the study. Note that you should avoid attempting to address issues of how the data might be used, who might be interested in the data, or any evaluative comments about the worth or usefulness of the study.

Is this deposit a new edition, extract, update, or special version of an existing ICPSR data collection?

Yes
 No

The legal deposit agreement (see Table 1) addresses intellectual property, confidentiality, and permissions to reformat, archive, preserve, and disseminate deposited content. While depositors retain ownership of their data, it is important that they give permission so ICPSR can transform the files for long-term preservation and access.

Table 1. ICPSR Deposit Form -- Deposit Agreement^{xiv}

1. I have implicit or explicit copyright to this work and have the right to make it publicly available through ICPSR.
2. I give my permission for the Data Collection to be used by ICPSR for the following purposes, without limitation:
 - o To disseminate copies of the Data Collection in a variety of media formats
 - o To promote and advertise the Data Collection in any publicity (in any form) for ICPSR
 - o To describe, catalog, validate and document the Data Collection
 - o To store, translate, copy or re-format the Data Collection in any way to ensure its future preservation and accessibility
 - o To incorporate metadata or documentation in the Data Collection into public access catalogues
3. I give my permission to ICPSR to enhance, transform and/or rearrange to the Data Collection, including the data and metadata, for any of the following purposes:
 - o Protect respondent confidentiality
 - o Improve usability
4. To the extent allowable by law or permitted by the sponsor of the data collection, in preparing this data collection for public archiving and distribution, I have removed all information directly identifying the research subjects in these data, and I have used due diligence in preventing information in the collection from being used to disclose the identity of research subjects.
5. I further agree to release and hold harmless ICPSR (including staff and the ICPSR Council) and the University of Michigan from any and all liability from claims arising out of any legal action concerning identification of research subjects, breaches of confidentiality, or invasions of privacy by or on behalf of said subjects.

To formally complete the deposit, the depositor electronically signs the document. If the depositor does not have permission to sign off on the deposit, she may complete the deposit and designate another person as the final signatory, who then receives a separate e-mail request to sign off on the deposit.

Behind the scenes, ICPSR's system runs virus checks, calculates checksums, identifies file formats, and records the technical metadata for all uploaded files. These are important steps to insure the integrity of deposited content; the technical details captured upon ingest can be compared against future states of the materials to insure long-term maintenance and continuing access. The deposited data are also transferred to secure storage.

After the deposit is submitted, the depositor receives two email notifications. The first is immediately after submission and simply confirms that the files were received. The second (see Figure 3) is sent that evening, and inventories the deposited content, including file name, format, and checksum.

Figure 3. Deposit Inventory Email Notification

From: <deposit@icpsr.umich.edu>
Subject: Deposit Inventory - John Doe

This message is automatically generated in response to files you (or your group) uploaded to ICPSR through the online deposit form. When ICPSR receives files through the data deposit form, we automatically generate a listing of the files received. This list also includes a description of the format of the files uploaded as detected by our software. And for data deposited in SAS, SPSS, or Stata formats, we provide a count of cases and variables in each data file.

Please review the following information carefully to be sure that it corresponds to what you intended to submit to ICPSR. Please contact us with any discrepancies or questions that you have (deposit@icpsr.umich.edu).

Thank you,
ICPSR Acquisitions Staff

Deposit Title: Test Data
Deposit Number: 12345
[Link: http://www.icpsr.umich.edu/cgi-bin/ddf2?key=VE82mvr1ztthefakeurl&page=suppl](http://www.icpsr.umich.edu/cgi-bin/ddf2?key=VE82mvr1ztthefakeurl&page=suppl)
Listing of Deposited Files

ingest.txt was scanned on 01-OCT-15.
We think it's: ASCII text, with CRLF line terminators

Review of the Submission

Immediately after deposit, ICPSR staff receive an e-mail notification. This signals staff to review the deposit using a Web-based 'deposit viewer' (see Figure 4).

Figure 4. ICPSR Deposit Viewer

Deposit 36437 Search for another Deposit Search

History • Metadata • Files • Physical Items • Emails

Deposit ID	Title and PI	Depositor	Source	Status	Owner	Assigned to	Created	Modified	Signed	Files Uploaded	Linked to Study Metadata Project(s)
d36437	Test John Doe (Test College)	Authorized to sign: • Jared Lyle lyle@umich.edu Signer [Not required]: • Expected: jl • Actual: Jared Lyle	ddf2 	Signed	TEST	Michael Shove 	2015-11-06	2015-11-06	2015-11-06	2015-11-06: 1	

History of Deposit 36437 [add a diary entry](#)

Timestamp	Event	User	Comment
2015-11-06 01:16:10	Status: Signed	Jared Lyle	edit
2015-11-06 01:16:10	Status: Started	Jared Lyle	edit
2015-11-06 01:16:10	Assigned to: Michael Shove	Jared Lyle	

Metadata for Deposit 36437 [show empty metadata fields](#)

