New Curation Software: Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation

Limor Peer & Stephanie Wykstra

IASSIST, June 2015

Talk outline:

- Background: about IPA & ISPS, and our data policies/repositories
- Data curation – what we do
- More detail on code checks
- Curation workflow and partnership with Colectica
- Software demo
- Conclusion

Background: about our groups

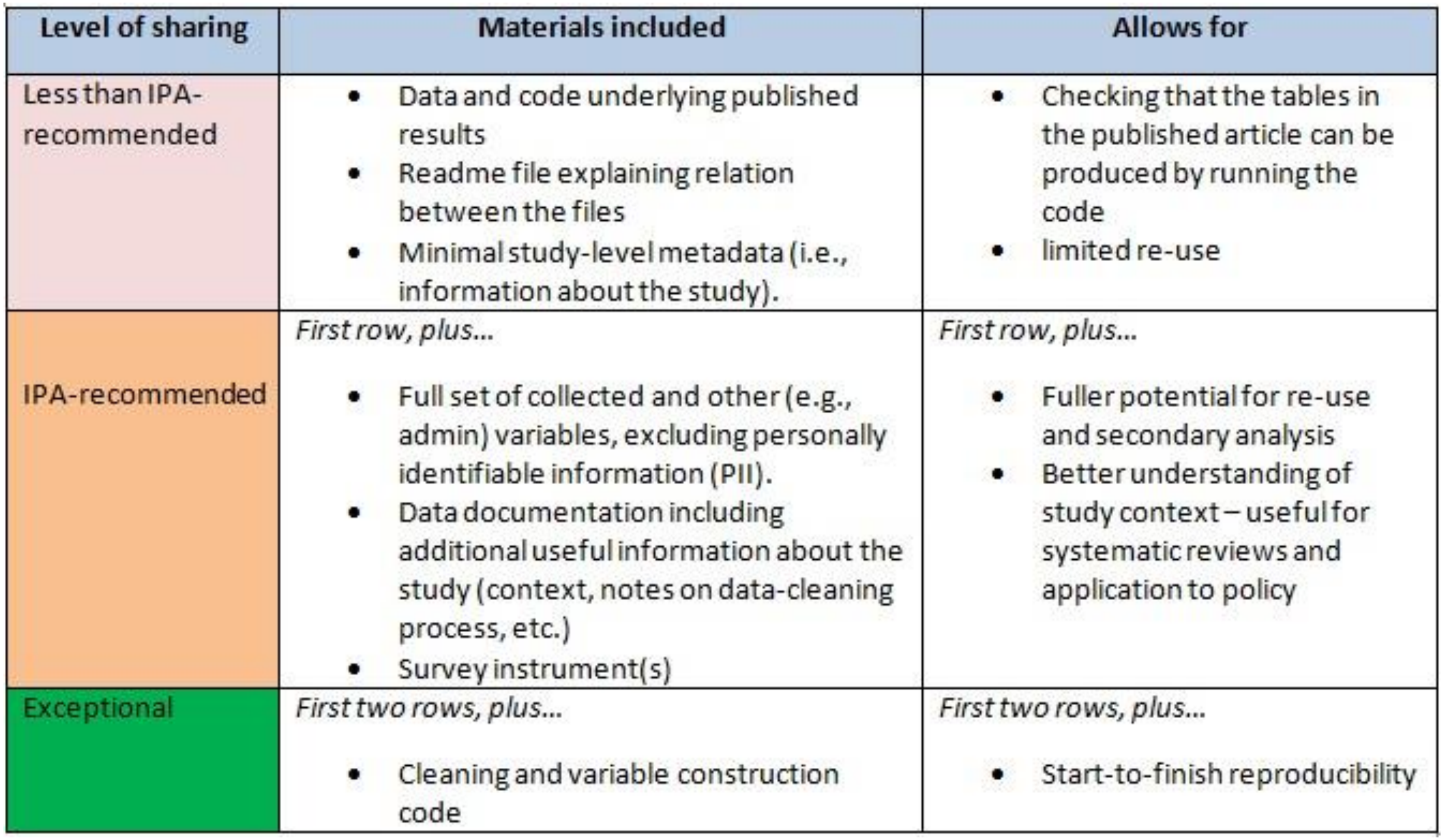
- Innovations for Poverty Action (IPA) works with academic researchers and partners (NGOs, governments) to carry out randomized controlled trials of poverty-alleviation programs in developing countries. Founded in 2002.
  - 300+ ongoing and completed projects.

