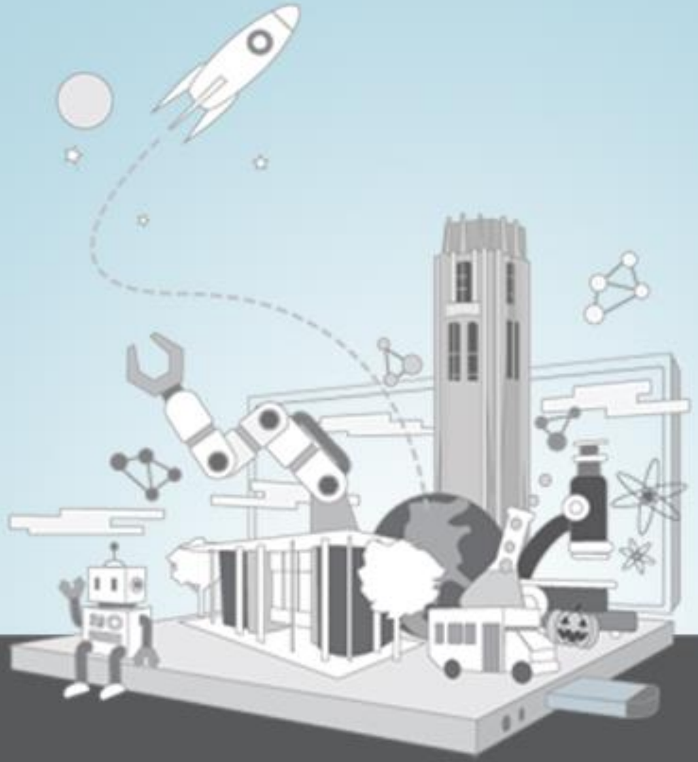
Developing software to prepare social science research data and code for sharing and preservation

Limor Peer

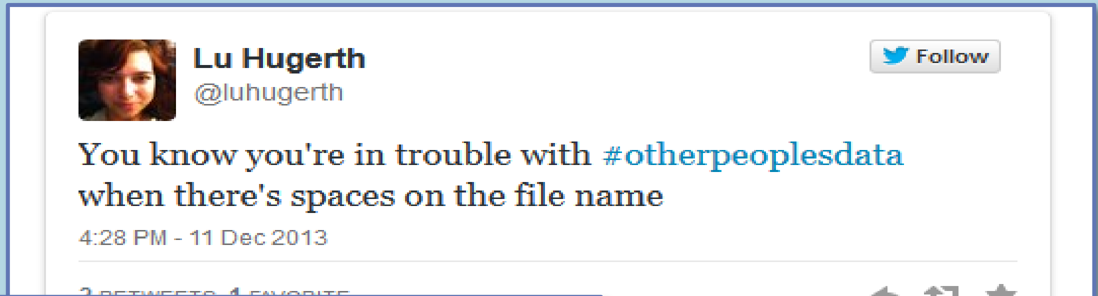
Institution for Social and Policy Studies, Yale University



Data sharing

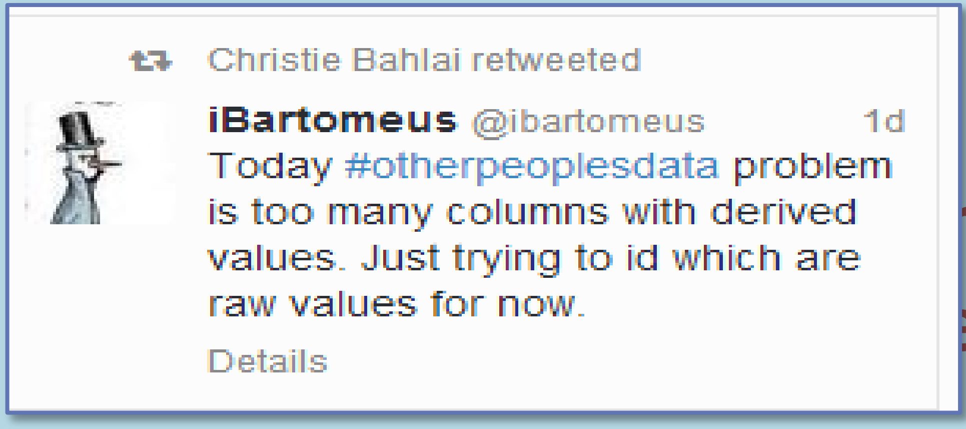
Source: Nature (GARY WATERS/IKON IMAGES/CORBIS)

Using other people's data…

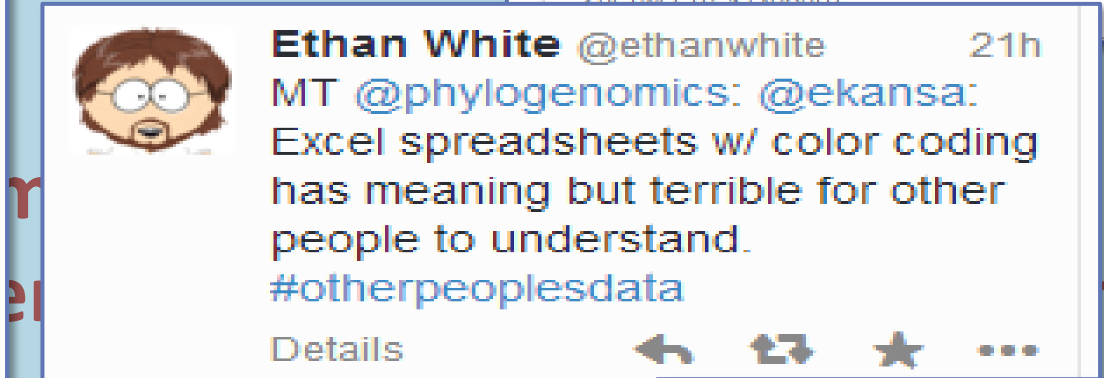

Lu Hugerth
@luhugerth

You know you're in trouble with #otherpeoplesdata when there's spaces on the file name

4:28 PM - 11 Dec 2013


Christie Bahlai retweeted

iBartomeus @ibartomeus 1d
Today #otherpeoplesdata problem is too many columns with derived values. Just trying to id which are raw values for now.

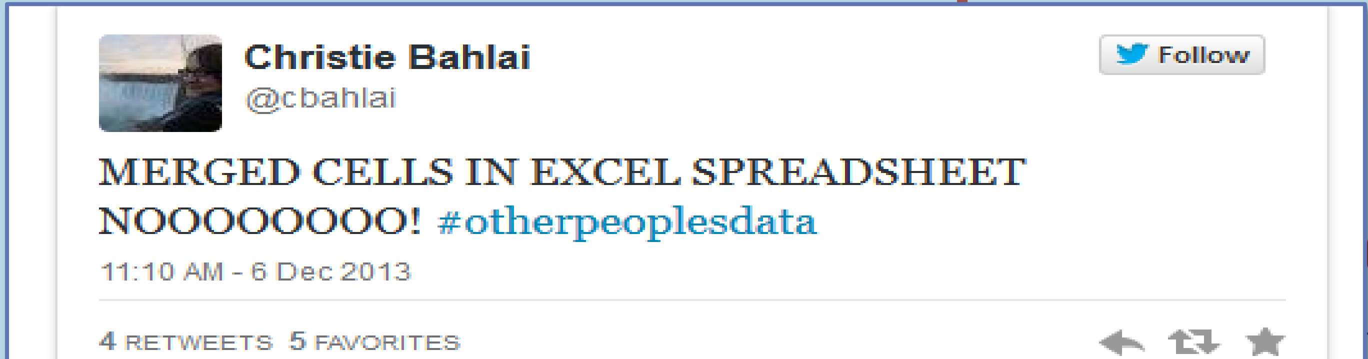

Ethan White @ethanwhite 21h
MT @phylogenomics: @ekansa: Excel spreadsheets w/ color coding has meaning but terrible for other people to understand. #otherpeoplesdata

r

e

n]