Section	Fieldname	Field	Value
1. Deposit Arrival	source	How the deposit entered the system	ddf2
	create	Date user began deposit entry	06-NOV-15
	chdate	Date of last change to deposit entry	06-NOV-15

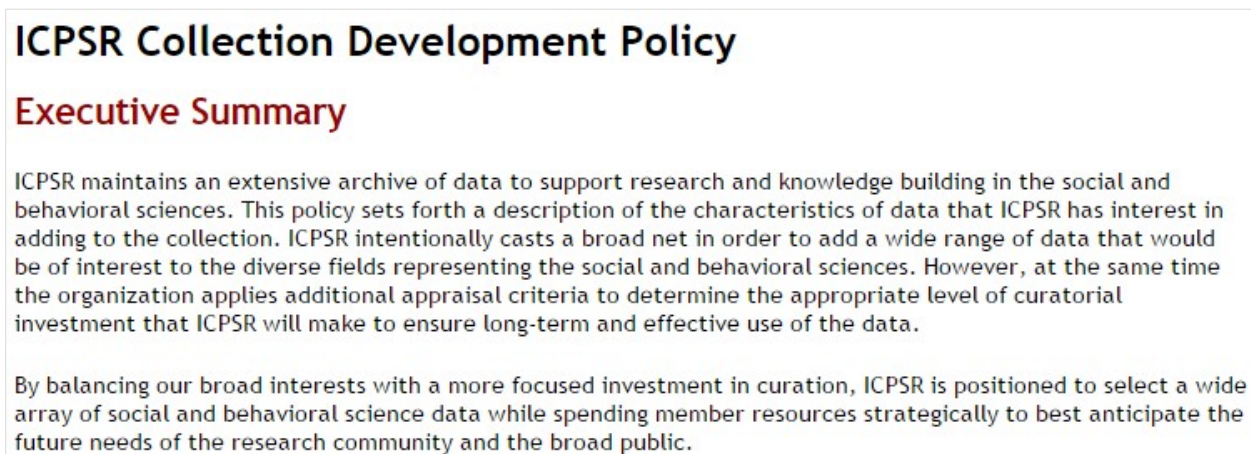
The deposit viewer contains the same metadata that was submitted by the depositor, although augmented by additional technical metadata and notes internal to ICPSR, including the unique ID automatically assigned to each deposited object and a record of all communication with the depositor.

Using the metadata and transferred files tracked in the deposit viewer, staff review the submission for completeness -- i.e., they insure that everything expected upon submission has been received. This includes manual review of files and metadata to check the completeness and functionality. Depositors occasionally upload unintended files, including family photos, draft documentation files, and superseded data. Similarly, depositors sometimes upload only partial documentation -- e.g., submitting all codebooks but forgetting the questionnaires and user

guides. We're only human. Review of content allows staff to negotiate with depositors to update and correct any unintended errors or omissions.

Staff also evaluate the collection against the ICPSR Collection Development Policy (see Figure 5).

Figure 5. ICPSR Collection Development Policy^{xv}



ICPSR Collection Development Policy

Executive Summary

ICPSR maintains an extensive archive of data to support research and knowledge building in the social and behavioral sciences. This policy sets forth a description of the characteristics of data that ICPSR has interest in adding to the collection. ICPSR intentionally casts a broad net in order to add a wide range of data that would be of interest to the diverse fields representing the social and behavioral sciences. However, at the same time the organization applies additional appraisal criteria to determine the appropriate level of curatorial investment that ICPSR will make to ensure long-term and effective use of the data.

By balancing our broad interests with a more focused investment in curation, ICPSR is positioned to select a wide array of social and behavioral science data while spending member resources strategically to best anticipate the future needs of the research community and the broad public.

While ICPSR “casts a broad net in order to add a wide range of data that would be of interest to the diverse fields representing the social and behavioral sciences....at the same time the organization applies additional appraisal criteria to determine the appropriate level of curatorial investment that ICPSR will make to ensure long-term and effective use of the data.”^{xvi} The policy defines what ICPSR will and won't accept. Not all data have long-term value, enough supporting information to enable secondary analysis, or can be economically preserved.^{xvii} Content may be rejected or redirected to another, more appropriate repository.

Once staff are sure deposited content is complete and adheres to collection development guidelines, the overall collection is assigned an internal tracking number -- referred to as a 'study

number' -- and moved into the second phase of the ingest process: creating an archival package that is the basis for preservation and access.

Future Improvements

We continue to make and plan revisions for the ICPSR ingest process. These refinements are typically made to increase usability, transparency, or metadata. Some ideas for improvements include:

Update the deposit interface to minimize 'metadata friction'

As explained in a 2011 article by Edwards, Mayernik, Batcheller, Bowker, and Borgman, “Every movement of data across an interface comes at some cost in time, energy, and human attention....[and] represents a point of resistance where data can be garbled, misinterpreted, or lost....Research scientists’ main interest, after all, is in using data, not in describing them for the benefit of invisible, unknown future users, to whom they are not accountable and from whom they receive little if any benefit.”^{xviii} While ICPSR makes every effort to capture as many details as possible at the time of submission, we also seek to make the process streamlined and minimally invasive – all with the goal of eliciting from the depositor a complete and self-explanatory data collection.^{xix} Can we reduce the number of fields, buttons clicked, or files to upload? Is an online form the best mechanism for accepting deposits? Would it be better to not have the depositor complete an online form but instead convey metadata by phone or video conference with an archive staff member? This might capture more complete and accurate information for the archive, and be a better experience for the depositor. Sometimes our attempts at computer-mediated approaches obstruct rather than improve communication.

