

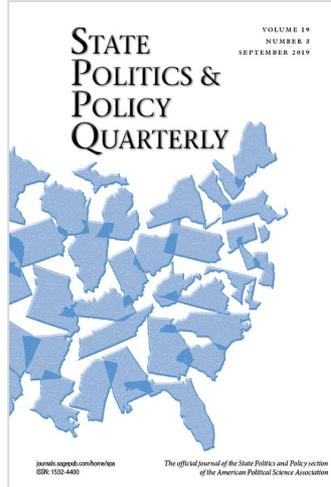
Streamlining data curation and verification workflows

Thu-Mai Christian, Assistant Director for Archives
Data-PASS Pre-APSA Workshop | August 28, 2019

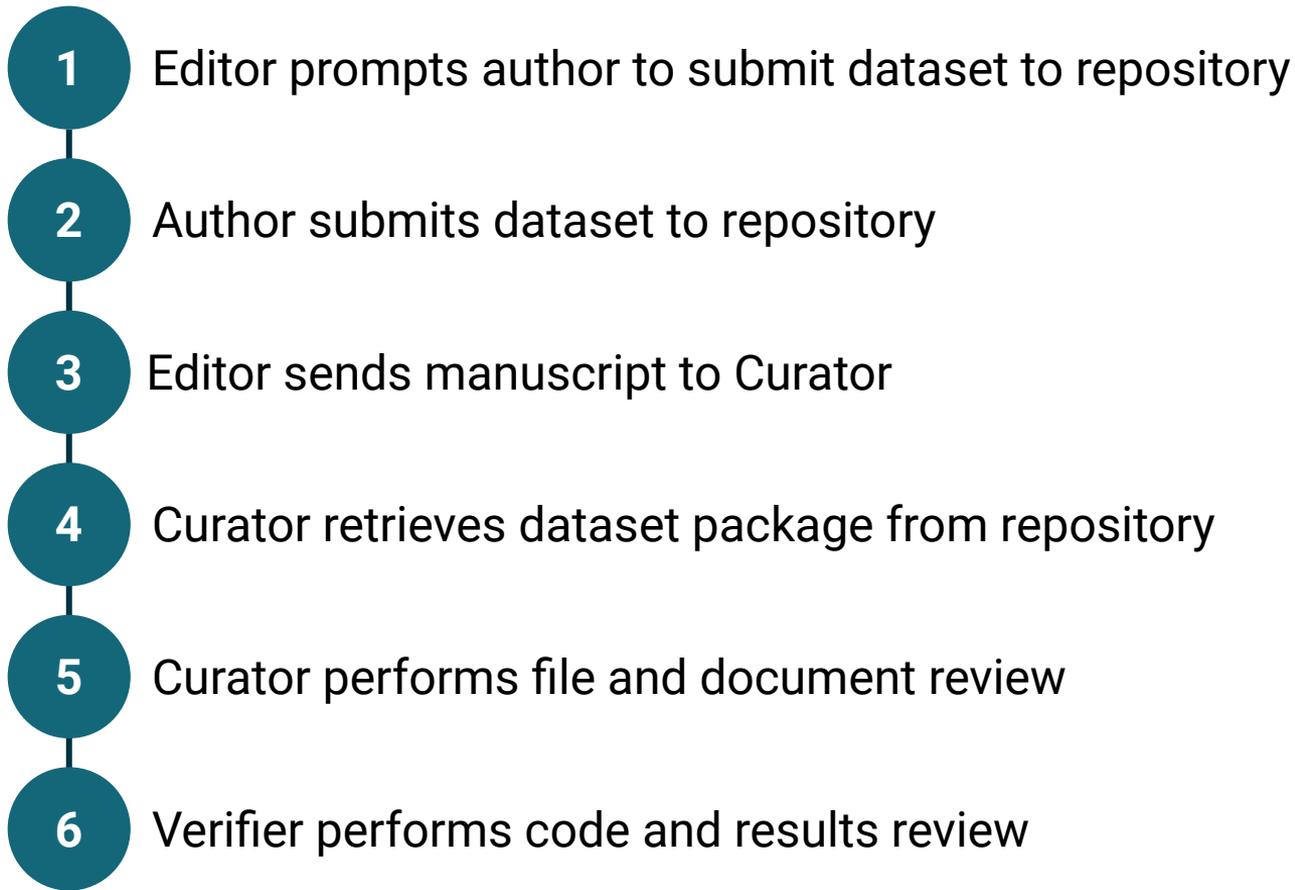
THE ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