- Institution for Social And Policy Studies (ISPS) works with academic researchers primarily in political science. Founded in 1968.
  - 100+ studies archived, many of them field experiments.

Background: about our groups

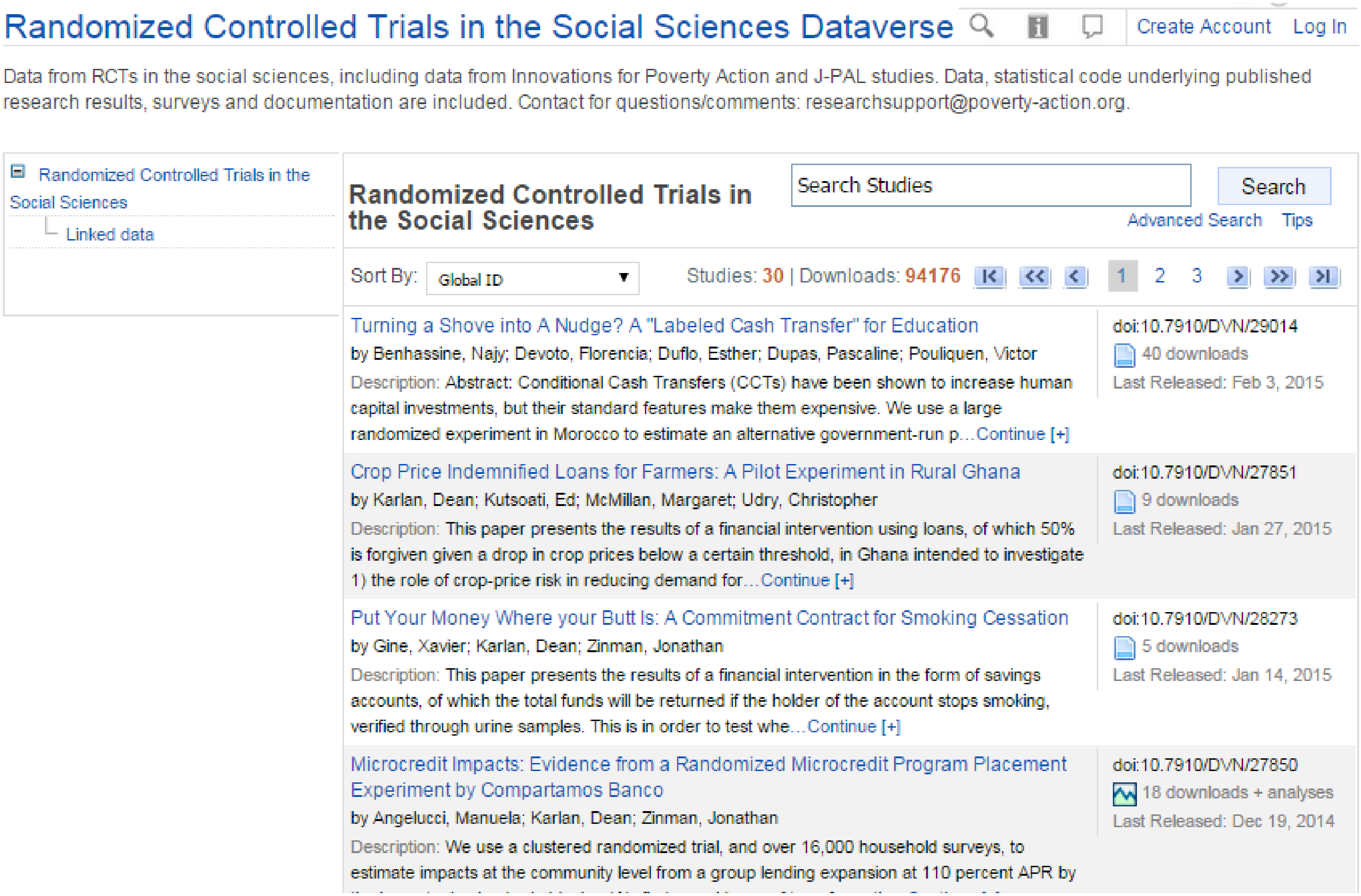
- Both ISPS and IPA have data-sharing policies and data repositories