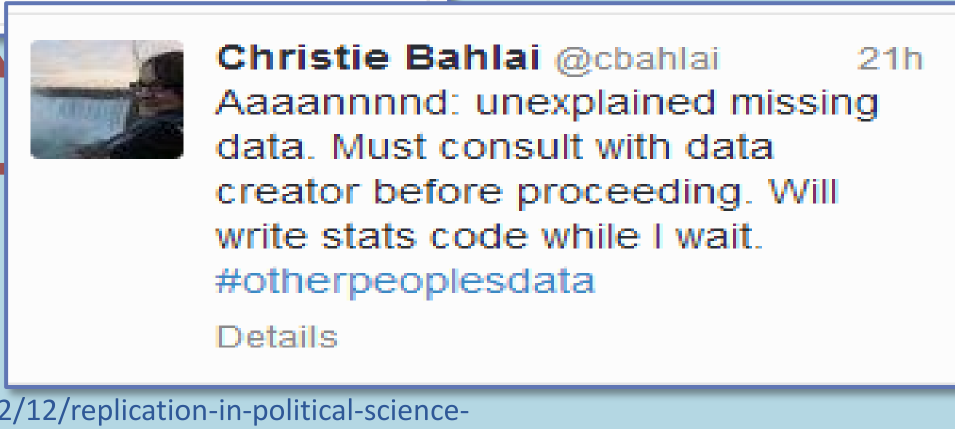
attempts were the lack


Christie Bahlai
@cbahlai

MERGED CELLS IN EXCEL SPREADSHEET NOOOOOOOO! #otherpeoplesdata

11:10 AM - 6 Dec 2013

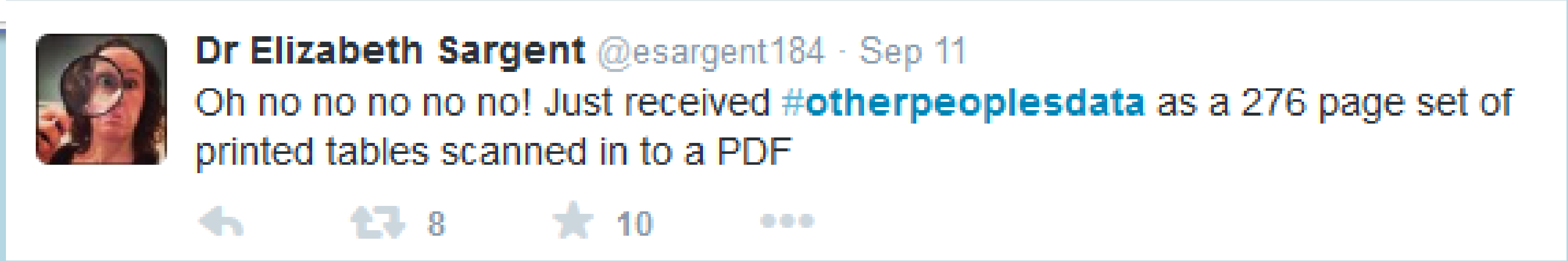
4 RETWEETS 5 FAVORITES


Christie Bahlai @cbahlai 21h
Aaaannnnd: unexplained missing data. Must consult with data creator before proceeding. Will write stats code while I wait. #otherpeoplesdata

insu

on”

-cage/wp/2014/02/12/replication-in-political-science-


Dr Elizabeth Sargent @esargent184 · Sep 11
Oh no no no no no! Just received #otherpeoplesdata as a 276 page set of printed tables scanned in to a PDF

Unusable data = lost data

Image: Shutterstock.com/Lightspring
http://slashdot.org/topic/datacenter/neglect-causes-massive-loss-of-irreplaceable-research-data

Usable data:

Intelligently open

Independently understandable

Outline for today

IF:

Shared (and/or preserved) data may not be usable

THEN:

Make data usable = data curation

Project to develop curation software

- Background
- Requirements
- The software
- The architecture at Yale

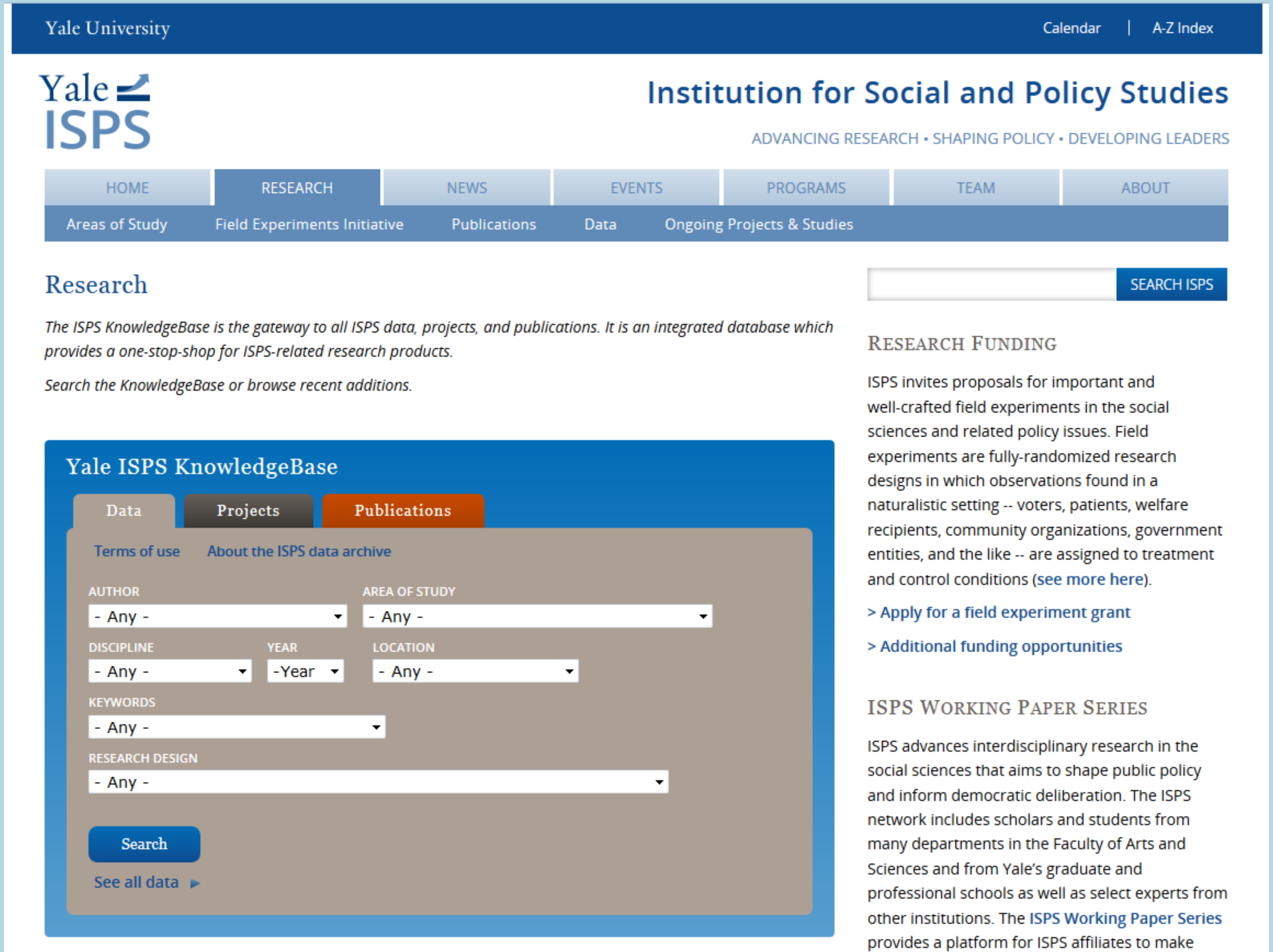
Data curation

Active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education.

Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation.

-- The University of Illinois' Graduate School of Library and Information Science

ISPS Data Archive



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

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 -

See all data ▸

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).

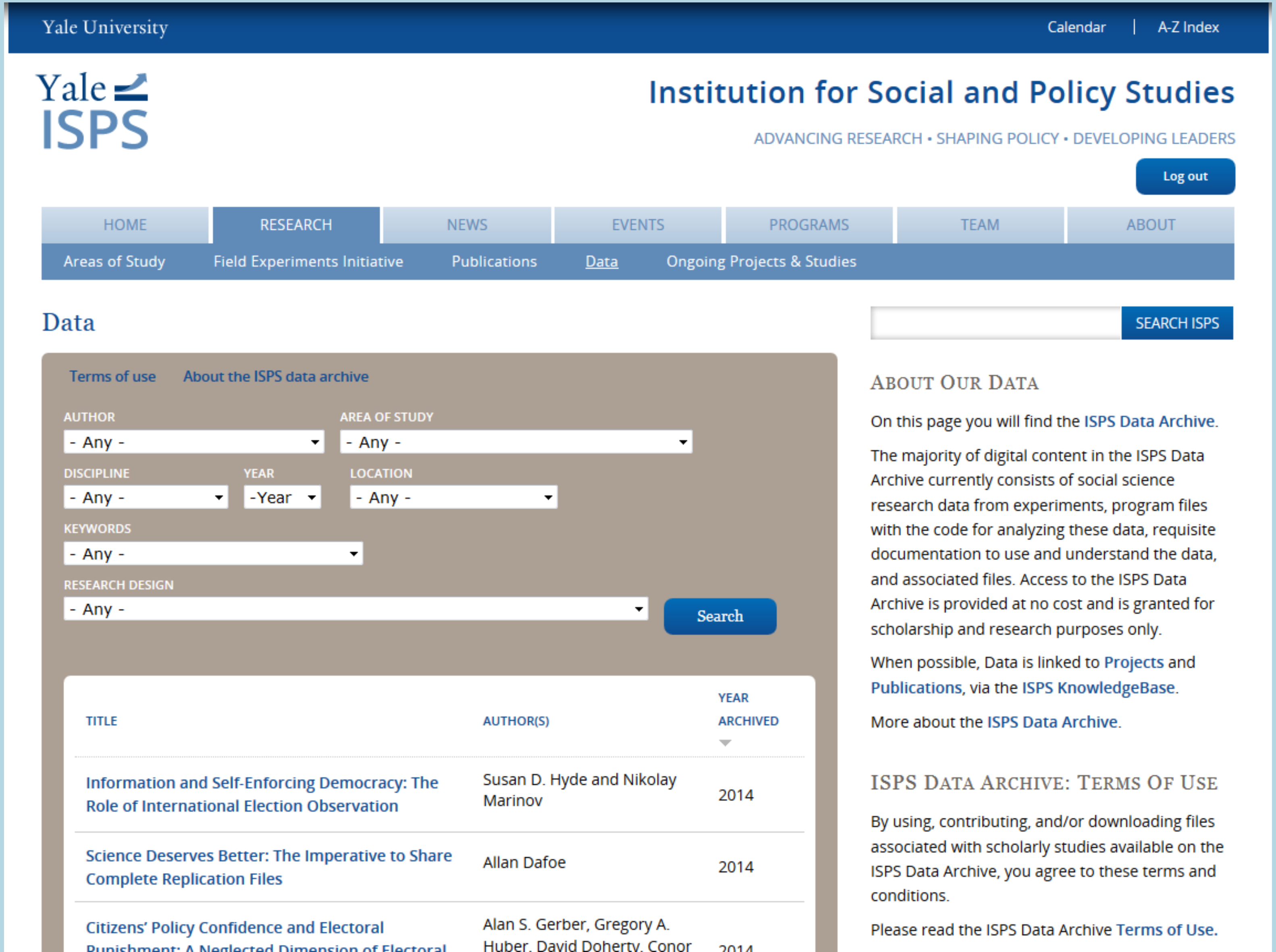
> Apply for a field experiment grant

> Additional funding opportunities

ISPS Working Paper Series

ISPS advances interdisciplinary research in the social sciences that aims to shape public policy and inform democratic deliberation. The ISPS network includes scholars and students from many departments in the Faculty of Arts and Sciences and from Yale's graduate and professional schools as well as select experts from other institutions. The ISPS Working Paper Series provides a platform for ISPS affiliates to make

ISPS Data Archive



Yale University

Calendar | A-Z Index

Yale ISPS

Institution for Social and Policy Studies

ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS

Log out

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 -

Search

TITLE	AUTHOR(S)	YEAR ARCHIVED
Information and Self-Enforcing Democracy: The Role of International Election Observation	Susan D. Hyde and Nikolay Marinov	2014
Science Deserves Better: The Imperative to Share Complete Replication Files	Allan Dafoe	2014
Citizens' Policy Confidence and Electoral Punishment: A Neglected Dimension of Electoral	Alan S. Gerber, Gregory A. Huber, David Doherty, Conor	2014

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.

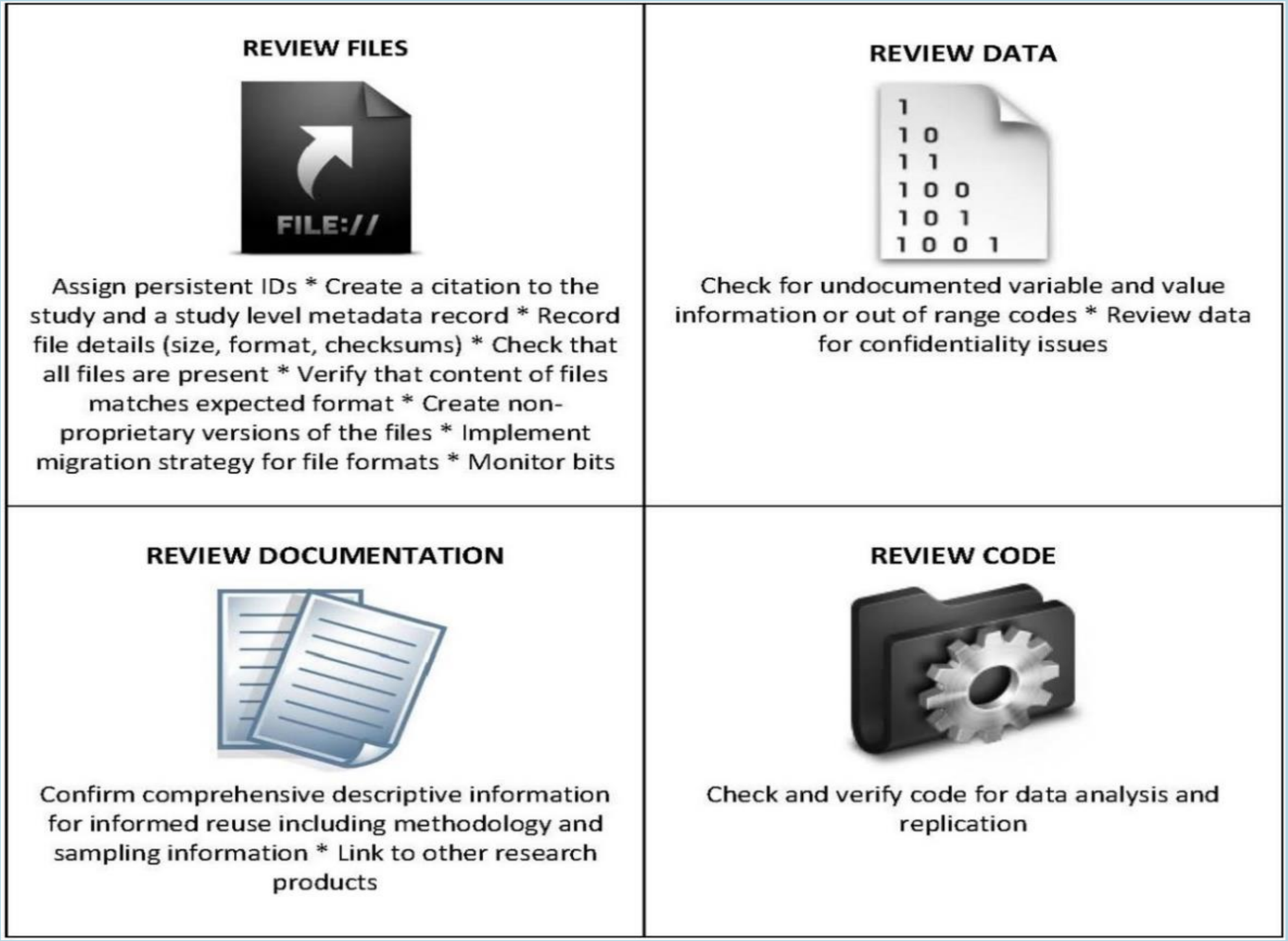
More about the ISPS Data Archive.