Provide instantaneous notifications to depositors

While ICPSR provides notifications after deposits are initiated and completed, the messages can be distributed several hours after the point of contact. This might mean that notifications and updates are ignored, discarded, or misinterpreted by the depositor. For deposit inventories, for instance, which include file formats and checksums, immediate notice could make it easier for depositors to spot discrepancies or errors.

Conclusion

Trust is based on transparent adherence to and certification against community standards. Archival trust is initiated with the ingest process, particularly the portion of ingest dealing with acquisition of content. Examples of submission and ‘pre-ingest’ activities implemented at ICPSR have been provided. Possible future refinements also have been discussed. These activities help insure long-term access to valuable research data.

Acknowledgements

Thanks to Cole Whiteman for providing the ICPSR Pipeline Process diagram. Numerous ICPSR staff have developed and refined the pipeline, including the ingest process, over the past decade. This chapter would not have been possible without their dedicated work.

ⁱ Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data System Standards, CCSDS 650.0-M-2. Magenta Book. Issue 2. Washington, D.C.: CCSDS, June 2012. [Also published as ISO 14721:2012.]

-
- ⁱⁱ Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report. Mountain View, CA: RLG, May 2002.
- ⁱⁱⁱ *Ibid.* See pages 44-45.
- ^{iv} King, Gary. 1995. "Replication, Replication," *PS: Political Science and Politics*, 28: 444–452. Accessed January 13, 2016, <http://dx.doi.org/10.2307/420301>.
- ^v Inter-university Consortium for Political and Social Research (ICPSR). Sustaining Domain Repositories for Digital Data: A Call for Change from an Interdisciplinary Working Group of Domain Repositories. Position statement. "Sustaining Domain Repositories Meeting," Ann Arbor, MI, June 24-25, 2013. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2013. DOI: 10.3886/Sustaining Domain Repositories Digital Data.
- ^{vi} John Holdren. Memorandum for the Heads of Executive Departments and Agencies. Subject: Increasing Access to the Results of Federally Funded Scientific Research. Washington, DC: Executive Office of the President, Office of Science and Technology Policy. February 22, 2013). Accessed January 13, 2016, www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
- ^{vii} Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Chicago, IL: CRL, The Center for Research Libraries, February 2007.
- ^{viii} Audit and Certification of Trustworthy Repositories. Recommendation for Space Data System Practices, CCSDS 652.0-M-1. Magenta Book. Issue 1. Washington, D.C.: CCSDS, September 2011. [Also published as ISO 16363:2012.]
- ^{ix} ICPSR Audit Report. Center for Research Libraries. October 24, 2006. Accessed January 13, 2016, http://www.crl.edu/sites/default/files/d6/attachments/pages/ICPSR_final.pdf

^x Data Seal of Approval web site. Accessed January 13, 2016,

https://assessment.datasealofapproval.org/assessment_114/seal/html/

^{xi} International Council for Science (ICSU) World Data System web site. Accessed January 13,

2016, <http://www.icsu-wds.org/>

^{xii} While physical materials may be submitted by mail, all electronic content -- the bulk of deposited items -- are transferred via the online form.

^{xiii} ICPSR Deposit Data web page. Accessed January 13, 2016,

<http://www.icpsr.umich.edu/icpsrweb/deposit/>

^{xiv} ICPSR Deposit Data web page. Accessed January 13, 2016,

<http://www.icpsr.umich.edu/icpsrweb/deposit/>

^{xv} ICPSR Collection Development Policy, July 7, 2015. Accessed January 13, 2016,

<http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/policies/colldev.html>

^{xvi} ICPSR Collection Development Policy, July 7, 2015. Accessed January 13, 2016,

<http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/policies/colldev.html>

^{xvii} Gutmann, M, Schürer, K., Donakowski, D., Beedham, H. “The Selection, Appraisal, and Retention of Digital Social Science Data.” *Data Science Journal*, Volume 3, 30 December 2004.

^{xviii} Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker and Christine L. Borgman. *Social Studies of Science* 2011 41: 667. Originally published online 15 August 2011. DOI: 10.1177/0306312711413314.

^{xix} A well-prepared data collection “contains information intended to be complete and self-explanatory” for future users. See National Longitudinal Survey of Youth. (2010). *NLSY97 Documentation*. In *NLSY97 User’s Guide* (Chapter 3.3). Accessed January 11, 2016,

<https://web.archive.org/web/20100727011626/http://www.nlsinfo.org/nlsy97/97guide/chap3.htm>