IPA's data publication policy



Level of sharing	Materials included	Allows for
Less than IPA-recommended	- Data and code underlying published results - Readme file explaining relation between the files - Minimal study-level metadata (i.e., information about the study).	- Checking that the tables in the published article can be produced by running the code - limited re-use
IPA-recommended	*First row, plus…* - Full set of collected and other (e.g., admin) variables, excluding personally identifiable information (PII). - Data documentation including additional useful information about the study (context, notes on data-cleaning process, etc.) - Survey instrument(s)	*First row, plus…* - Fuller potential for re-use and secondary analysis - Better understanding of study context – useful for systematic reviews and application to policy
Exceptional	*First two rows, plus…* - Cleaning and variable construction code	*First two rows, plus…* - Start-to-finish reproducibility

IPA's data repository (via Dataverse)



Randomized Controlled Trials in the Social Sciences Dataverse

Data from RCTs in the social sciences, including data from Innovations for Poverty Action and J-PAL studies. Data, statistical code underlying published research results, surveys and documentation are included. Contact for questions/comments: researchsupport@poverty-action.org.

Randomized Controlled Trials in the Social Sciences

Sort By: Global ID ▾

Studies: 30 | Downloads: 94176

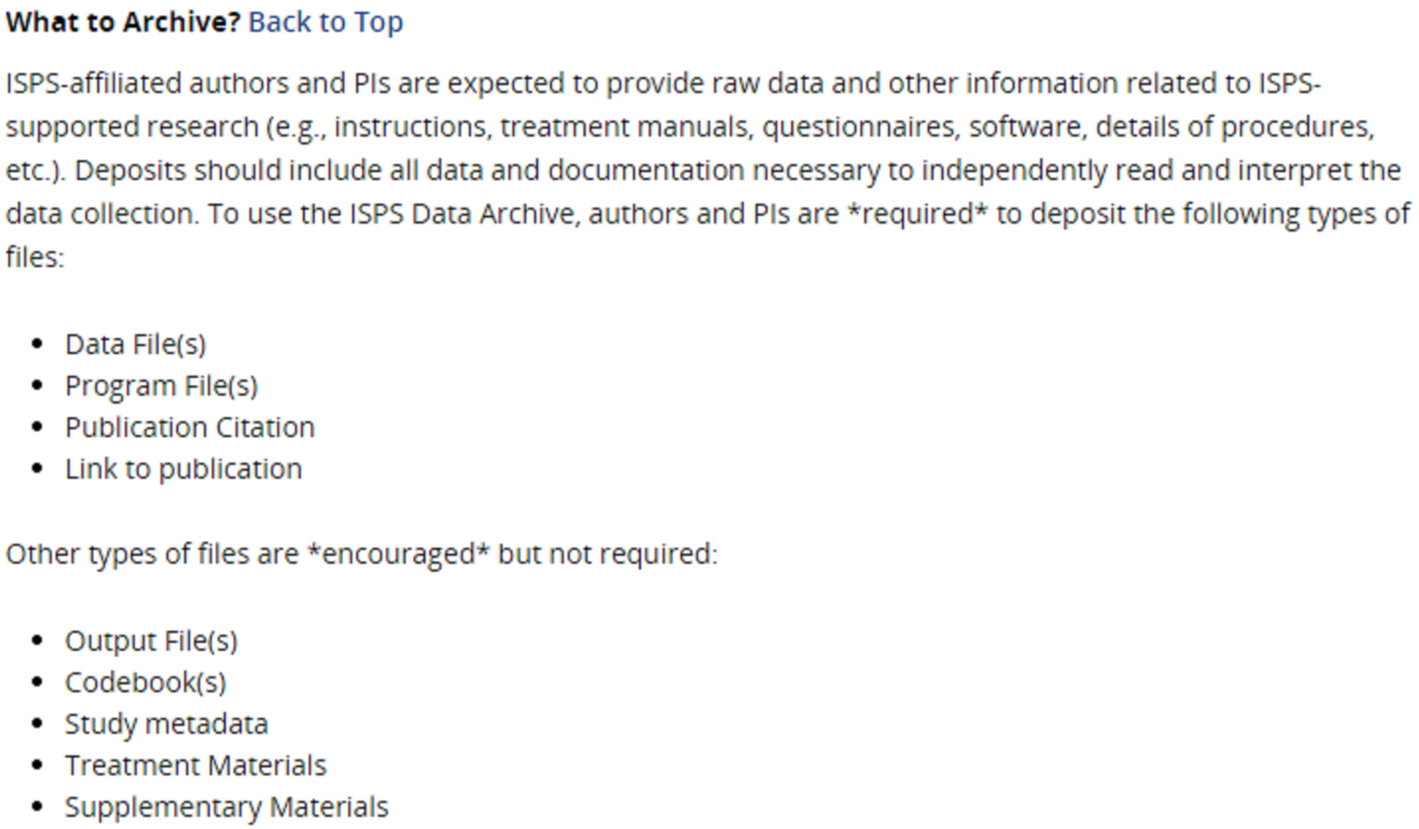
Turning a Shove into A Nudge? A "Labeled Cash Transfer" for Education
by Benhassine, Najy; Devoto, Florencia; Duflo, Esther; Dupas, Pascaline; Pouliquen, Victor
Description: Abstract: Conditional Cash Transfers (CCTs) have been shown to increase human capital investments, but their standard features make them expensive. We use a large randomized experiment in Morocco to estimate an alternative government-run p...Continue [+]
doi:10.7910/DVN/29014
40 downloads
Last Released: Feb 3, 2015

