Designing the Workflow for a Digital Data Repository

Cole Whiteman
University of Michigan - Ann Arbor, colew@umich.edu

Follow this and additional works at: http://docs.lib.purdue.edu/ddad2011
Designing the Workflow for a Digital Data Repository

Cole Whiteman
University of Michigan
Inter-university Consortium for Political and Social Research (ICPSR)
colew@umich.edu

Symposium on Data-Driven Approaches to Droughts (DDAD2011)
Wednesday, June 11, 2011
Case Study: ICPSR

- The “Inter-university Consortium for Political and Social Research”
- Hosted by the Institute for Social Research at the University of Michigan
- One of the world’s oldest and largest social science data archives, est. 1962
- Distributed data on punch cards, then on reel-to-reel tape, now on the web

Explore: http://www.icpsr.umich.edu
<table>
<thead>
<tr>
<th><strong>Membership</strong></th>
<th>~700 member institutions worldwide.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Holdings</strong></td>
<td>~7000 studies, consisting of ~65,000 datasets, consisting in turn of ~550,000 files.</td>
</tr>
<tr>
<td><strong>Downloads</strong></td>
<td>~612,500 dataset downloads last year.</td>
</tr>
<tr>
<td><strong>Users</strong></td>
<td>~19,800 active user accounts. (“Active” = downloaded or otherwise accessed something.)</td>
</tr>
</tbody>
</table>
ICPSR’s data repository enables social scientists, as users, to:

• Explore, understand, evaluate original findings in detail
• Replicate
• Extend original findings with secondary analyses

And, as depositors, to:

• Offload the ongoing task of preserving data accessibility
• Protect it from damage, loss, obsolescence
• Offload user support
• Fulfill archival obligations to funding agencies
ICPSR: Outsider’s View

World of Political and Social Research

Studies are...

1. Submitted to ICPSR
   Transmitted by an Investigator via:
   • Post,
   • Electronic mail, or
   • Electronic file transfer

2. Obtained by ICPSR
   Downloaded from a publicly accessible web/ftp site

3. Posted to the Web
   • Thoroughly Q/A'd
   • Available on the Web
   • Catalogued on the Web
   • Announced on Website and by email to user groups

4. Archived
   in long-term storage
The ICPSR Portal

ICPSR
INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

Our Mission
ICPSR provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community.

Announcements
- ICPSR awarded Data Seal of Approval
- New Releases through 2011-06-12
- Give us feedback
- ICPSR participating in IASSIST conference
- Changes to authentication
- ICPSR Internship selections announced

Featured Site (view all)
Integrated Fertility Survey Series (IFSS)

IFSS offers data and tools for examining issues related to families and fertility in the United States spanning five decades. IFSS encompasses the Growth of American Families (GAF), National Fertility Surveys (NFS), and National Surveys of Family Growth (NSFG), as well as a single dataset of harmonized variables across all ten surveys.
Access to a Data Collection
Guide to Social Science Data Preparation and Archiving

Best Practice Throughout the Data Life Cycle

4th edition
Planning for Preservation over the Data Life Cycle

1. Anticipate archiving costs and challenges
2. Create a data management plan
3. Follow best practices for data and documentation
4. Manage master datasets and work files
5. Determine file formats to deposit
6. Comply with dissemination standards and formats
7. Set up support for data users
Workflow Building Blocks

Task

Owner

input

output

Other input

Other output
Using the Workflow to Organize Resources

Step 2: Deposit Study Material

The Depositor -- that is, the Investigator or someone depositing material on behalf of the Investigator -- prepares and submits study materials to ICPSR for dissemination and preservation. The Investigator signs the deposit, which transfers custody over it to ICPSR.

Typically, the Depositor performs the deposit directly into the ICPSR web site’s Deposit Form.