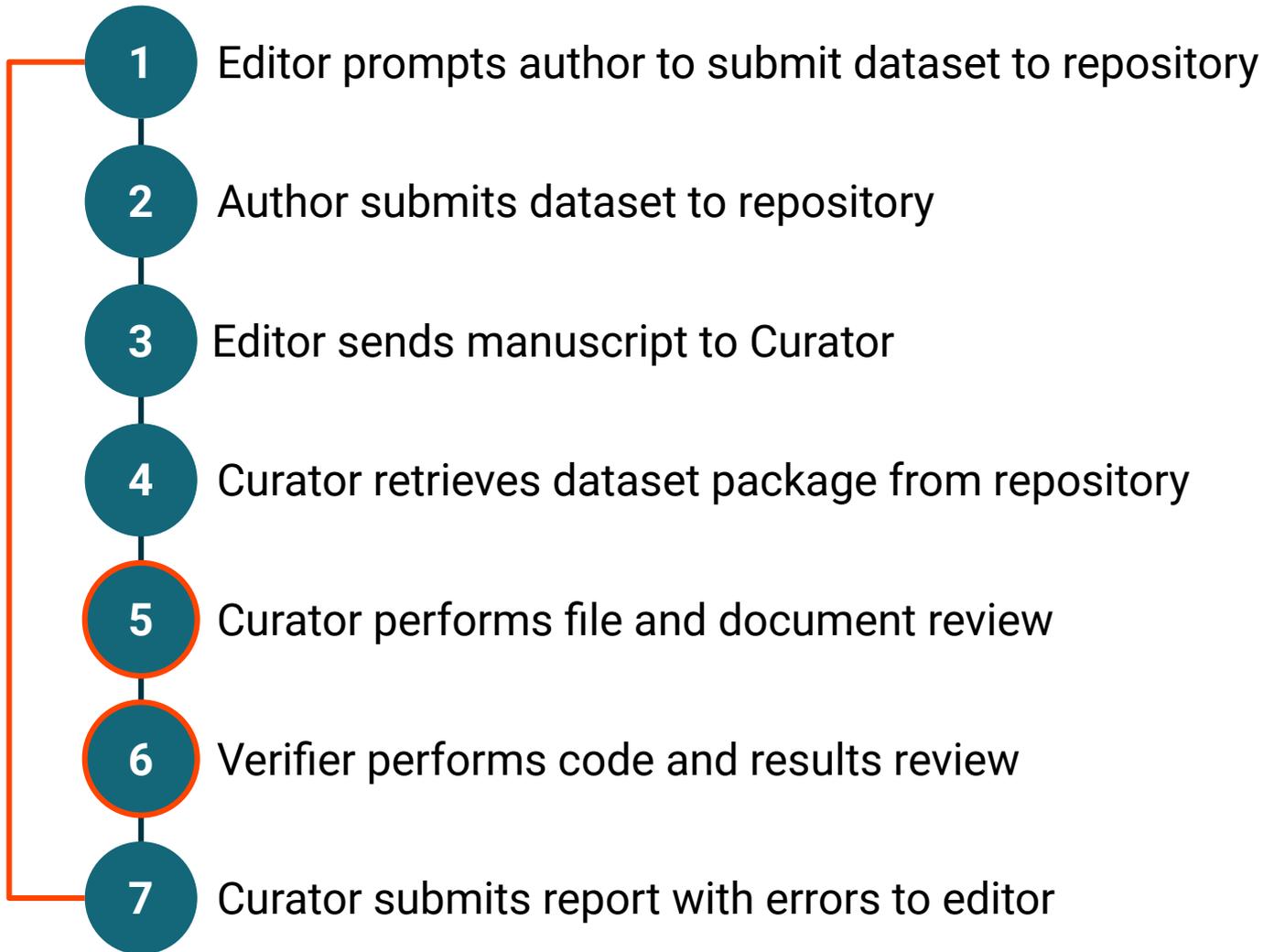


THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



“When the final draft of the manuscript is submitted, the materials will be verified to confirm that they do, in fact, reproduce the analytic results reported in the article.”





