6-22-2011

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

http://docs.lib.purdue.edu/ddad2011/10

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
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>Scale of ICPSR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Membership</strong></td>
</tr>
<tr>
<td>~700 member institutions worldwide.</td>
</tr>
<tr>
<td><strong>Holdings</strong></td>
</tr>
<tr>
<td>~7000 studies, consisting of ~65,000 datasets, consisting in turn of ~550,000 files.</td>
</tr>
<tr>
<td><strong>Downloads</strong></td>
</tr>
<tr>
<td>~612,500 dataset downloads last year.</td>
</tr>
<tr>
<td><strong>Users</strong></td>
</tr>
<tr>
<td>~19,800 active user accounts. (&quot;Active&quot; = 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

- Investigators
- Current Researchers
- Future Researchers

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

then

3. Posted to the Web
   - Thoroughly QA'd
   - Available on the Web
   - Catalogued on the Web
   - Announced on Website and by email to user groups

and

4. Archived
   in long-term storage

Hardcopy + Digital Linear Tape
The ICPSR Portal

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

See More Announcements...

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

Great Plains Population and Environment Data: Agricultural Data, 1870-1997 [United States]

Principal Investigator(s): Gutmann, Myron P.

Summary: The data in this series of studies were assembled by an interdisciplinary research team led by Myron Gutmann of the University of Michigan between 1995 and 2004, as part of a research project funded by the National Institute of Child Health and Human Development (Grant Number R01 HD33445 to the University of Michigan). The goal of the project was to assemble information about approximately 500 counties in 12 states of the Great Plains of the United States, and then to analyze those data in order (View details)

Persistent URL: http://dx.doi.org/10.3886/ICPSR04254

Access Notes

- This study was originally processed, archived, and disseminated by DSDR, a project funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).

What Can I Do With This Collection?

- Explore
  - View study description
  - Browse documentation files (login not required)
  - View related literature (37)
  - View related literature for the entire series
Guidelines for Depositors

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

input

output

Other input

Other output
Deeper Levels of Detail
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

1. Identify Data Collection
   - Investigator
   - ICPSR

2. Deposit Study Material

Title and owner of a data collection intended for deposit at ICPSR
Step 2: Deposit Study Material

1. Identify Data Collection
2. Deposit Study Material
3. Review and Plan

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

Deposit Study Material

Deposited files of data and documentation

Database record of deposit metadata

Signed online deposit form

Depositor Investigator or Processor
Step 3: Review and Plan

2 Deposit Study Material

3 Review and Plan
   Processor Supervisor

4 Build Data Set

5 Build Study Description

6 Build Document Set

Signed Deposit
   SIP

Deposited files of data and documentation

Database record of deposit metadata

Signed online deposit form

Study linked to Deposit

Processing Plan (implicit or explicit)
Step 4: Build Data Set

3. Review and Plan

4. Build Data Set

5. Build Study Description

6. Build Document Set

7. Assemble, Review, Turn Over

8. Prepare Datasets for Online Analysis

9. Data Files

10. Hermes Output for SDA

- Study linked to Deposit
- Processing Plan (implicit or explicit)
- Processed data, rendered into a full product suite
- Fileset (part) names
- Hermes-produced Codebooks
Step 5: Build Study Description
Step 6: Build Document Set

1. Study linked to Deposit
2. Processing Plan (implicit or explicit)
3. Review and Plan
4. Build Data Set
5. Build Study Description
6. Processor Conversion
7. Assemble, Review, Turn Over
Step 7: Assemble, Review, Turn Over

- **Build Data Set**
  - Data Files
    - Processed data, rendered into a full product suite
  - Metadata
    - Metadata database records
    - Study Description XML File
    - Study Description PDF File

- **Build Study Description**
  - Processor Supervisor
  - 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

- **Build Document Set**

- **Step 7: 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

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

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

7 Assemble, Review, Turn Over
8 Release the Study
9 Announce the Study
Step 9: Announce the Study
Step 10: Prepare for Online Analysis
Step 11: Release for Online Analysis

10. Prepare Datasets for Online Analysis

11. Release Datasets for Online Analysis
   Supervisor

8. Release the Study

Dissemination Web Server
http://www.icpsr.umich.edu/
Step 12: Access the Study

- Step 12:
  - Access the Study

Dissemination Web Server

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

Download

- User
  - 12
  - Access the Study

- Dissemination Web Server

- Download DIP
  - 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
About ICPSR’s Workflow

- **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