ICPSR

American Economic Association (AEA) Data & Code Repository at openICPSR

Maggie Levenstein
Lars Vilhuber (Cornell University and American Economic Association)
Chelsea Goforth
Jared Lyle

9 June 2020

ICPSR INSTITUTE FOR SOCIAL RESEARCH UNIVERSITY OF MICHIGAN

NATIONAL MEDAL for Museum and Library Service

OpenICPSR and reproducibility

AEA data sharing policy 2005

- 2013: Herndon, Ash & Pollin on Reinhart & Rogoff 2010
- 2016 *Science*

OpenICPSR

- Initiated in 2013, following publication of OSTP memo and multiple meetings with editors of social science journals
- Economic History Association, American Education Research Association

ICPSR

AEA Data Editor Workflow and Methods

Quick Outline

- Workflow from the user perspective
- Workflow from the Data Editor's perspective
- Desired outcomes
- Tasks
- Skills required
- Training
- Scale

Required outcomes

- Archival record of the scholarly materials
  - Implies versioning
- Permanent identifier (PID)
  - DOI
- Linkage to and from article
  - Human-visible
  - Metadata (CrossRef + DataCite)
  - Across versions
- Transparent and efficient inspection of materials
  - Understand materials without downloading (multi-100-GB deposits…)
- Proper identification of roles
  - Author, Distributor, Publisher

ICPSR

AMERICAN ECONOMIC ASSOCIATION

CORNELL UNIVERSITY FOUNDED A.D. 1865

ILR School

Desired attributes

- Journal workflow integration
  - Without prescribing a specific repository!
- Pre-publication access for replicators
- Role-based access
  - Ability to delegate certain tasks
- Partial automation
  - Of repeated tasks
  - Of communicating between systems
  - Of computational checks
  - Of metadata and layout checks
  - Of downloads

ICPSR

AMERICAN ECONOMIC ASSOCIATION
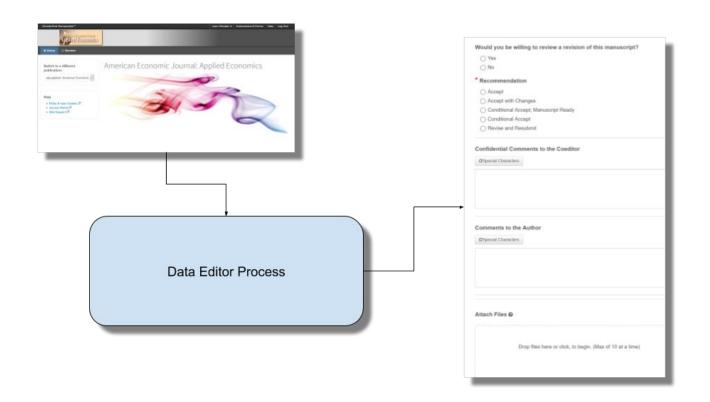
ILR School

Workflow from user perspective

Workflow from Data Editor perspective

Real Workflow...

JIRA ties it together…
with lots of manual sweat

(1) JIRA identifier
(2) Manuscript Central / ScholarOne identifier
(3) openICPSR project number (when available)
(4) Direct link to deposit (even if not on openICPSR)

Tasks

Metadata

- List datasets
- Check data citations
- Check "data availability statements"
- Check metadata on openICPSR
- Check structure on openICPSR

Data and Code

- Download code
- Download data (if separate)
- Assess code completeness
    - Data preparation code
    - Analysis code
- Run code
- Compare results

Write report

Approve report

Submit report

ICPSR

AMERICAN ECONOMIC ASSOCIATION

CORNELL UNIVERSITY FOUNDED A.D. 1865

ILR School

Metadata on data deposit

Required elements

Title
(Suggested: "*Data and code for: (NAME OF PAPER)*")
"Principal Investigators"
(=Authors; these need not be in the same order)
Abstract
JEL Classification
Manuscript Number

Structure:
- README in TXT, PDF
- No ZIP files
- No empty directories

Highly recommended

Subject Terms (e.g., "Machine Learning",
"Randomized Control Trial", "Nudges", …)
Geographic coverage (e.g, "United States", "Florida,
U.S.", "Indonesia", …)
Time period(s) (e.g., "1982-2008")

Suggested

Collection date(s)
Universe (e.g., "All households in Canada",
"Manufacturing establishments in Indonesia", …)
Data Type(s)
Data Source
Units of Observation

