Digital Lifecycle Research & Consulting

AN OPEN SOURCE, DDI-BASED DATA CURATION SYSTEM FOR SOCIAL SCIENCE DATA

NADDI 2015 – Madison, WI

Two Partners, a Consultant, and a Software Developer

ipa
INNOVATIONS FOR
POVERTY ACTION

Yale
ISPS

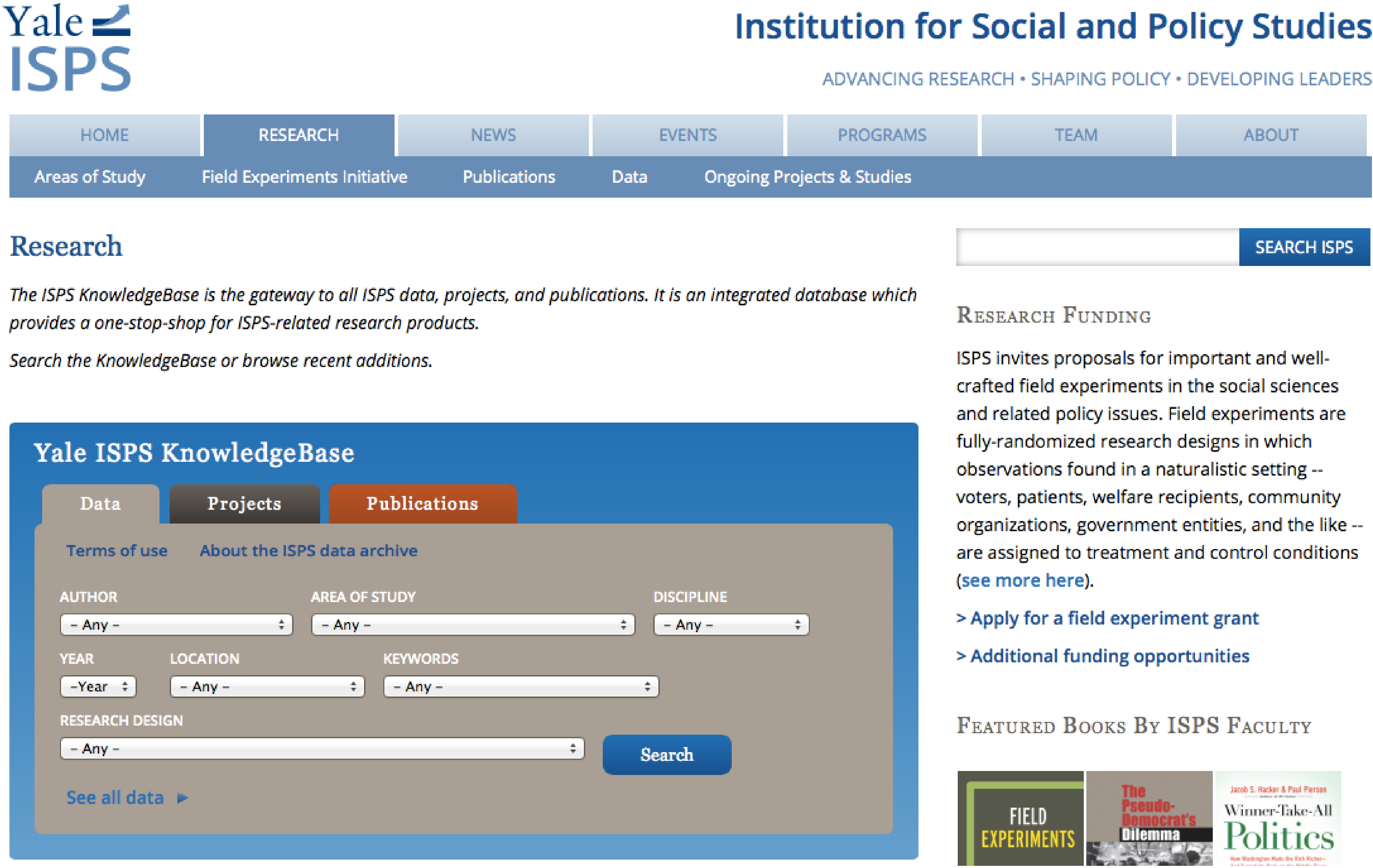
Digital Lifecycle Research & Consulting

colectica

STRATEGIC PLAN ›› 2013–2018

MORE EVIDENCE, LESS POVERTY

ipa
INNOVATIONS FOR
POVERTY ACTION



Institution for Social and Policy Studies

ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS

Research

The ISPS KnowledgeBase is the gateway to all ISPS data, projects, and publications. It is an integrated database which provides a one-stop-shop for ISPS-related research products.