Crop Price Indemnified Loans for Farmers: A Pilot Experiment in Rural Ghana
by Karlan, Dean; Kutsoati, Ed; McMillan, Margaret; Udry, Christopher
Description: This paper presents the results of a financial intervention using loans, of which 50% is forgiven given a drop in crop prices below a certain threshold, in Ghana intended to investigate 1) the role of crop-price risk in reducing demand for...Continue [+]
doi:10.7910/DVN/27851
9 downloads
Last Released: Jan 27, 2015

Put Your Money Where your Butt Is: A Commitment Contract for Smoking Cessation
by Gine, Xavier; Karlan, Dean; Zinman, Jonathan
Description: This paper presents the results of a financial intervention in the form of savings accounts, of which the total funds will be returned if the holder of the account stops smoking, verified through urine samples. This is in order to test whe...Continue [+]
doi:10.7910/DVN/28273
5 downloads
Last Released: Jan 14, 2015

Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco
by Angelucci, Manuela; Karlan, Dean; Zinman, Jonathan
Description: We use a clustered randomized trial, and over 16,000 household surveys, to estimate impacts at the community level from a group lending expansion at 110 percent APR by
doi:10.7910/DVN/27850
18 downloads + analyses
Last Released: Dec 19, 2014

ISPS Data Archive policy



What to Archive? Back to Top

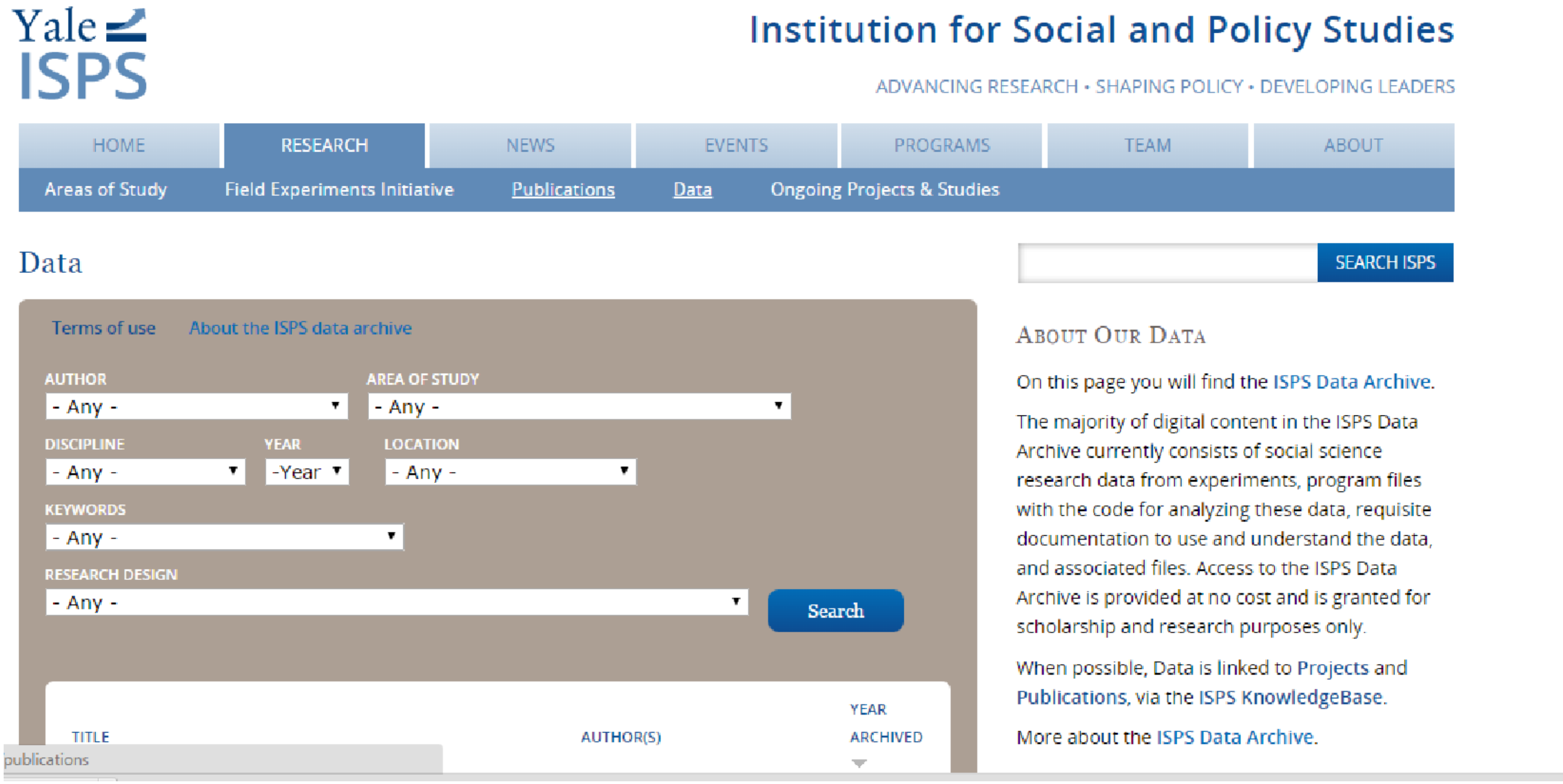
ISPS-affiliated authors and PIs are expected to provide raw data and other information related to ISPS-supported research (e.g., instructions, treatment manuals, questionnaires, software, details of procedures, etc.). Deposits should include all data and documentation necessary to independently read and interpret the data collection. To use the ISPS Data Archive, authors and PIs are *required* to deposit the following types of files:

