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
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
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
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
Workflow from user perspective
Workflow from Data Editor perspective

Data Editor Process

Would you be willing to review a revision of this manuscript?
- Yes
- No

Recommendation
- Accept
- Accept with Changes
- Conditional Accept, Manuscript Ready
- Conditional Accept
- Revise and Resubmit

Confidential Comments to the Editor

Comments to the Author

Attach Files
- Drop files here or click to begin. (Max of 10 at a time)
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
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**

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**

Structure:

- README in TXT, PDF
- No ZIP files
- No empty directories
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
AEA Data & Code Repository Walkthrough
AEA Data & Code Repository

Depositing Data in the AEA Data and Code Repository

The American Economic Association journals require authors to deposit data and materials with a community-recognized or general repositories. The AEA Data and Code Repository at ICPSR serves that purpose. Please see the AEA’s Data and Code Availability Policy and data citation guidance at the Sample References page for more details. Authors are required to include a citation pointing to the deposit in the reference section of the final version of the article sent to the AEA. The openICPSR repository automatically generates a citation when the data are “published.”

Deposits should include all data, annotated program code, command files, and documentation that is needed to replicate the findings from the authors’ submitted article.

- Data should be comprehensively documented (see ICPSR’s Guide to Social Science Data Preparation and Archiving, 5th Edition for guidance). The author is responsible for removing identifying information from the data to protect confidentiality. Neither the AEA nor ICPSR review submissions for disclosure risk.
- Program code and command files should be annotated to facilitate replication and ensure clear correspondence between code and figures, tables, and analyses in the published article.
- Authors retain ownership and copyright to the data and code. Authors are required to affirm that they have the right to publish and redistribute the material. However, ICPSR requires a license for distribution of data.
- An open license is required by the AEA in order to allow others to re-use the data and code in particular for replication. Authors can select from several license options, including CC-BY 4.0 for data and Modified BSD for software and code. If an author would like to use multiple licenses or create a customized license, she should select the “other” license option and upload a LICENSE file alongside the data and documentation.

By depositing in the AEA Data and Code Repository, the depositors allow the AEA staff to add keywords and other metadata which are important for proper indexing in linking. Any other changes are subject to the license chosen for the materials.

View more extensive (unofficial) guidance if.

Start Your Deposit
Create New Project
Project Workspace

- Upload Files/Import From Zip
- Enter Project Metadata
- Project Administration Tools
Upload Files

Please upload files by dragging and dropping files into the designated area or by selecting the "Add files..." button and then navigating to the file(s) of your choice. Note that .ZIP files uploaded through the "Upload Files" button will remain as .ZIP files in your workspace. If you want to extract contents of a .ZIP file to your project workspace, close this window and upload using the "Import From ZIP" button. Please be advised that password protected files are not supported at this time.

ICPSR allows users to share their data up to a 30000 MB limit. Requests to exceed the 30000 MB limit should be sent to ICPSR Acquisitions and should address expected size. 30000 MB space remaining for uploads.
Upload Files

Test AEA Project

- TestFile1.docx
- TestFile2.txt
- TestFile3.csv

Actions

- View
- Edit
- Share
- Permissions
- Settings
- Delete
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

Terms and Conditions

Once published, information about this data collection will remain permanently in the repository. If any information is incomplete or incorrect, please return to the previous screen to make the necessary changes.

Assessing Disclosure Risk

Can individuals be identified from information in this Data Collection? If the data were made public, could someone use a combination of variables (e.g., age, sex, race, occupation, geography) to find individuals in a publicly available database?

- Yes □ No □ (required)

Does this Data Collection include sensitive information? Would the release of individually identifiable information create a risk of harm (e.g., psychological distress, social embarrassment, financial loss) greater than the risks that people experience in everyday life?

- Yes □ No □ (required)

Distribution Method

Data Collections that are unsuitable for public download due to disclosure risk and data sensitivity issues may be published as restricted data. Users will be required to apply for access to the data, may be required to pay a fee, and there may be a wait time before users can access the restricted data.

- Public download □ Restricted Access □

Ensure about whether to publish as restricted data or public download! Read our FAQ on the pros and cons of restricted data.

Delayed Dissemination

Do you wish to delay the release of your data? If yes, you are able to delay release for up to three years. Delayed releases immediately publish a citation and descriptive metadata, but the data are not available until the release date.

- Yes □ No □

License

Which license would you like to apply to this Data Collection? If you would like to use multiple licenses or create a customized license, be sure to upload a LICENSE file alongside the data and documentation within the project workspace and then select the Other license option below.

- Creative Commons Attribution 4.0 International (CC BY 4.0) License □

Deposit Agreement

I understand that, by depositing data into ICPSR, I agree to the following:

1. I hereby irrevocably authorize the Regents of the University of Michigan (the "University") on behalf of ICPSR, a non-exclusive license, in perpetuity, to use the deposited data for the following purposes, without limitation:

   □
Published Project

**Data and Code for: Impact of Foreign Official Purchases of US Treasuries on the Yield Curve**

**Principal Investigator(s):** Erin L. Wolcott, Middlebury College

**Version:** V1

---

**Project Citation:**

**Usage Metrics**
- **Overall Project Metrics**
  - 17 Views
  - 6 Downloads
  - 1 Publications

**Published Versions**
- V1 (2020-05-14)

**Export Metadata**
- **Dublin Core**
- **DOI** 2.3

---

**Project Description**

**Summary:** These files contain the programs and data for the journal article “Impact of Foreign Official Purchases of US Treasuries on the Yield Curve.” AEA Papers and Proceedings.

**Abstract:** Foreign governments went from owning a tenth of publicly available U.S. Treasury notes and bonds in 1985 to over half in 2008. Recently, foreign governments have reduced their positions. I find foreign official purchases have depressed medium-term yields, despite conventional wisdom pointing towards the long end of the yield curve. To examine effects over the entire yield curve, I embed a structural vector autoregression of macroeconomic variables into an affine term structure model. With segments of the yield curve increasingly determined by international financial markets, it may be more difficult for the Federal Reserve to implement its interest rate policy.

**Funding Sources:** National Science Foundation Graduate Research Fellowship Program (DGE-1744066)

---

**Scope of Project**

---

**Download**
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

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

Current project status displayed prominently on workspace.
Sneak Peek Into New Workflow

Status changes, including any comments, are reflected in the project communication log, which also triggers an email notification.
Questions?
maggiel@umich.edu
lars.vilhuber@cornell.edu
cgoforth@umich.edu
lyle@umich.edu