1 Editor prompts author to submit dataset to repository

2 Author submits dataset to repository

3 Editor sends manuscript to Curator

4 Curator retrieves dataset package from repository

5 Curator performs file and document review

6 Verifier performs code and results review

- 
- 7 Curator enhances metadata in repository
 - 8 Curator publishes dataset record in repository
 - 9 Curator submits final report with data citation to Editor
 - 10 Curator updates dataset record in repository with article citation

- 7 Curator enhances metadata in repository
- 8 Curator publishes dataset record in repository
- 9 Curator submits final report with data citation to Editor
- 10 Curator updates dataset record in repository with article citation



6 hours per manuscript

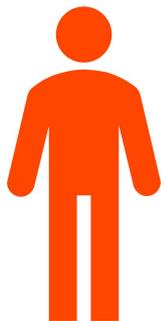
2.11 resubmissions per manuscript



- ✗ Missing or incomplete codebooks and README files
- ✗ Missing data and syntax files
- ✗ Missing citation to original dataset
- ✗ No file descriptions
- ✗ Use of file formats not optimal for preservation



- ✗ Use of absolute file paths
- ✗ Missing package installation scripts
- ✗ Missing seed command
- ✗ Code execution errors
- ✗ Output mismatches
- ✗ Rounding errors



AUTHOR



EDITOR



CURATOR



VERIFIER



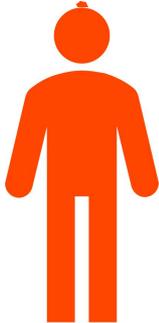
edit data
and code



submit
manuscript



submit
dataset



AUTHOR



EDITOR



CURATOR



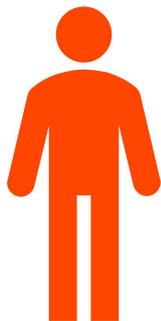
VERIFIER



manage
publication
workflow

correspond
with Curator

receive
repository
notifications



AUTHOR



EDITOR



CURATOR



VERIFIER

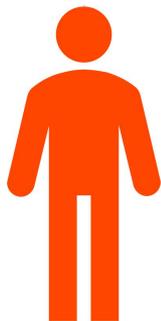


correspond
with Editor
+ Verifier

retrieve
+ publish
dataset

report results
+ track
manuscripts

perform file
review



AUTHOR



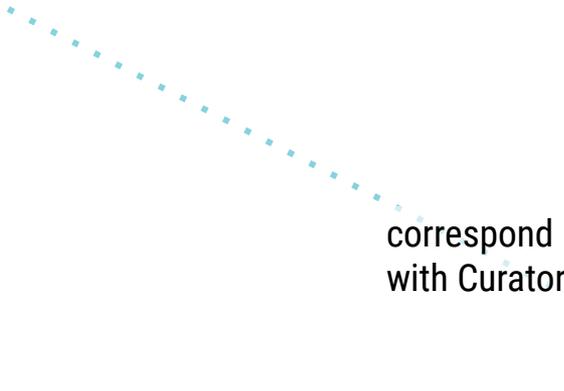
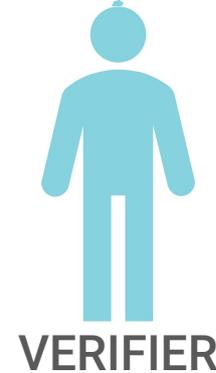
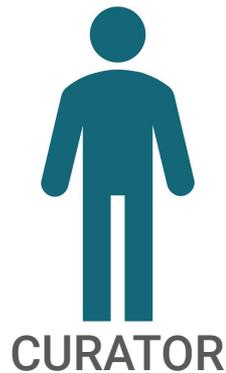
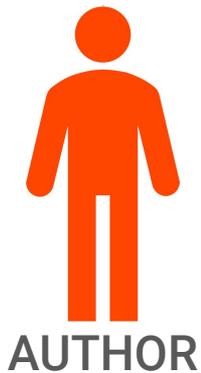
EDITOR