- Data File(s)
- Program File(s)
- Publication Citation
- Link to publication

Other types of files are *encouraged* but not required:

- Output File(s)
- Codebook(s)
- Study metadata
- Treatment Materials
- Supplementary Materials

ISPS Data Archive



Yale ISPS
Institution for Social and Policy Studies
ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS
HOME
RESEARCH
NEWS
EVENTS
PROGRAMS
TEAM
ABOUT
Areas of Study
Field Experiments Initiative
Publications
Data
Ongoing Projects & Studies
SEARCH ISPS
Search
publications

Data

Terms of use About the ISPS data archive

AUTHOR: - Any -

AREA OF STUDY: - Any -

DISCIPLINE: - Any -

YEAR: -Year

LOCATION: - Any -

KEYWORDS: - Any -

RESEARCH DESIGN: - Any -

TITLE	AUTHOR(S)	YEAR ARCHIVED

About Our Data

On this page you will find the ISPS Data Archive.

The majority of digital content in the ISPS Data Archive currently consists of social science research data from experiments, program files with the code for analyzing these data, requisite documentation to use and understand the data, and associated files. Access to the ISPS Data Archive is provided at no cost and is granted for scholarship and research purposes only.

When possible, Data is linked to Projects and Publications, via the ISPS KnowledgeBase.

More about the ISPS Data Archive.

Data curation

- “Maintaining, preserving and *adding value* to research data.” (Digital Curation Center).

Data curation

- Adding value for which purposes?
  - Reproducible research: ability to re-generate research results from data and code.
  - Data re-use for secondary analysis/ meta-analysis

- Data curation:
  - Making research materials independently usable, in order to allow reproducibility and re-use.