ISPS Data Archive: Terms of Use

By using, contributing, and/or downloading files associated with scholarly studies available on the ISPS Data Archive, you agree to these terms and conditions.

Please read the ISPS Data Archive Terms of Use.

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

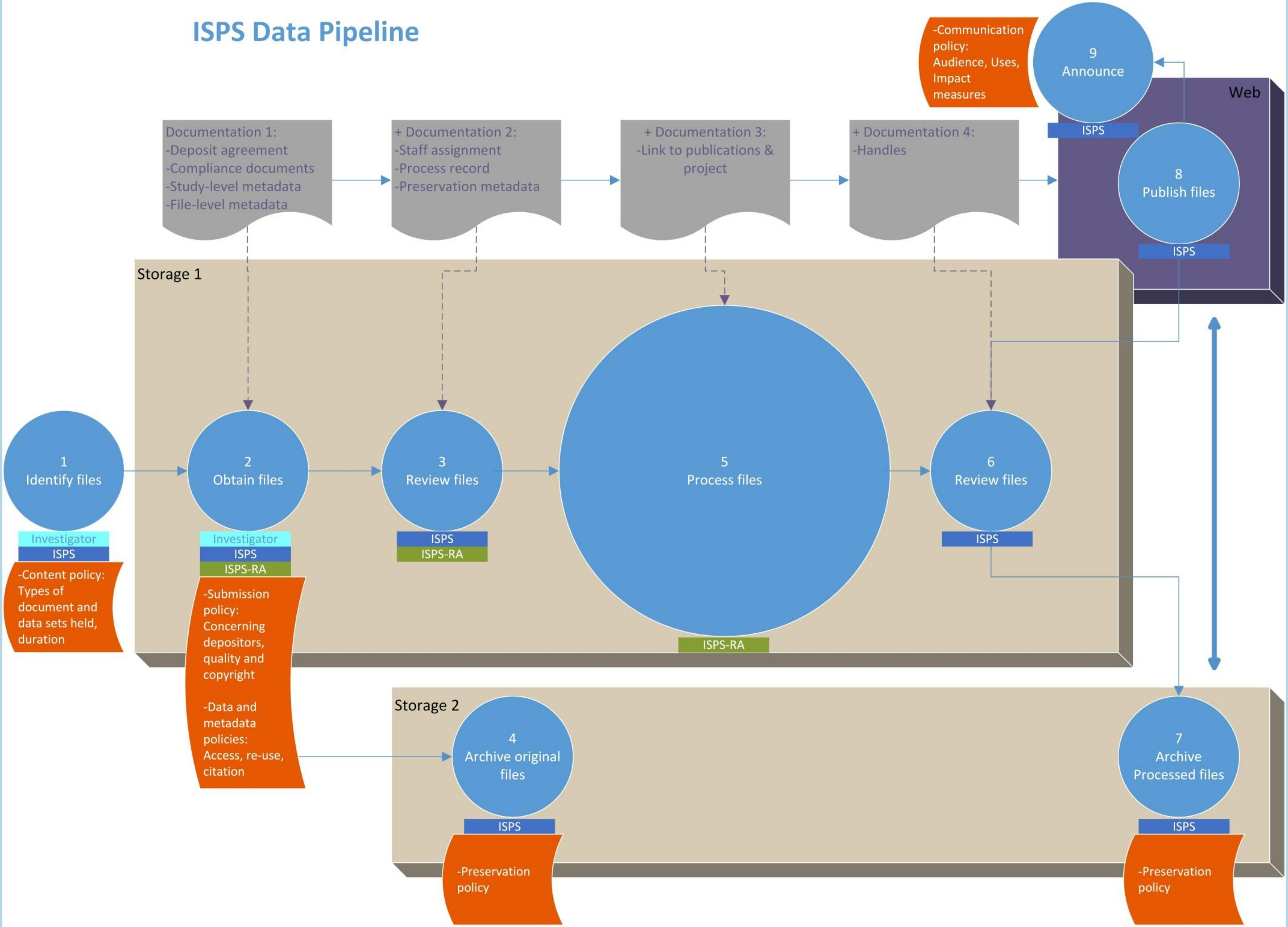
Confirm comprehensive descriptive information for informed reuse including methodology and sampling information * Link to other research products

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 Data Pipeline



Documentation 1:
-Deposit agreement
-Compliance documents
-Study-level metadata
-File-level metadata