Details
- Submitting a Study
- Submission Criteria
- Data Collection Preparation Tips
- Mode of Transmission
- Contact ICPSR Acquisitions Staff
- Obtaining a Study
- Depositing Data and Documentation

Resources
- ICPSR Guide to Social Science Data Preparation and Archiving, 4th Edition
- ICPSR Data Deposit Inventory (broken)
- Data Deposit Form
- Handling Removable Media (broken)
- Receiving Physical Materials (broken)
- Virus Protection
- Deposit Viewer
- Deposit View FAQ
- Secure Package Handling Instructions
Step 1: Identify Data Collection

Title and owner of a data collection intended for deposit at ICPSR

1 Identify Data Collection
   Investigator or ICPSR

2 Deposit Study Material
Step 2: Deposit Study Material

1. Identify Data Collection
2. Deposit Study Material
   - Depositor Investigator
   - or Processor
3. Review and Plan
Step 3: Review and Plan
Step 4: Build Data Set
Step 5: Build Study Description
Step 6: Build Document Set

1. Review and Plan
2. Build Document Set
3. Processor Conversion
4. Build Data Set
5. Build Study Description
6. Study linked to Deposit
7. Document covers
8. Hermes-produced Codebooks
9. Document Files
10. Assemble, Review, Turn Over
Step 7: Assemble, Review, Turn Over

- **Data Files**
  - Processed data, rendered into a full product suite

- **Metadata**
  - Metadata database records
  - Study Description XML File
  - Study Description PDF File

- **Document Files**

- **Processor’s Turnover Directory**
  - Data
  - Documentation
  - Setups
  - Processing history
  - DDI XML
  - Variable-level metadata

- **Processor Supervisor**

- **Build Document Set**

- **Build Study Description**

- **Build Data Set**

- **Assemble, Review, Turn Over**

- **Release the Study**
Step 8: Release the Study

Processor's Turnover Directory
- AIP
  - Data
  - Documentation
  - Setups
  - Processing history
  - DDI XML variable-level metadata

Metadata
- Study Description XML File
- Study Description PDF File
- Metadata database records

Release the Study

Release Coordinator

Release Datasets for Online Analysis
- Data files, Doc files, and Study Description XML file is installed in web tree
- Study is marked as "available"
- Search engine is re-indexed

Dissemination Web Server
- http://www.icpsr.umich.edu/

Assemble, Review, Turn Over

Announce the Study
Step 9: Announce the Study

8
Release the Study

9
Announce the Study
Web Support
Step 10: Prepare for Online Analysis
Step 11: Release for Online Analysis

- Prepare Datasets for Online Analysis
- Release for Online Analysis
- Database record of SDA component
- SDA Files installed in web tree
- Release Datasets for Online Analysis Supervisor
- Dissemination Web Server
  - http://www.icpsr.umich.edu/
**Step 12: Access the Study**

**Dissemination Web Server**

http://www.icpsr.umich.edu/

**Search & Browse**
- Study is discoverable on the ICPSR public web site via search engine
- Study Description is visible

**Download**
- Study files are downloadable
  - Selected data and doc files
  - All other doc files for this study and the selected datasets
  - Study Description PDF file
  - Bibliography file
  - Terms of Use file
  - File manifest

**Analyze Data Online**

**Access the Study User**
• **Two Pipelines:** one for input (Deposits) and one for output (Studies). Typically 1 Deposit → 1 Study, but often m→1, 1→n, or even m→n.

• **Processing:** ICPSR devotes significant effort to build packages for dissemination and archiving from deposits.

• **Storage:** ICPSR maintains separate Ingest Storage, Archival Storage and Dissemination Storage.
Typical Repository Design Decisions

• Primary focus of effort: delivery or preservation?
  – Shall we make the deposit available for access as quickly as possible, and then prep it for preservation? Or preserve it first, then open up access?

• Design priority: ease of deposit, ease of preservation, or ease of delivery?
  – Requirements for acceptance: narrow and rigid, or wide and flexible?
  – Shall we accept only certain formats? Certain types of content?
  – What metadata do we accept? What metadata do we require?

• Preservation commitment: bitstream or intellectual content?
  – What do we do with holdings when their data format becomes obsolete?

• Who can deposit data?
  – Anyone? Or only certain qualified contributors?

• Who can access data? Can anyone access everything?
  – Or shall certain material have access restrictions?
Thanks!

Cole Whiteman
Process Architect

ICPSR
University of Michigan
Institute for Social Research
P.O. Box 1248, Ann Arbor, MI 48106-1248

734-615-7935
colew@umich.edu