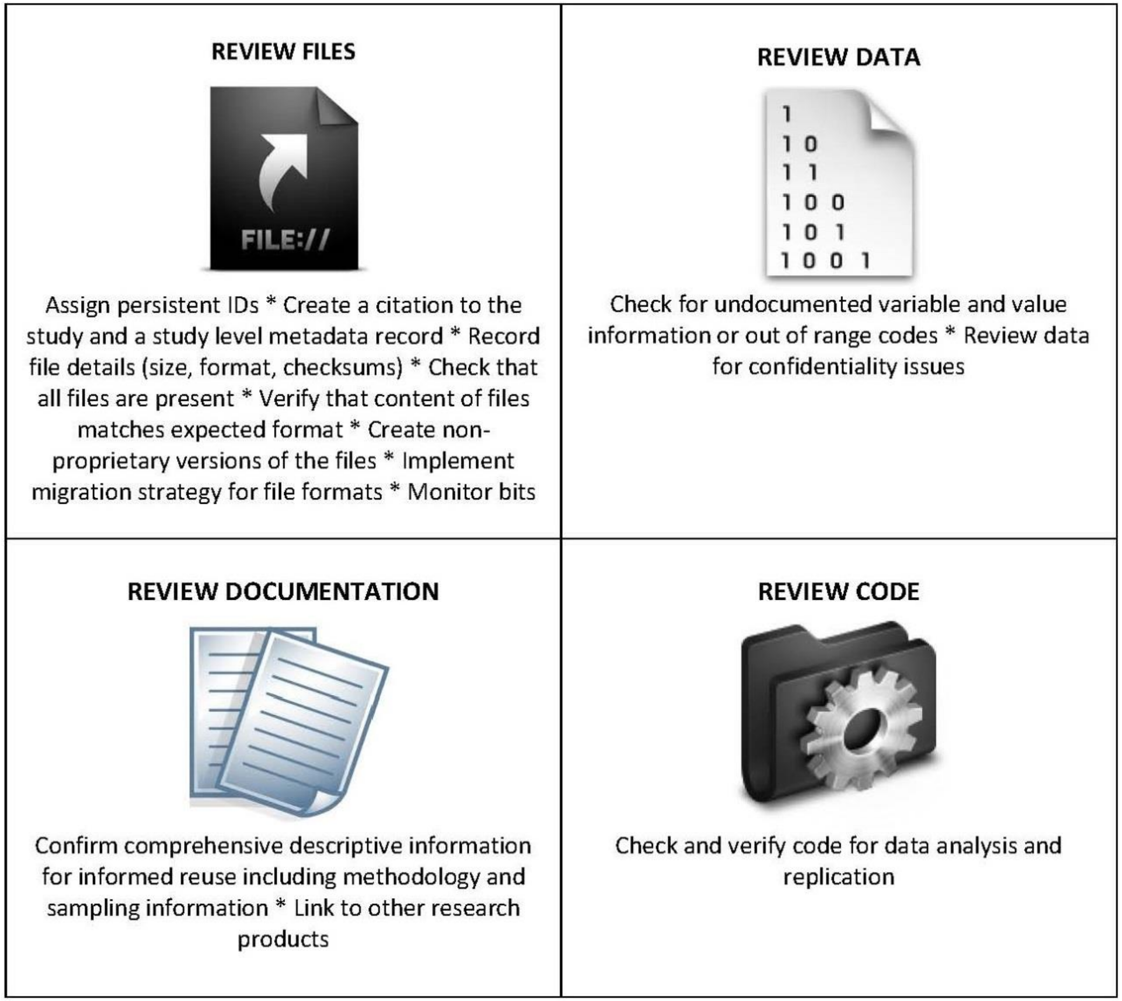
Data curation for independent usability:

- Check for missing variable labels and value codes.
- Compare questionnaire and data.
- Ensure there is no personally-identifiable information (PII)
- Share key study-level metadata.
- Create open formats.
- Confirm code executes.
- Confirm code produces reported results.

Code checks

- **What?**
  - Running the code and confirming that the output matches the tables in the paper.
- **Purpose?**
  - Catching discrepancies i.e., differences between output and publication tables.
  - Often, we have problems *running* the code initially and need to diagnose the issue e.g., different versions of commands, syntax errors, user-generated commands needed, missing code for some published tables.