+ Documentation 2:
-Staff assignment
-Process record
-Preservation metadata

+ Documentation 3:
-Link to publications & project

+ Documentation 4:
-Handles

Storage 1

1 Identify files
Investigator
ISPS
-Content policy: Types of document and data sets held, duration

2 Obtain files
Investigator
ISPS
ISPS-RA
-Submission policy: Concerning depositors, quality and copyright

-Data and metadata policies: Access, re-use, citation

3 Review files
ISPS
ISPS-RA

5 Process files
ISPS-RA

6 Review files
ISPS

Storage 2

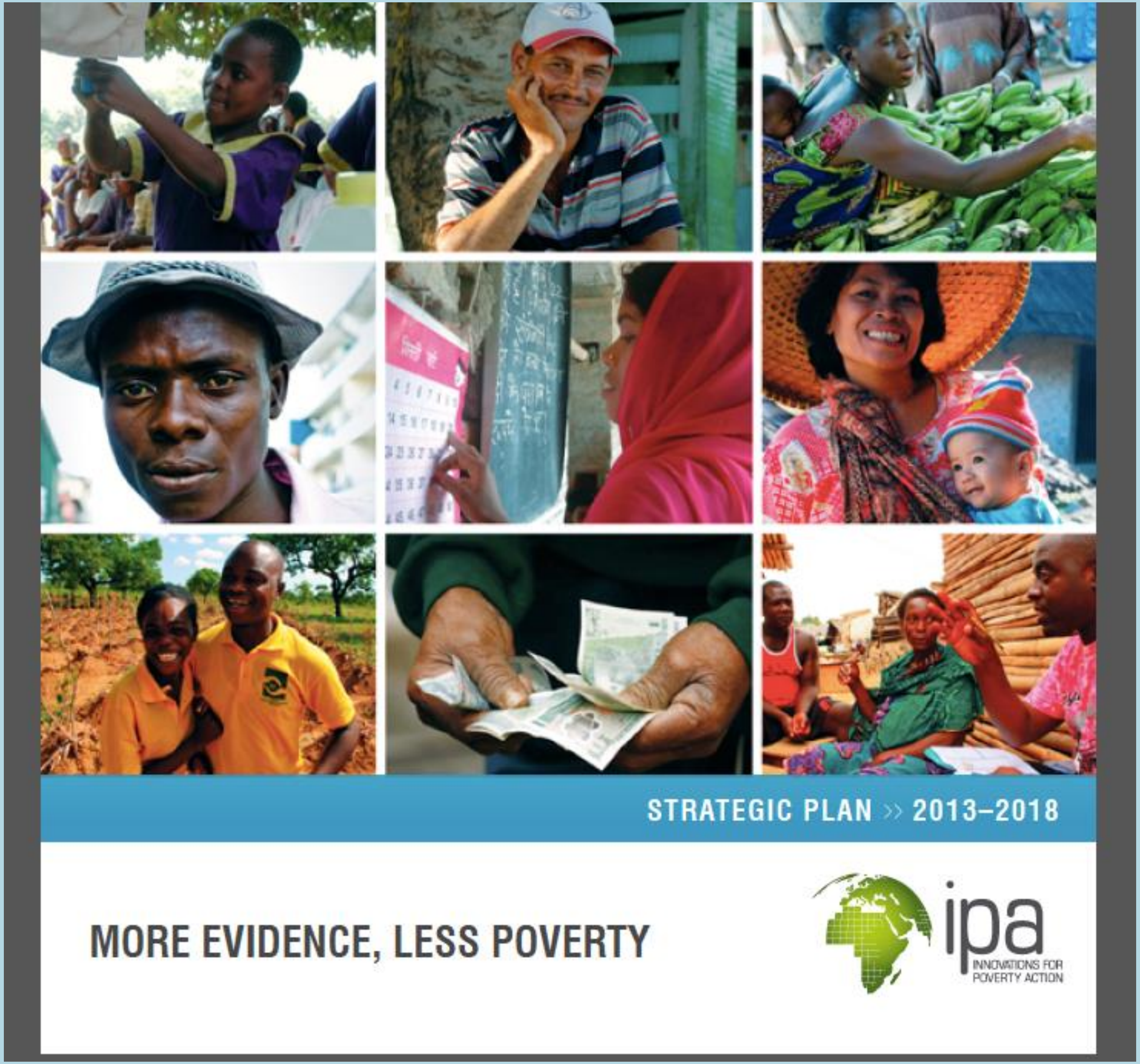
4 Archive original files
ISPS
-Preservation policy

7 Archive Processed files
ISPS
-Preservation policy

Web

8 Publish files
ISPS

9 Announce
ISPS
-Communication policy: Audience, Uses, Impact measures



STRATEGIC PLAN » 2013–2018

MORE EVIDENCE, LESS POVERTY

ipa
INNOVATIONS FOR
POVERTY ACTION

Two Research Organizations

The two organizations have in common…

- Similar content: Data from randomized controlled trials in the social sciences
- Similar approach to data sharing and preservation: Focus on replication, review data and code pre-publication

Cross-fertilization…

- Build on ISPS Data Archive curation standards and practices
- Maintain key aspects of ISPS UI such as linked publications, data, and code
- Build on IPA ability to prepare data earlier in the lifecycle (e.g., pre analysis)
- Allow IPA network to access software from distributed research sites

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

Curator software: Making data usable

A software platform that leverages the DDI Lifecycle and structures the curation workflow, including checking data for confidentiality and completeness, creating preservation formats, and reviewing and verifying code.