Search the KnowledgeBase or browse recent additions.

Yale ISPS KnowledgeBase

Research Funding

ISPS invites proposals for important and well-crafted field experiments in the social sciences and related policy issues. Field experiments are fully-randomized research designs in which observations found in a naturalistic setting -- voters, patients, welfare recipients, community organizations, government entities, and the like -- are assigned to treatment and control conditions (see more here).

> Apply for a field experiment grant

> Additional funding opportunities

Featured Books By ISPS Faculty

Two Research Organizations

Institution for Social and Policy Studies (Yale)

- Data preparation at end of research project
- Replication
- Field Experiments
- Linked publications, data, and code

Innovations for Poverty Action

- Data preparation before analysis and at end of research project
- Project hosting from distributed research sites
- Lifecycle data management

Why make something new?

The Dataverse Network™ Project

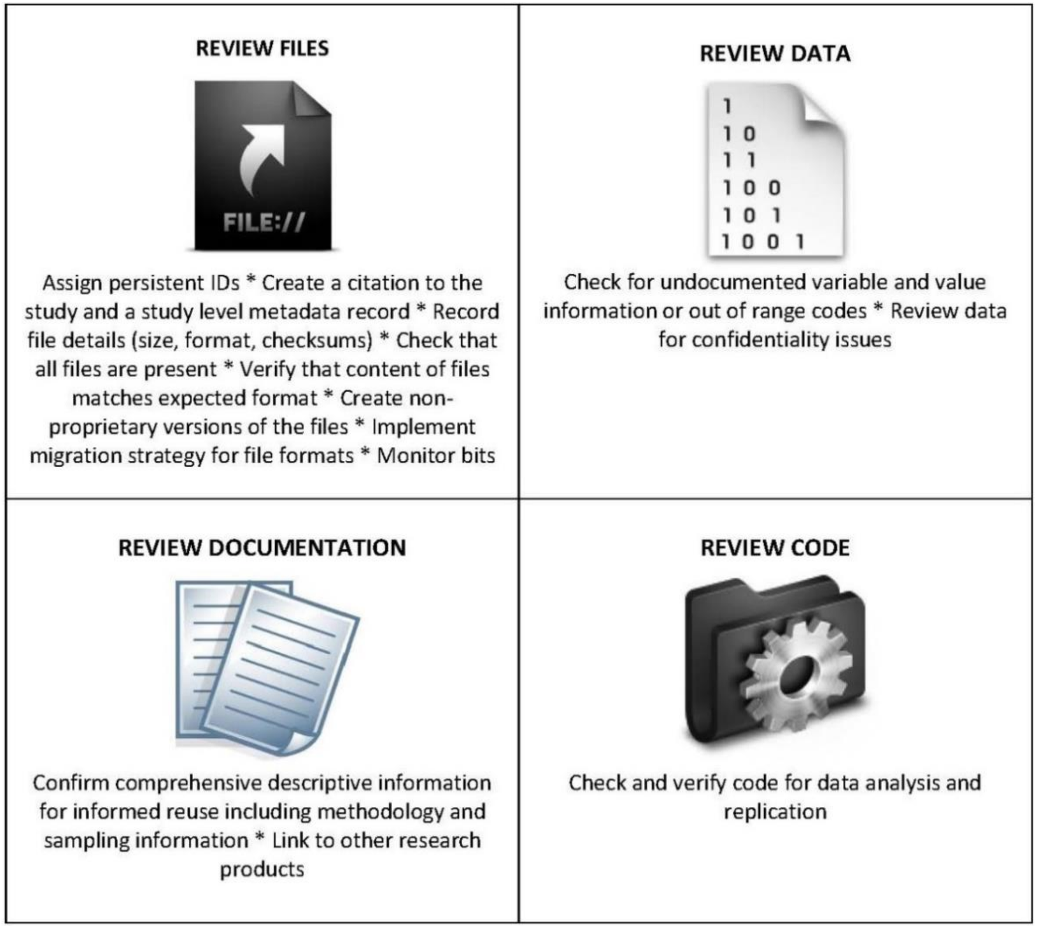
Fedora™

hydra

colectica

The Need: Curation Management

Data Quality Review



REVIEW FILES

Assign persistent IDs * Create a citation to the study and a study level metadata record * Record file details (size, format, checksums) * Check that all files are present * Verify that content of files matches expected format * Create non-proprietary versions of the files * Implement migration strategy for file formats * Monitor bits