ICPSR

AMERICAN ECONOMIC ASSOCIATION

ILR School

Skills

- Ability to read through non-standardized documents and structure
    - Data sources do not have a fixed location
    - Data citations are nearly absent
- Ability to run code in "minimally invasive" mode
    - Not code consultants!
    - Reproduction runs needs to be … reproducible!
- Multiple computer languages
    - Stata
    - Matlab
    - R, Python, Julia, etc.
- Diverse computing environments

Training

Typical team member:

- **Undergraduate** in social sciences, computer science, statistics, engineering
- Typically **junior**, sometimes **sophomore** when they join

Training

- Intensive start session ("bootcamp")
- Continuous training over 5-7 weeks
- On-the-job training after that

Scale

As of Monday, June 8 2020

- 796 assessments for all 8 journals
- Currently 18 RAs (+ 1 GRA)
- About dozen reproducibility checks with confidential data
    - Mostly third parties

Challenges

- Scale
- Recurring training
  - Of RAs
  - Of Authors!
- Coordination across multiple econ journals
  - Addresses one of the training issues
- Software not available
  - Very few cases
- Integration/ streamlining of workflow

ICPSR

AMERICAN
ECONOMIC
ASSOCIATION

ILR School

AEA Data & Code Repository Walkthrough

AEA Data & Code Repository

Create New Project

Project Workspace

Upload Files

Upload Files

Metadata Fields

Metadata Fields

Metadata Fields

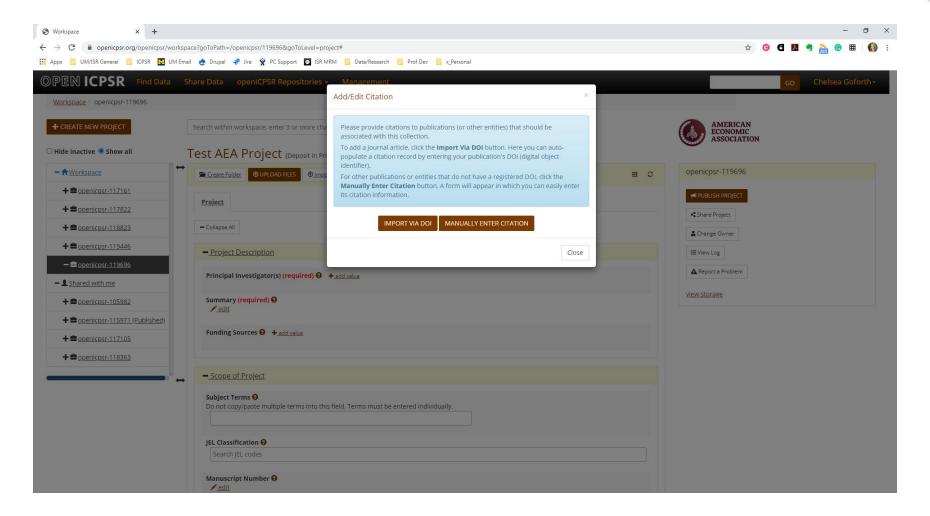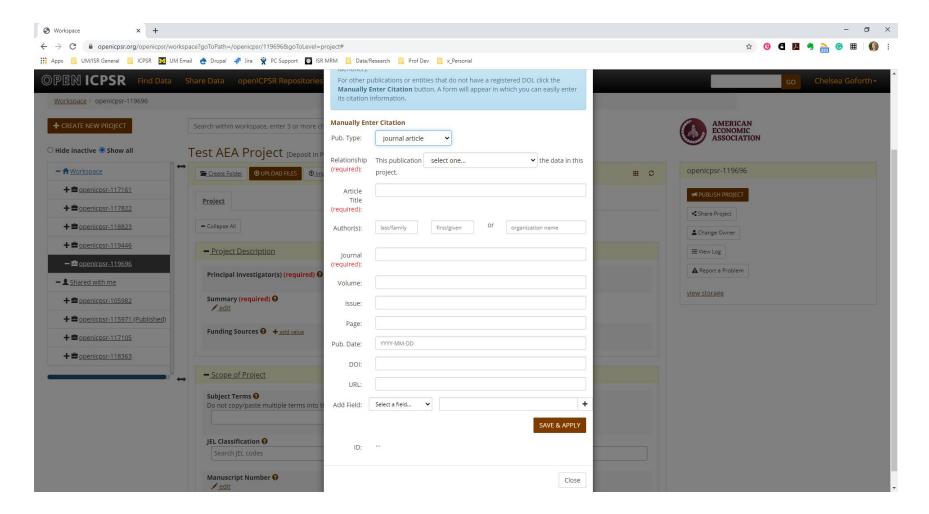# Metadata Fields