Data Quality Review Framework



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

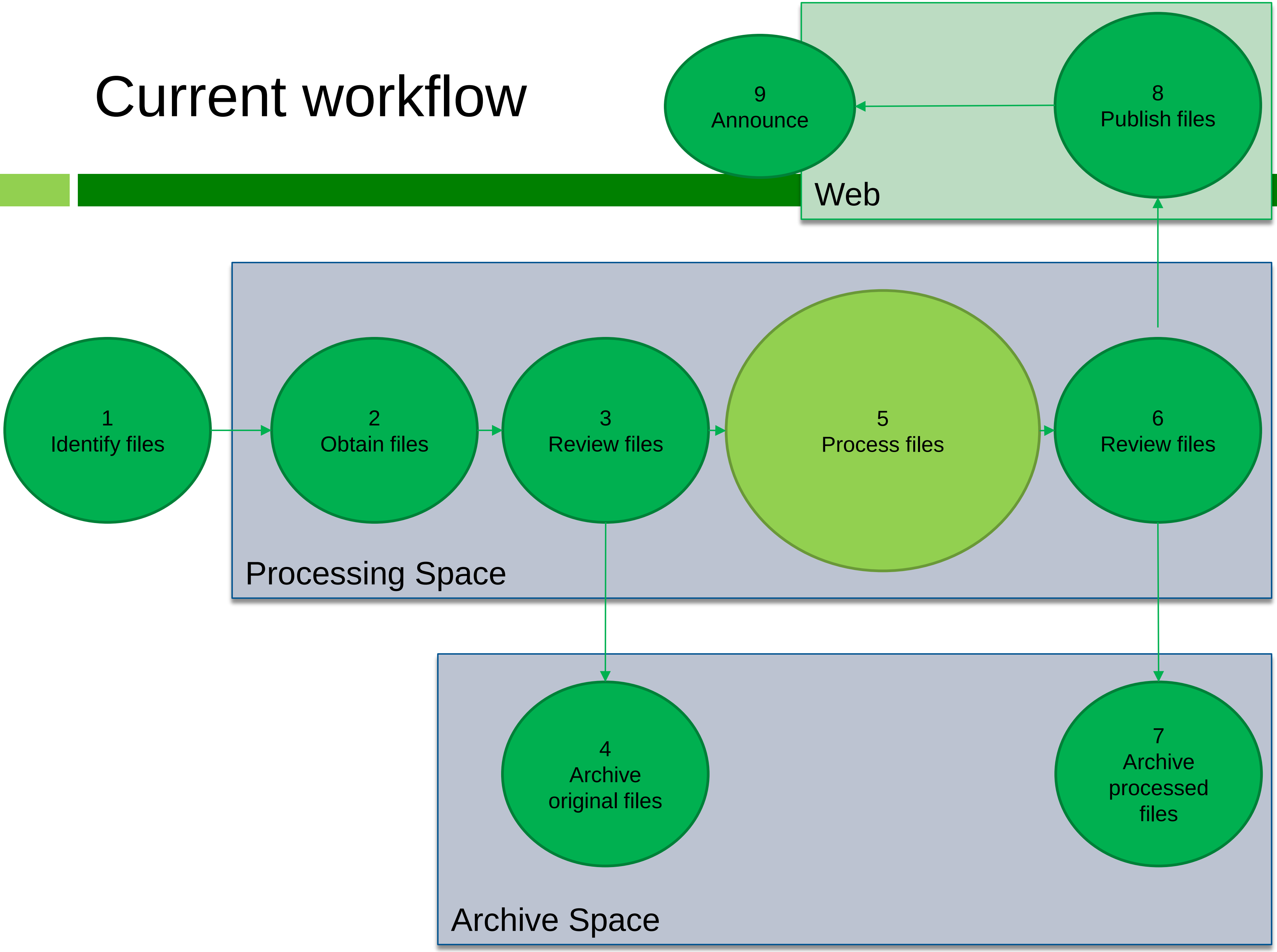
Confirm comprehensive descriptive information for informed reuse including methodology and sampling information * Link to other research products