CURATOR

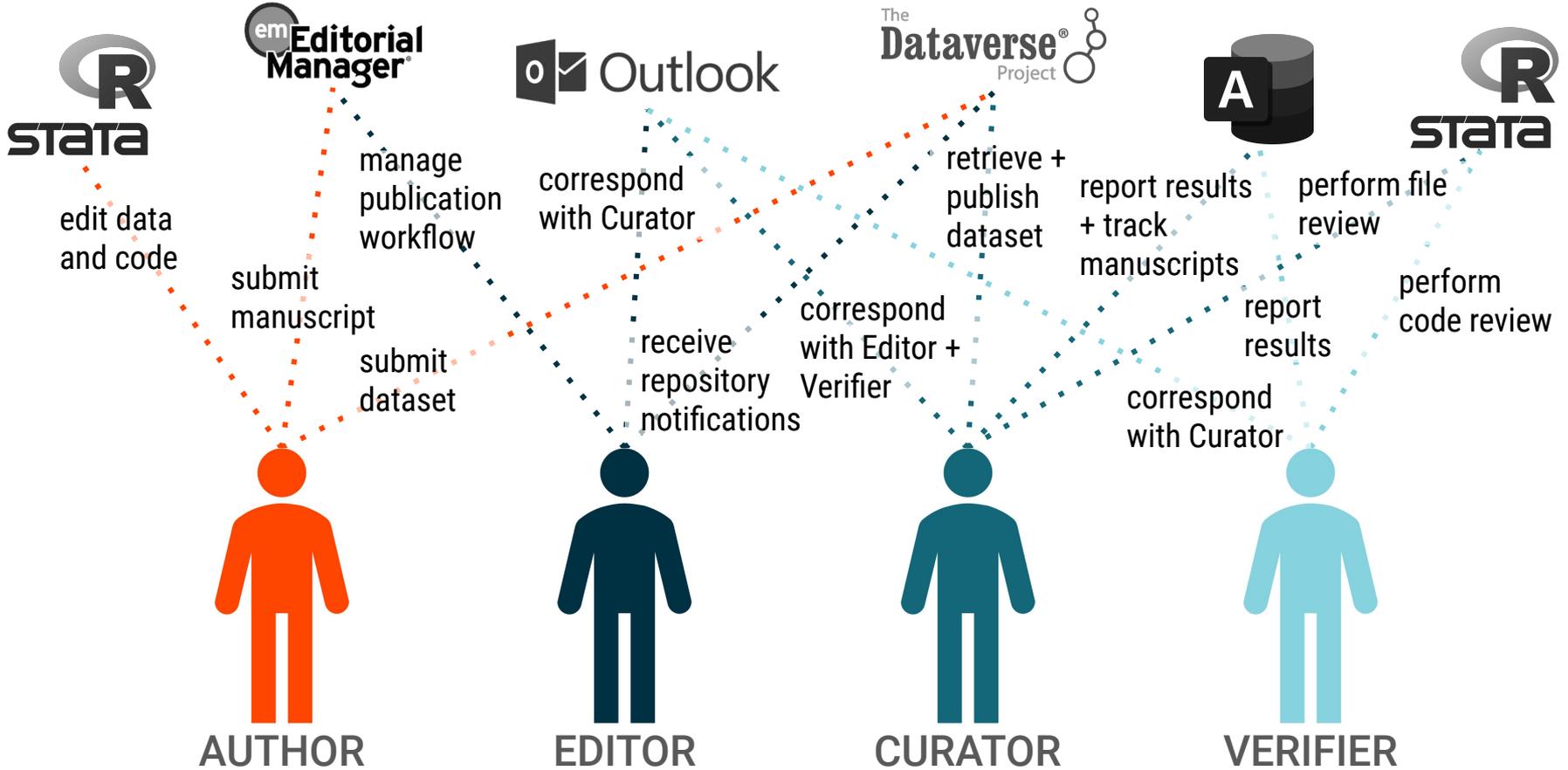


VERIFIER



report results

perform code review





- ★ Policies with clearly articulated requirements
- ★ Comprehensive policy guidance to support compliance
- ★ Training in data management and reproducible research standards and best practices



- ★ Technology integrations to streamline verification workflows
- ★ Automated workflow tracking
- ★ Enabling tools that support data management and reproducible research standards and best practices



- ★ Policies with clearly articulated requirements
- ★ Comprehensive policy guidance to support compliance
- ★ Training in data management and reproducible research standards and best practices

AMERICAN JOURNAL OF POLITICAL SCIENCE
QUANTITATIVE DATA VERIFICATION CHECKLIST
Version 1.2, May 23, 2016

The following checklist can assist authors in preparing necessary exhaustive or applicable to all replication instructions provided in the *AJPS Guidelines for Preparation*.

Supporting Documentation and Information:

- Includes a README file (.txt format) containing and any other important information regarding need to be run, etc.)
- Includes a Codebook (.pdf format) with variable dataset(s) and value labels for categorical variables

Title stata.com
codebook — Describe data contents

Syntax Menu Description Options
Remarks and examples Stored results References Also see

Syntax

codebook [varlist] [if] [in] [, options]

Replication Data for: Donors, Primary Elections, and Polarization in the United States Version 1.0

Analyze Kujala, Jordan, 2019, "Replication Data for: Donors, Primary Elections, and Polarization in the United States", <https://doi.org/10.7910/DVN/QYG8Z1>, Harvard Dataverse, V1, UNF:6:LCy0jZPmAlkxvCCy69Q= [fileUNF] Cite Dataset -
Learn about Data Citation Standards