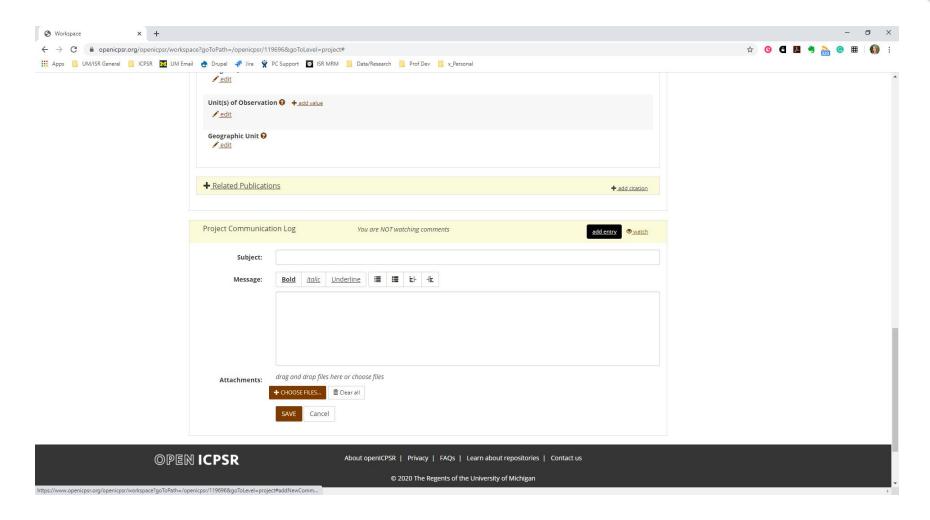# Metadata Fields

Metadata Fields

Related Publications

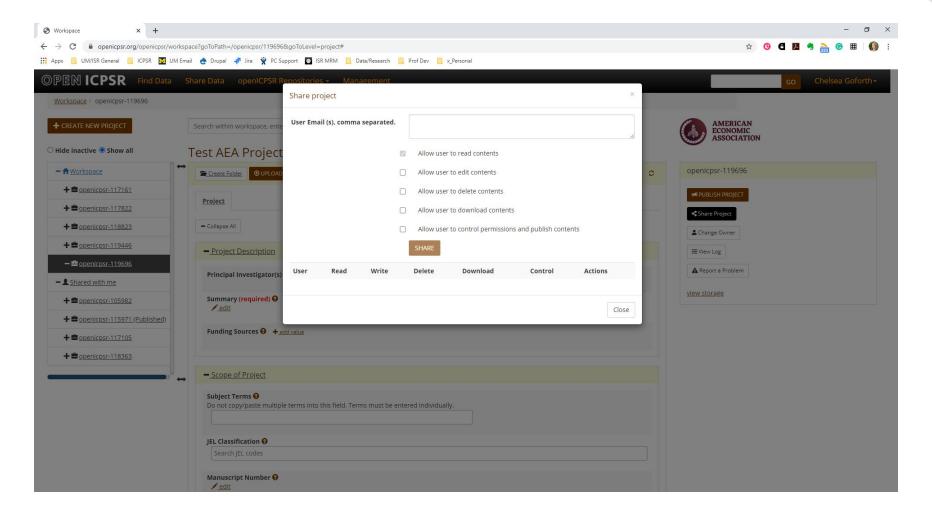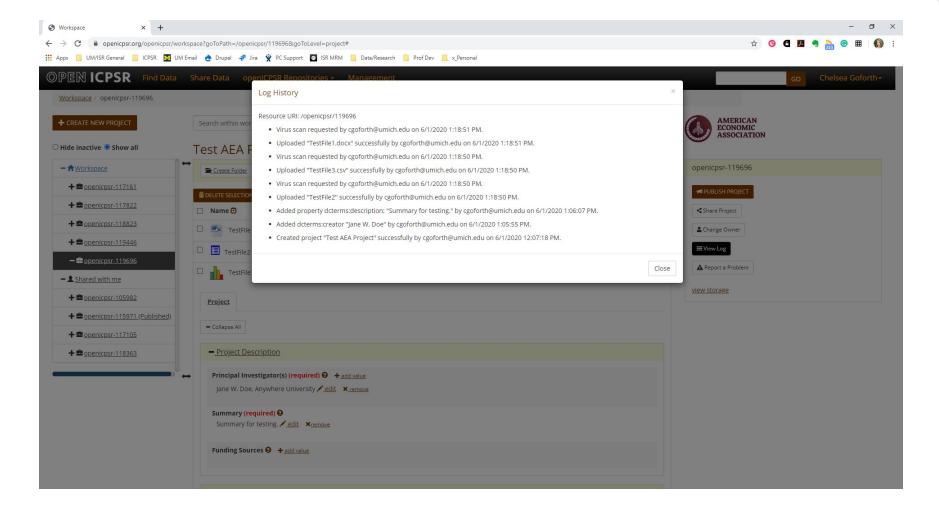# Related Publication(s)