REVIEW CODE

Check and verify code for data analysis and replication

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

Current workflow



Curation tool: Workflow objectives

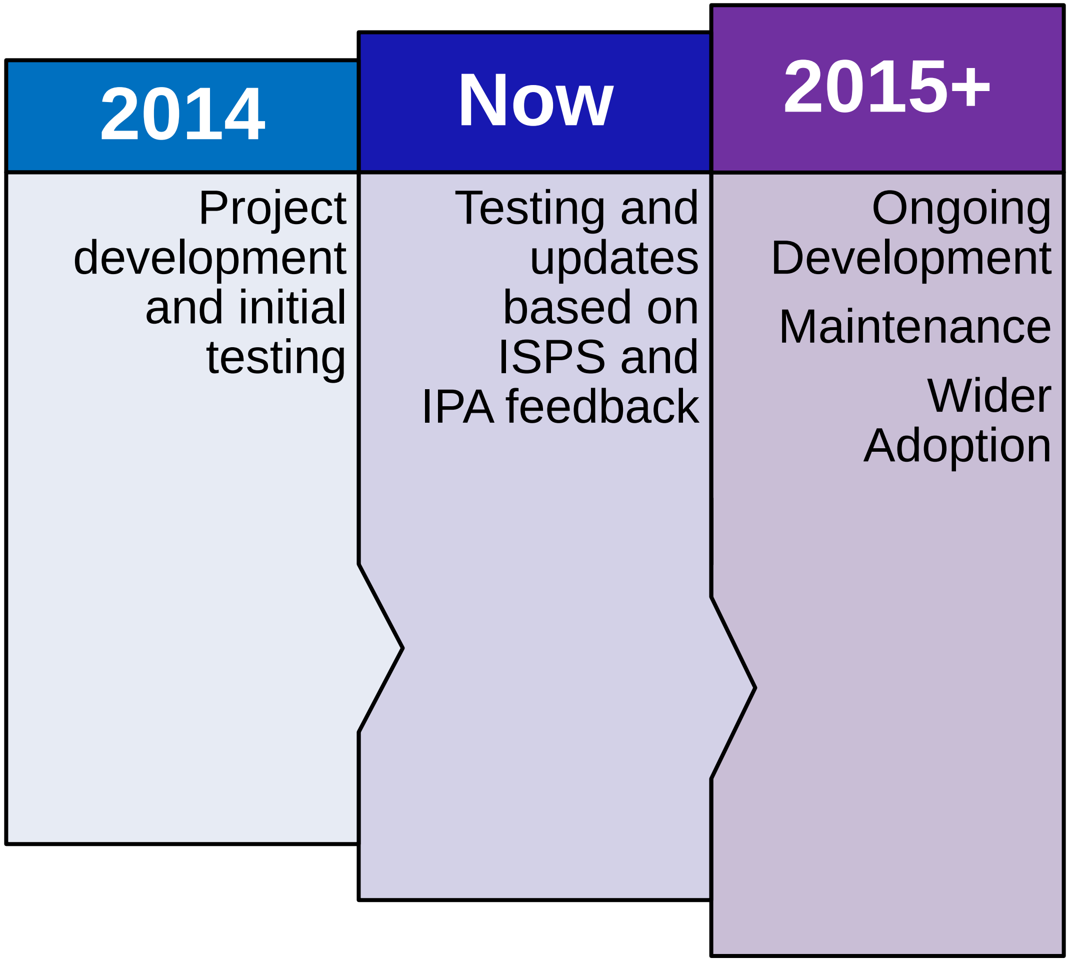
- Automate as many curation tasks as possible
- Technically integrate curation tasks using existing tools
- Track changes to files and metadata

And… to do this using a structured but flexible workflow from deposit to publication and preservation of data and code.

Curation tool: Key features

- Leverages DDI Lifecycle
  - Machine executable, open structured format supports research transparency
  - Plays a part in review tasks – study, file & var levels
- Modular, open-source
  - Could be adapted to changing needs, research methods, dissemination platforms, and preservation solutions
  - Could be used by repositories, researchers, and research staff

Timeline



2014

Project development and initial testing

Now

Testing and updates based on ISPS and IPA feedback

2015+

Ongoing Development

Maintenance

Wider Adoption

Users & Roles

- **DEPOSITOR** Creates a new catalog, adds files, and submits for curation
- **ADMIN** Assigns curator
- **CURATOR** Performs review steps and requests publication.
- **APPROVER** Reviews the record, the history, and approves for publication and/or preservation.

Demonstration: Preliminary!

Curation tool: Summary

- ✓ Structures and tracks the curation workflow
- ✓ Helps automate parts of the data review pipeline
- ✓ Captures all metadata throughout the process
- ✓ Pushes out relevant information to pre-determined destinations
  - ✓ i.e., a user, the archive administrators, a Web based dissemination system, or preservation systems
- ✓ Can fit into repository and research workflows

Thank you!

Limor Peer, limor.peer@yale.edu, @l_peer

Stephanie Wykstra, swykstra@poverty-action.org, @Swykstr

In collaboration with:
Jeremy Iverson & Dan Smith, Colectica
Ann Green, Digital Lifecycle Research and Consulting
Niall Keleher, Innovations for Poverty Action
Yale Academic IT / Research Services & Yale Library