REVIEW DATA

Check for undocumented variable and value information or out of range codes * Review data for confidentiality issues

REVIEW DOCUMENTATION

Confirm comprehensive descriptive information for informed reuse including methodology and sampling information * Link to other research products

REVIEW CODE

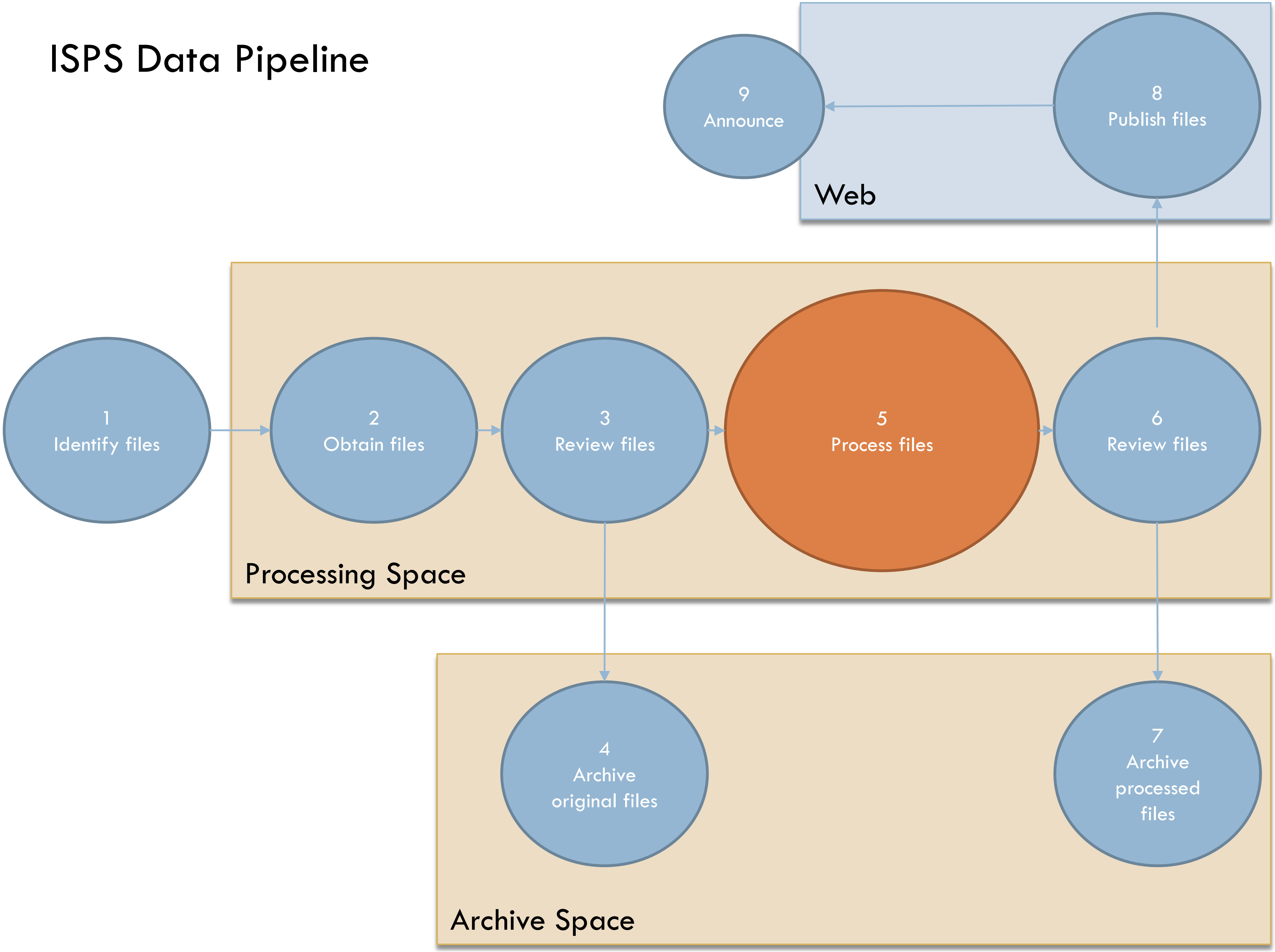
Check and verify code for data analysis and replication

Source: Peer, Green, and Stephenson. 2014. Committing to Data Quality Review. International Journal of Digital Curation, 19(1): 263-291. doi:10.2218/ijdc.v9i1.317