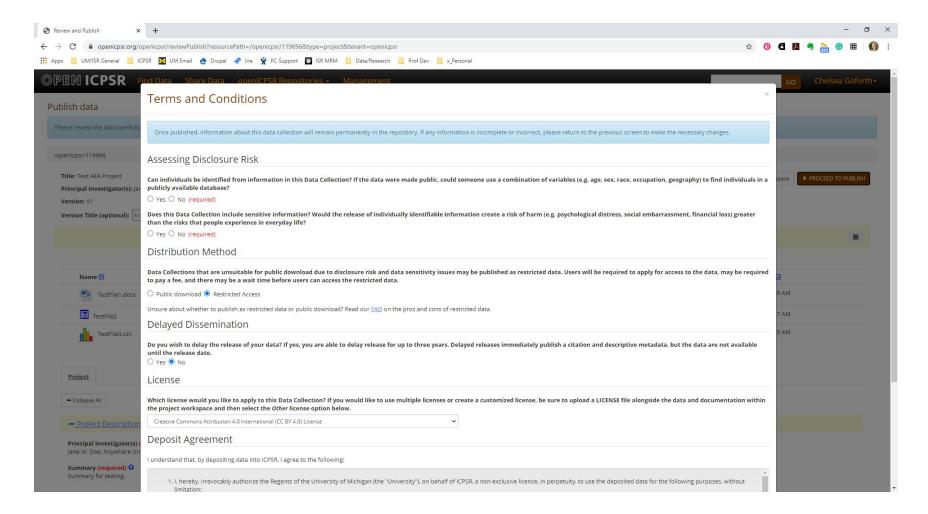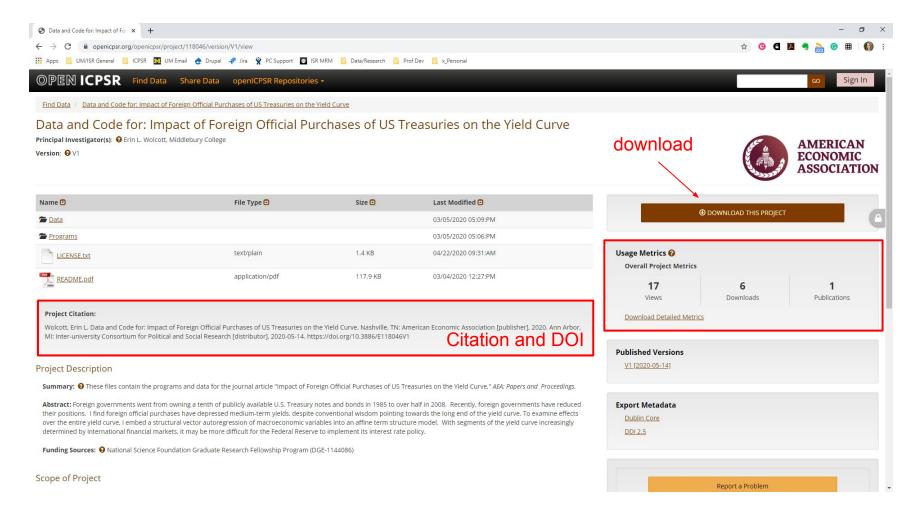# Project Communication Log