Description ⓘ I examine the influence of partisan donors on the district-level ideological polarization of congressional candidates in the United States. I use data from 2002-2010 U.S. House elections which provide for the placement of major party primary winners on the same ideological dimension as their primary, general election, and partisan donor constituencies. Using this unique data set, I find strong evidence that the influence of donors in nominating contests is a source of polarization in the United States. House nominees are more responsive to their donor constituencies than either their primary or general electorates. I also find some evidence that the lack of general election competition affects nominee externity. In safer districts, Democratic incumbents appear more responsive to donors. However, Republican donors seem to demand proximity regardless of district competitiveness. Overall, the polarizing effects of donor constituencies dominate any moderating effects resulting in ideologically extreme nominees and, ultimately, members of Congress. (2019-07-01)

Subject ⓘ Social Sciences

Keyword ⓘ Primaries, Congress, Elections, Representation, Donors, Polarization

Related Publication ⓘ Kujala, Jordan, [date] "Donors, Primary Elections, and Polarization in the United States." *American Journal of Political Science* Forthcoming. <http://ajps.org/>

Notes ⓘ This dataset underwent an independent verification process that replicated the tables and figures in the primary article. For the supplementary materials, verification was performed solely for the successful execution of code. The verification process was carried out by the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill.

The associated article has been awarded Open Materials and Open Data Badges. Learn more about the Open Practice Badges from the Center for Open Science.



header adds to the top of the output a header that lists the dataset name, the date that the dataset was last saved, etc.

notes lists any notes attached to the variables; see [D] [notes](#).

mv specifies that **codebook** search the data to determine the pattern of missing values. This is a CPU-intensive task.



**Confirmable Reproducible
Research Environment:**
Linking tools to promote
computational reproducibility



- ★ Technology integrations to streamline verification workflows
- ★ Automated workflow tracking
- ★ Enabling tools that encourage data management and reproducible research standards and best practices



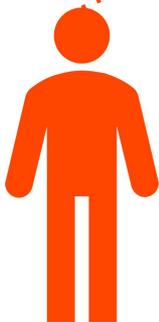
manuscript tracking



code execution



dataset publication



AUTHOR



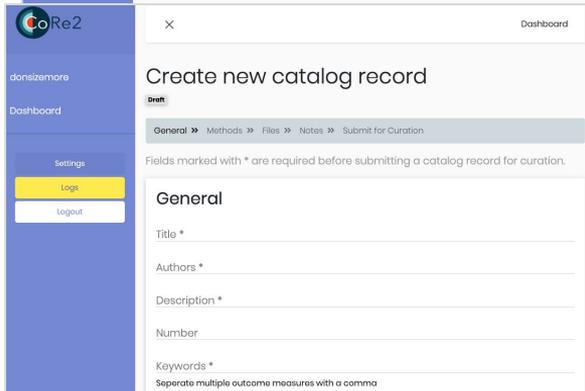
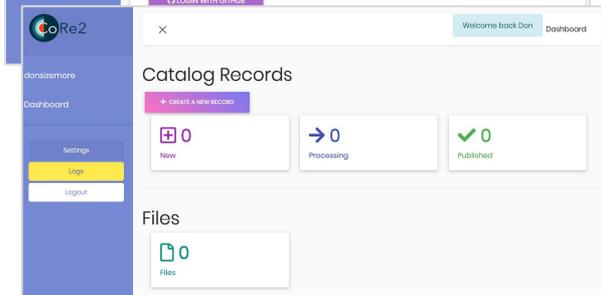
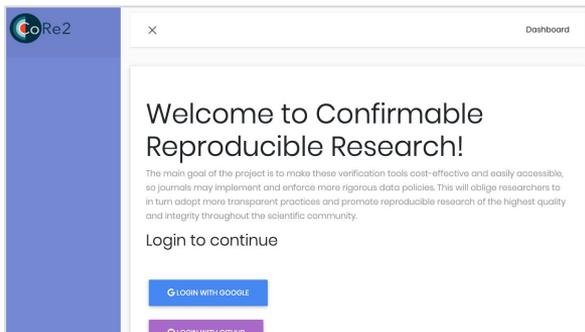
EDITOR



CURATOR



VERIFIER



- 1 Editor prompts author to submit dataset to CoRe2
- 2 Editor submits manuscript to CoRe2
- 3 Author uploads dataset to CoRe2 and runs code
- 4 Curator reviews package in CoRe2
- 5 Verifier reviews code in CoRe2
- 6 Curator enhances metadata and publishes dataset in CoRe2

Thu-Mai Christian
tlchristian@unc.edu



**Alfred P. Sloan
FOUNDATION**

Support for this research was provided by the Alfred P. Sloan Foundation (2018-11121). The views expressed here do not necessarily reflect the views of the Foundation.