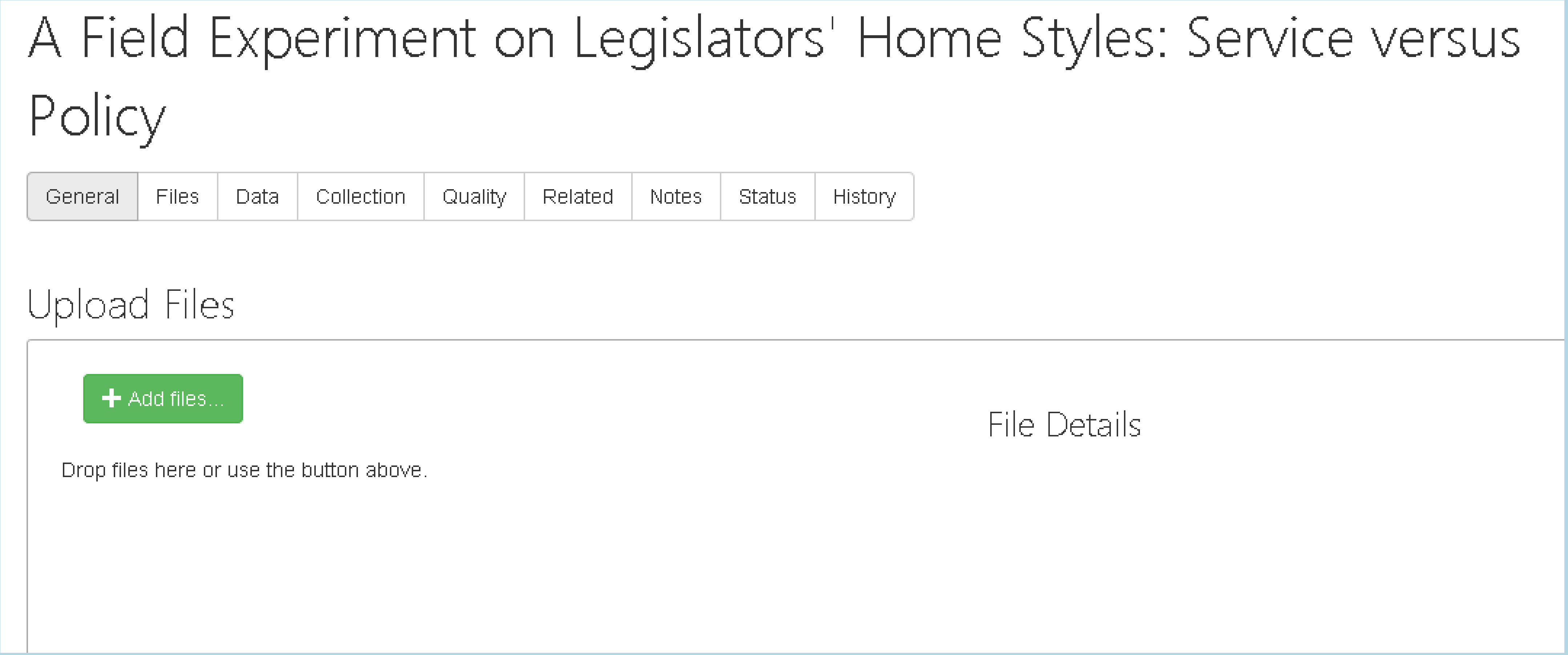
colectica

Experts in social science metadata

Involved in DDI development

Features

- Web-based
- Built on DDI 3.2
- Open Source
- Builds on Existing Tools



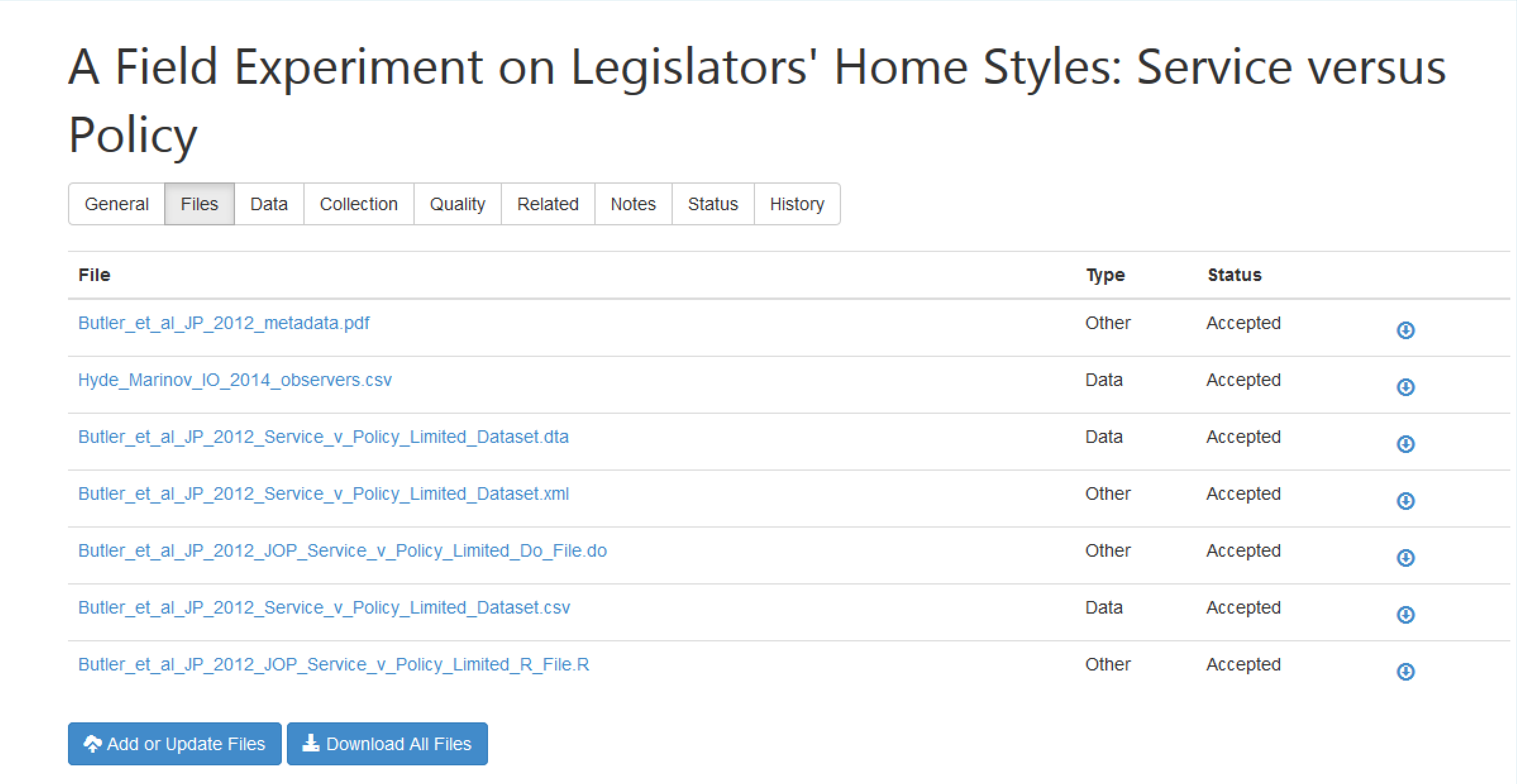
A Field Experiment on Legislators' Home Styles: Service versus Policy

Upload Files

Drop files here or use the button above.

File Details

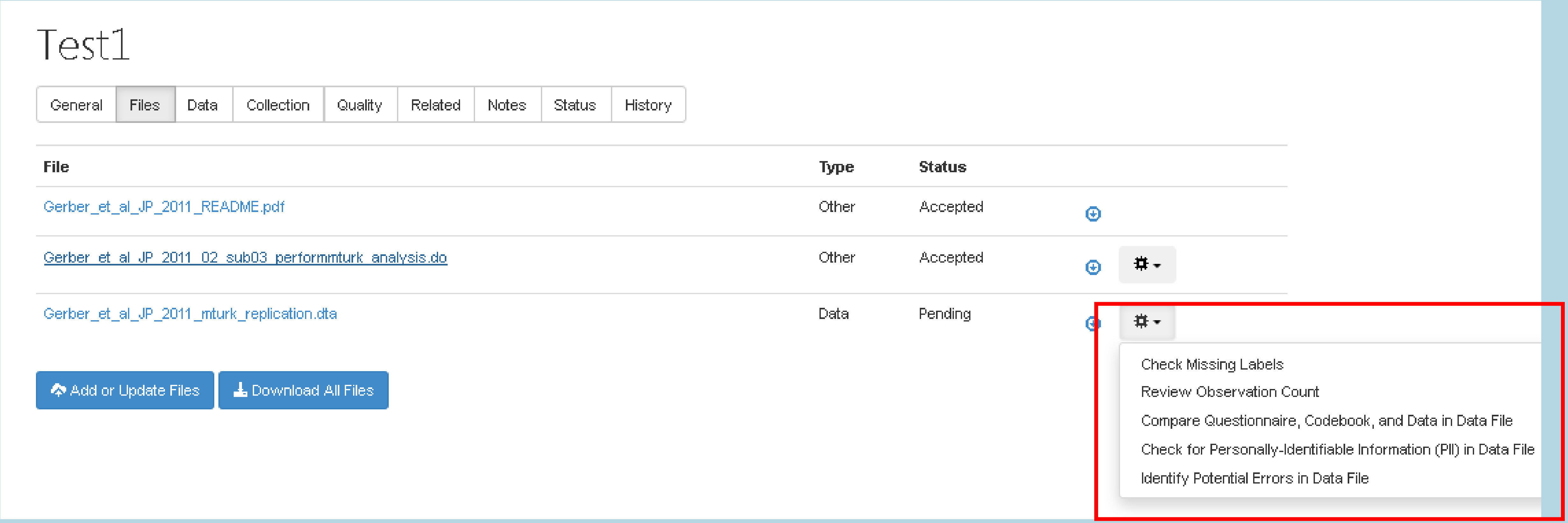
Colectica curator software for ISPS – IPA research data repository (beta)



A Field Experiment on Legislators' Home Styles: Service versus Policy

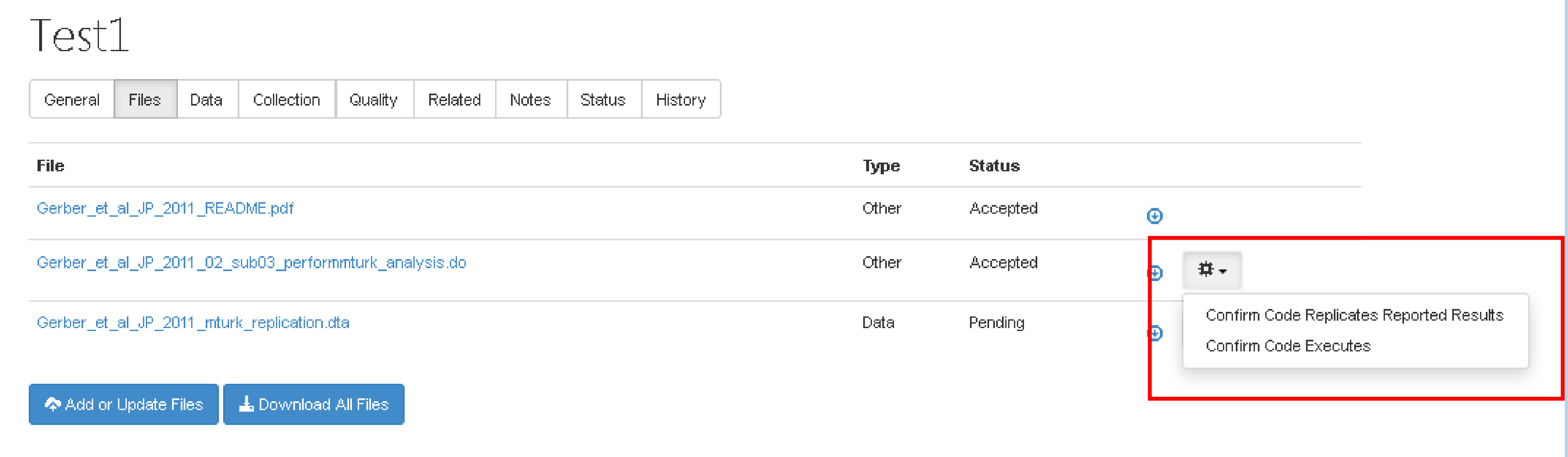
File	Type	Status
Butler_et_al_JP_2012_metadata.pdf	Other	Accepted
Hyde_Marinov_IO_2014_observers.csv	Data	Accepted
Butler_et_al_JP_2012_Service_v_Policy_Limited_Dataset.dta	Data	Accepted
Butler_et_al_JP_2012_Service_v_Policy_Limited_Dataset.xml	Other	Accepted
Butler_et_al_JP_2012_JOP_Service_v_Policy_Limited_Do_File.do	Other	Accepted
Butler_et_al_JP_2012_Service_v_Policy_Limited_Dataset.csv	Data	Accepted
Butler_et_al_JP_2012_JOP_Service_v_Policy_Limited_R_File.R	Other	Accepted