ISPS and IPA Requirements

- Curation workflow management (dashboard)
- Track changes to files (provenance)
- Integrate metadata production with data and code review and cleaning
- Preservation metadata and formats
- Secure storage and access
- Smooth transition to public dissemination of content
- Preference for open source solutions

ISPS Data Pipeline



Web

9
Announce

8
Publish files

Processing Space

1
Identify files

2
Obtain files

3
Review files

5
Process files

6
Review files

Archive Space

4
Archive
original files

7
Archive
processed
files

Processing Steps

1. Assign staff to study and files
2. Move original files to Archive space
3. Make copies of processed files and move to collaborative space
4. Identify related publications and projects
5. Rename all copied files for public dissemination according to ISPS Data Archive naming conventions
6. Check and complete variable-level metadata for each data file
7. Compare variable information, check for additional variables and recoded variables, check variable/value labels
8. Check all files for confidential and other sensitive information
9. Run the statistical code and check against published results
10. Re-write statistical code in R and check replication
11. Communicate with PI as needed
12. Create new DDI-XML file with variable-level information
13. Create additional files by converting to readable formats (e.g., ASCII, PDF)
14. Update study- and file-level metadata record
15. Update tracking documents: process record / general study database / status document

Solution: Build a Curation Management System to Work with Existing Tools

StatTransfer

Neat Features

□ Built on DDI 3.2

□ Web-based

□ Open Source

<ddi>

open source initiative

User Roles

- Depositor
- Curator
- Administrator
- Machines
- Researchers

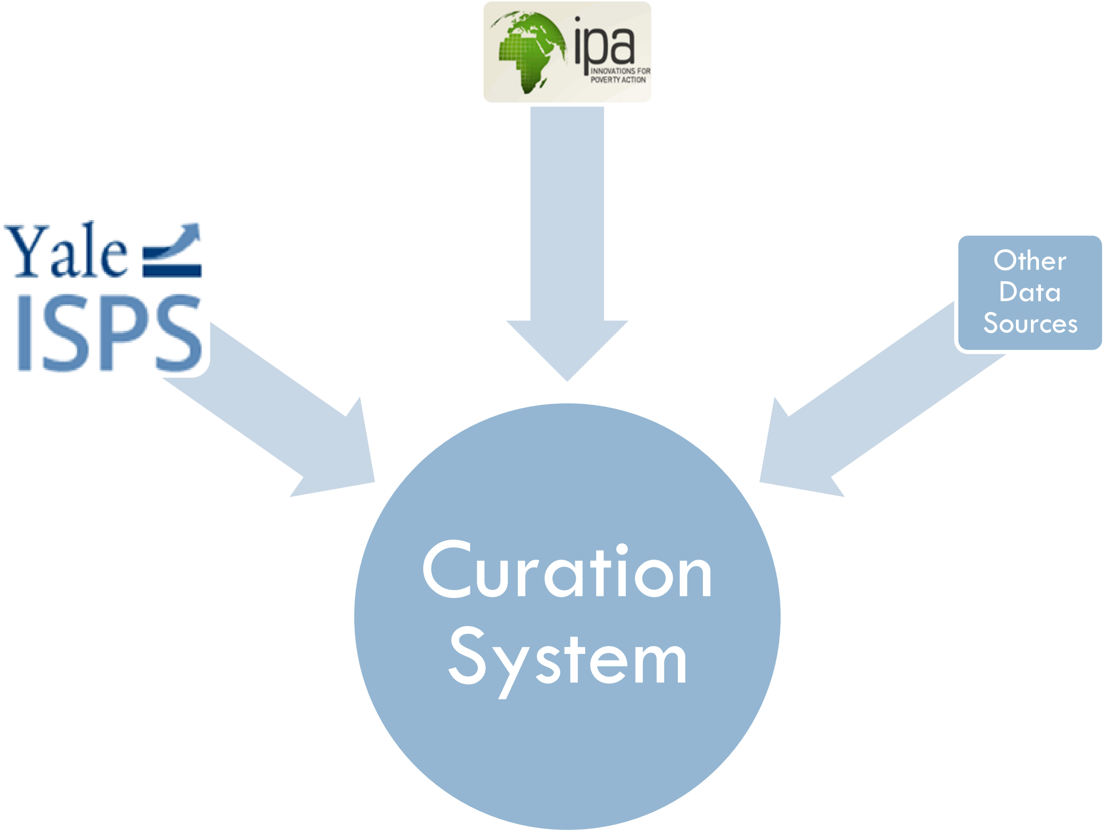
Processing: Example 1

- Goal: Ensure no missing variable labels
- Current Approach
  - Use Stata to open .dta file
  - Manually scan for missing labels
  - Use Stata to edit and save new copy of .dta file
  - Use Excel to make changes to metadata and "process record"

