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.¹ This chapter will focus on the portion of ingest dealing with acquisition of content -- also referred to as “submission and ‘pre-ingest’ activities”.² These activities include:³

- 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. 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. 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. Fourth, repositories themselves are looking for
operating standards for trustworthiness against which they may self-assess management, operations, and technologies.  

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 Approvalx and the ICSU World Data Systemxi. 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.
**Submission through a Deposit Form**

All electronic content is submitted to ICPSR via an online deposit form. 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

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.
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

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

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

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.”

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.

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.

---


iii Ibid. See pages 44-45.


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