Colectica curator software for ISPS – IPA research data repository (beta)


Test1
General
Files
Data
Collection
Quality
Related
Notes
Status
History
Add or Update Files
Download All Files
Check Missing Labels
Review Observation Count
Compare Questionnaire, Codebook, and Data in Data File
Check for Personally-Identifiable Information (PII) in Data File
Identify Potential Errors in Data File

File	Type	Status
Gerber_et_al_JP_2011_README.pdf	Other	Accepted
Gerber_et_al_JP_2011_02_sub03_performmturk_analysis.do	Other	Accepted
Gerber_et_al_JP_2011_mturk_replication.dta	Data	Pending

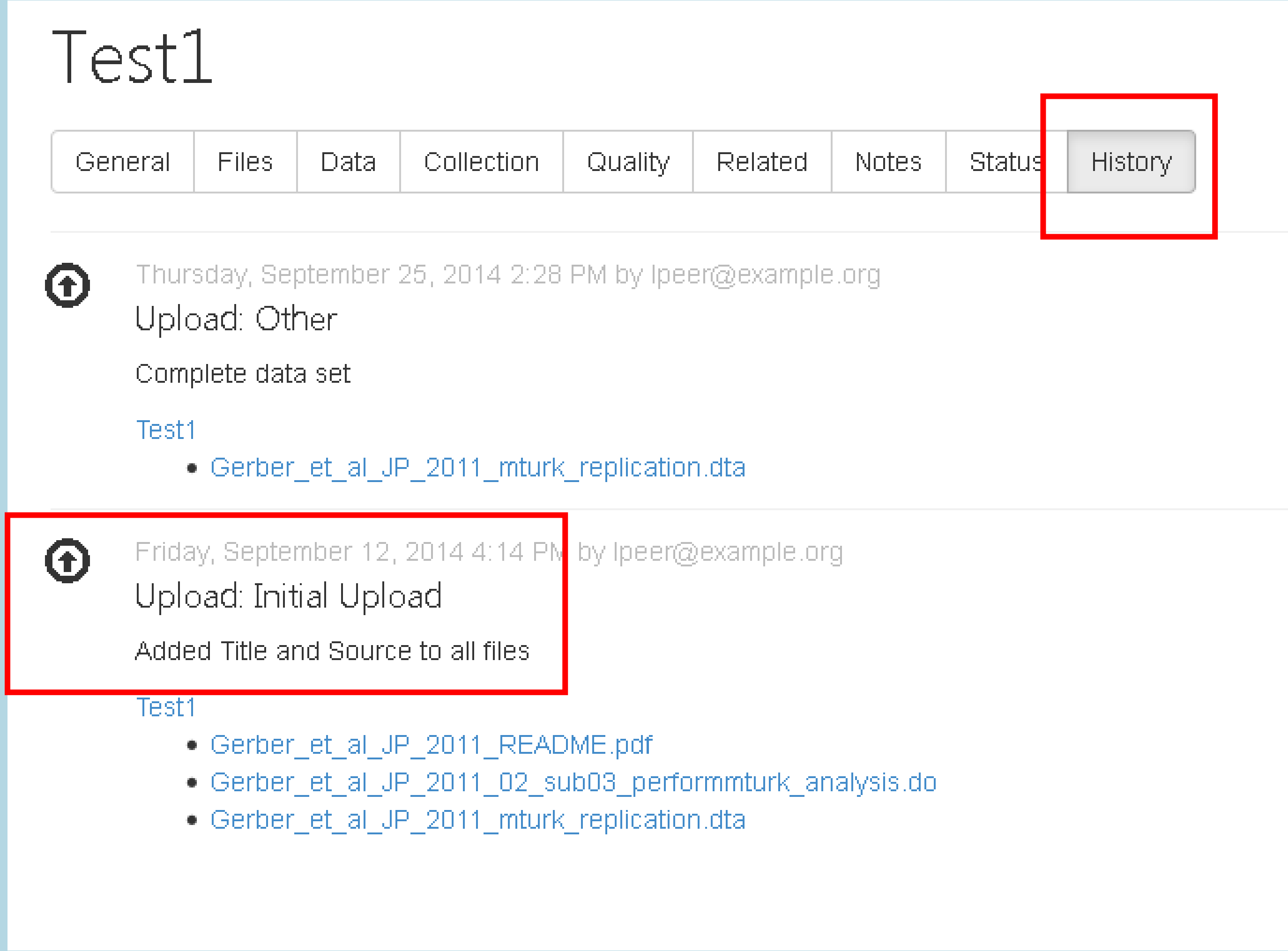
Colectica curator software for ISPS – IPA research data repository (beta)



Test1

File	Type	Status
Gerber_et_al_JP_2011_README.pdf	Other	Accepted
Gerber_et_al_JP_2011_02_sub03_performmturk_analysis.do	Other	Accepted
Gerber_et_al_JP_2011_mturk_replication.dta	Data	Pending

Colectica curator software for ISPS – IPA research data repository (beta)



Test1

Thursday, September 25, 2014 2:28 PM by lpeer@example.org

Upload: Other

Complete data set

Test1

- Gerber_et_al_JP_2011_mturk_replication.dta

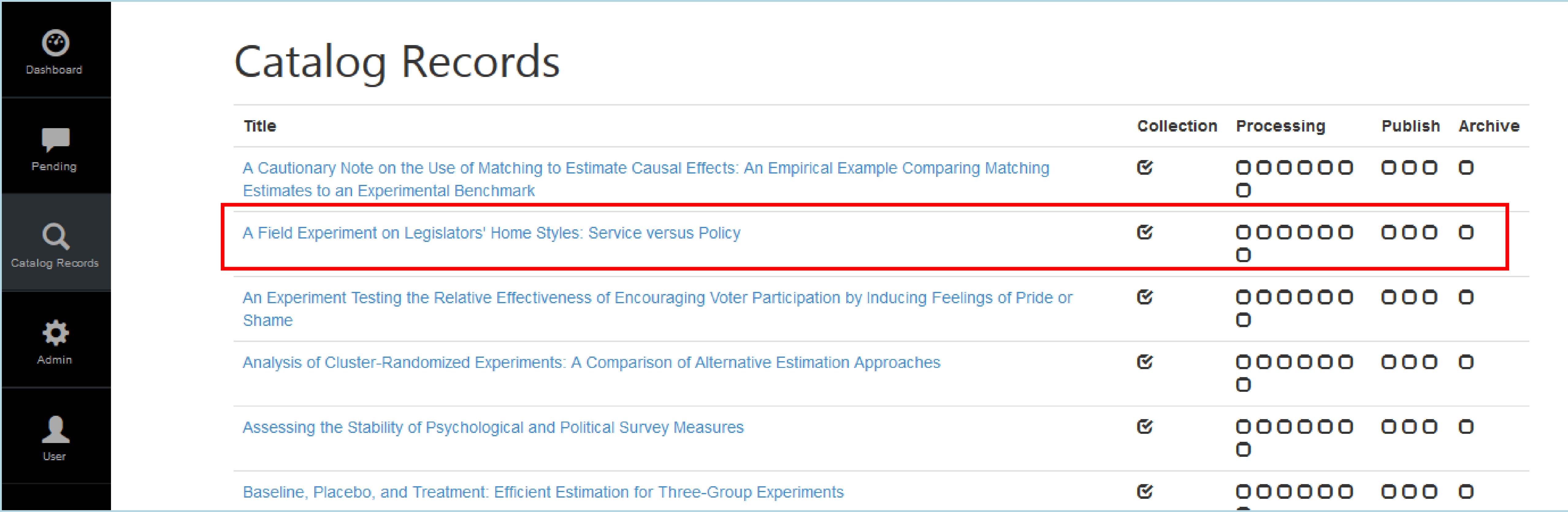
Friday, September 12, 2014 4:14 PM by lpeer@example.org