Share Project

View Log

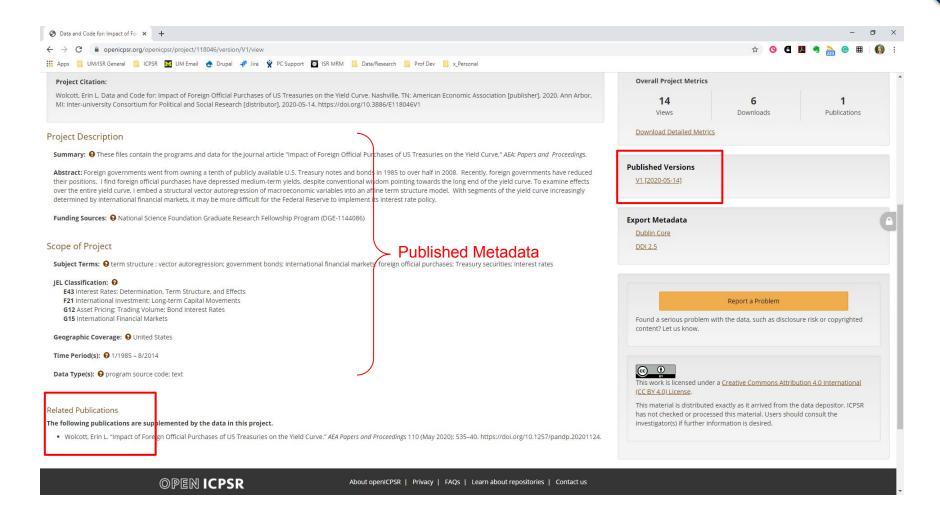# Publish - Terms & Conditions

Published Project

Citation and DOI

Wolcott, Erin L. Data and Code for: Impact of Foreign Official Purchases of US Treasuries on the Yield Curve. Nashville, TN: American Economic Association [publisher], 2020. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-05-14. https://doi.org/10.3886/E118046V1
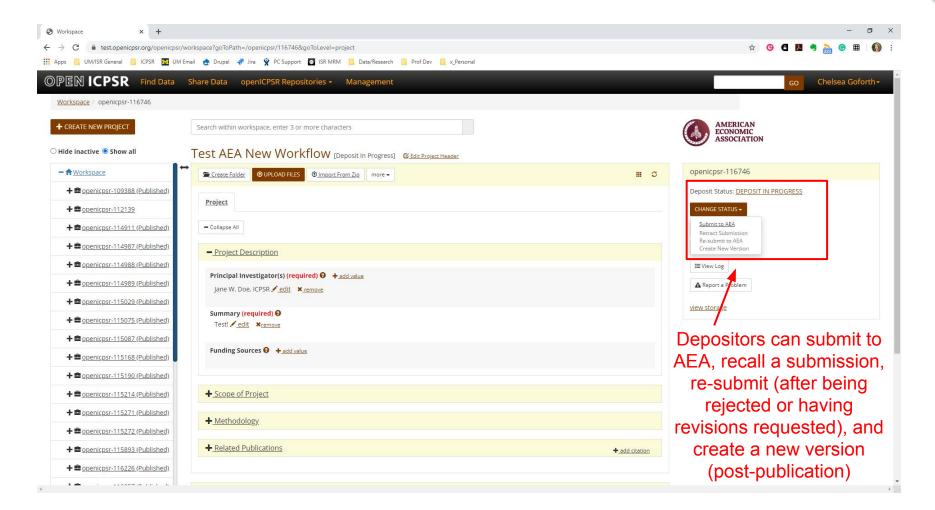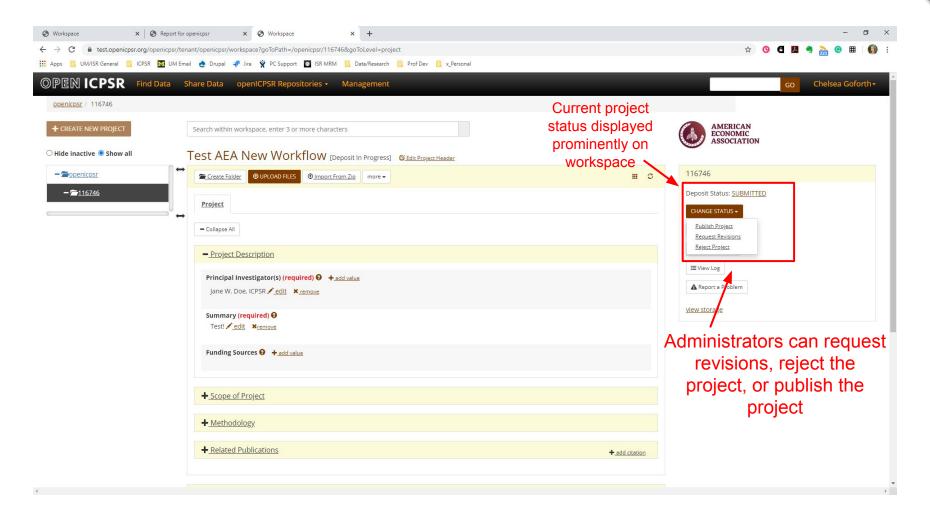
ICPSR

Published Project

Sneak Peek Into New Workflow

Depositors can submit to AEA, recall a submission, re-submit (after being rejected or having revisions requested), and create a new version (post-publication)

Sneak Peek Into New Workflow

Current project status displayed prominently on workspace

Administrators can request revisions, reject the project, or publish the project

Sneak Peek Into New Workflow

Status changes, including any comments, are reflected in the project communication log, which also triggers an email notification

ICPSR

Questions?

maggiel@umich.edu
lars.vilhuber@cornell.edu
cgoforth@umich.edu
lyle@umich.edu