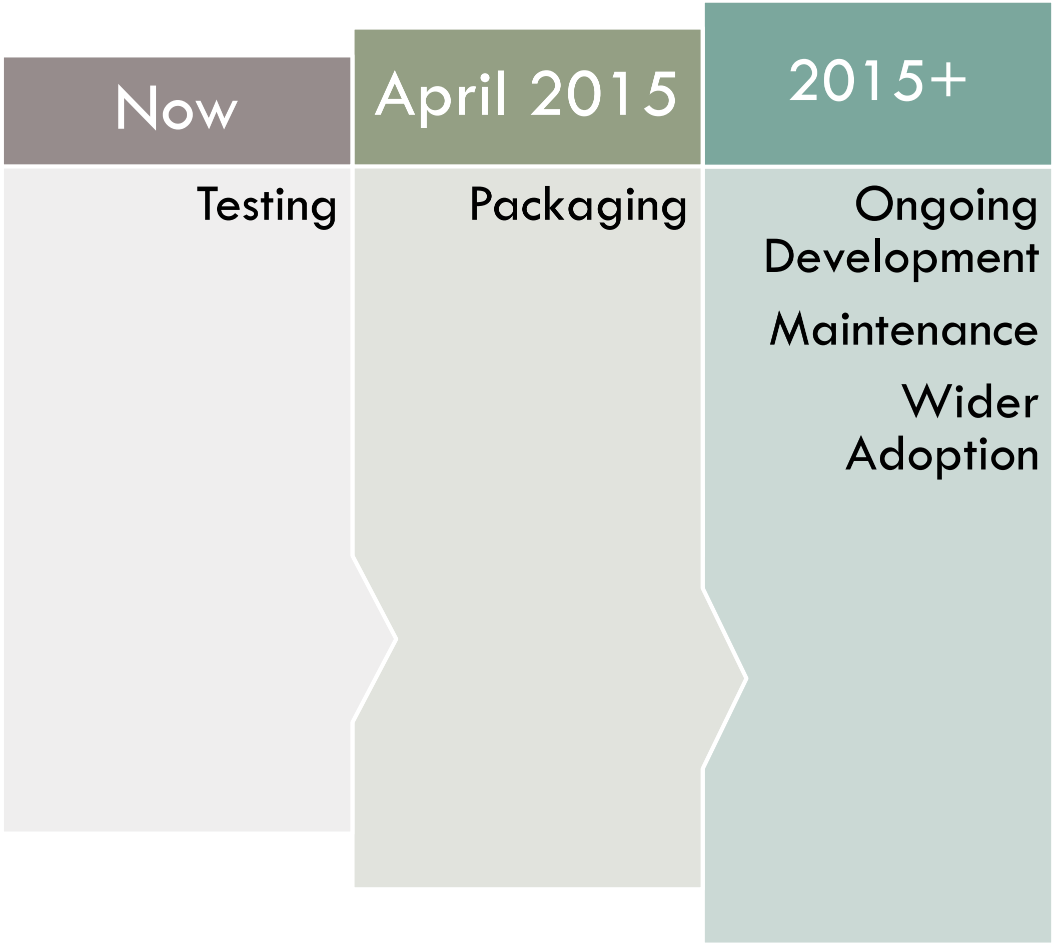
Processing: Example 1

- Goal: Ensure no missing variable labels
- New Approach
  - Curator opens Web application
  - Curator sees a list of variables with missing labels
  - Curator adds labels as appropriate
  - The system logs this information and generates a new .dta file

Data Migration



Timeline



Now

Testing

April 2015

Packaging

2015+

Ongoing Development

Maintenance

Wider Adoption

Demonstration

Thank you

Contributor	Organization	Email
Ann Green	Independent Consultant	green.ann@gmail.com
Jeremy Iverson	Colectica	jeremy@colectica.com
Niall Keleher	Innovations for Poverty Action	nkeleher@poverty-action.org
Limor Peer	Yale University	limor.peer@yale.edu
Dan Smith	Colectica	dan@colectica.com
Stephanie Wykstra	Innovations for Poverty Action	swykstra@poverty-action.org