Upload: Initial Upload

Added Title and Source to all files

Test1

- Gerber_et_al_JP_2011_README.pdf
- Gerber_et_al_JP_2011_02_sub03_performmturk_analysis.do
- Gerber_et_al_JP_2011_mturk_replication.dta

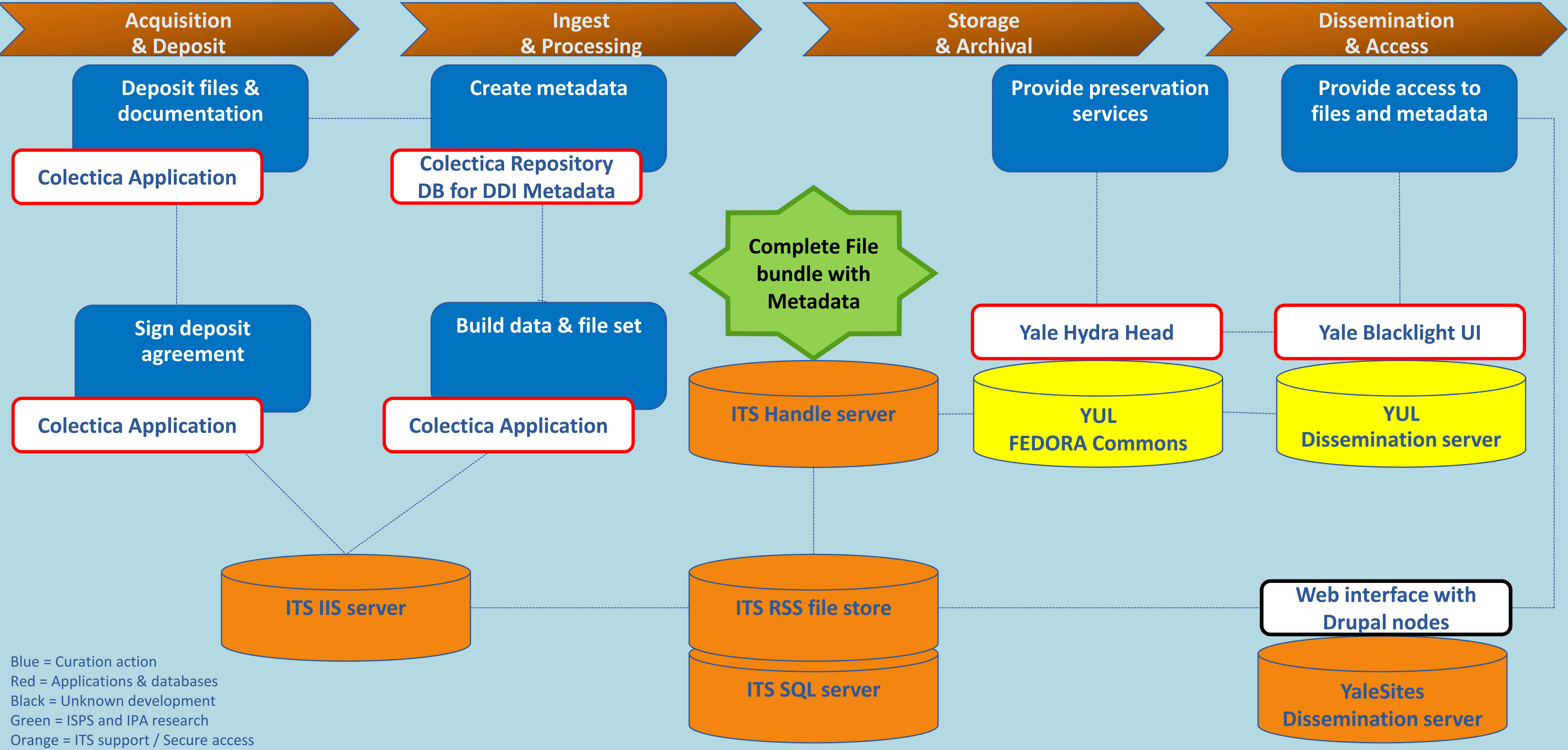


Catalog Records

Title	Collection	Processing	Publish	Archive
A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark	☑	□□□□□□□	□□□	□
A Field Experiment on Legislators' Home Styles: Service versus Policy	☑	□□□□□□□	□□□	□
An Experiment Testing the Relative Effectiveness of Encouraging Voter Participation by Inducing Feelings of Pride or Shame	☑	□□□□□□□	□□□	□
Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches	☑	□□□□□□□	□□□	□
Assessing the Stability of Psychological and Political Survey Measures	☑	□□□□□□□	□□□	□
Baseline, Placebo, and Treatment: Efficient Estimation for Three-Group Experiments	☑	□□□□□□	□□□	□

Colectica curator software for ISPS – IPA research data repository (beta)

Data curation



Acquisition & Deposit

Ingest & Processing

Storage & Archival

Dissemination & Access

Deposit files & documentation

Colectica Application

Sign deposit agreement

Colectica Application

Create metadata

Colectica Repository DB for DDI Metadata

Build data & file set

Colectica Application

Complete File bundle with Metadata

ITS Handle server

Provide preservation services

Yale Hydra Head

YUL FEDORA Commons

Provide access to files and metadata

Yale Blacklight UI

YUL Dissemination server

ITS IIS server

ITS RSS file store

ITS SQL server

Web interface with Drupal nodes

YaleSites Dissemination server

Blue = Curation action
Red = Applications & databases
Black = Unknown development
Green = ISPS and IPA research
Orange = ITS support / Secure access
Yellow = YUL IT support / Secure access

Technical components & support at Yale

ITS

Hardware	Windows Server (VM), 32GB RAM minimum (8 Cores), 100GB local disk for OS, applications and swap files
Software	Colectica suite of tools, statistical software, integrated APIs
Storage	RSS start at 500GB, read/write/no-execute access to one or more directories
Application hosting	WCF application and ASP.NET MVC web application on IIS, plus a SQL Server database (10GB), a Windows Service
Security	Federated identification

Library

Long-term preservation	Fedora Commons / Hydra
Discovery	Blacklight
Persistent links	Handle service (ODAI)

(Target) Timeline

Project Kickoff – February 2014

Development Plan – March to April

Design + Base Platform and Basic Workflow development – May to October

Full Workflow Development – November to December

Ongoing development and maintenance – January 2015+

Thank you!

limor.peer@yale.edu

@l_peer

In collaboration with:

Innovations for Poverty Action: Niall Keleher, Stephanie Wykstra

Digital Lifecycle Research and consulting: Ann Green

Colectica software company: Jeremy Iverson, Dan Smith

Yale ITS Academic IT / Research Services: Kiran Keshav, Themba Flowers, Paul Gluhosky

Yale Library IT: Michael Dula, Mike Friscia, Eric James

Yale Library CSSSI: Michelle Hudson, Jill Parchuck

and Yale ODAI and Office of General Counsel

2014 YALE TECHNOLOGY SUMMIT

Friday, October 31, 2014
Evans Hall, Yale School of Management

#YaleTechSummit2014

Presented by Yale Information Technology Services