Reproducibility and Quality Data: Presenting YARD

Limor Peer, PhD
Associate Director for Research
Institution for Social and Policy Studies
Yale University

MZES Open Social Science Conference | Mannheim, Germany | 25-27 January, 2019
Computational reproducibility is desirable and feasible.

"An experiment or analysis is preprodicible if it has been described in adequate detail for others to undertake it."

To achieve computational reproducibility, access to research output is necessary but not sufficient.

Producing quality research output for publication is a process – "curating for reproducibility" – that should be part of the research and data lifecycles.

YARD is a tool for managing that process.

Roadmap for this talk:
- Computational Reproducibility
- Access
- Preproducibility
- Quality Review
- YARD
The Yale Application for Research Data (YARD) is a proof of concept.

The concept is curating for reproducibility.

Curating for reproducibility is a process for producing quality research output for publication. It involves computational reproducibility.
computational reproducibility

- replicability
  - empirical reproducibility
  - methodological reproducibility

- repeatability
  - conceptual replication

- reproducibility
  - statistical reproducibility

- verification

- validation
  - direct replication
  - pre-reproducibility
  - reproducibility

http://cure.web.unc.edu/defining-reproducibility/
Computational reproducibility involves “calculation of quantitative scientific results by independent scientists using the original datasets and methods.”

Access to data and code is increasing

Social Sciences are at the forefront

“To the extent that researchers’ evidence-based knowledge claims rely on data they themselves generated or collected, they should,

- provide access to those data
- or explain why they cannot.”

https://www.dartstatement.org/
Access is necessary but not sufficient

Attempting to reproduce results you often find...

- Missing variable labels
- Insufficient documentation
- Omitted code
- Deviations in number of observations
- Unavailable software extensions
- Missing variables
- Incompatible datasets
How do we get from here...

To here...

Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment


http://hdl.handle.net/10079/zw3r2f9. ISPS Data Archive.

ISPS ID: D149

Related publications:

Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

Keyword(s): List experiment
computational reproducibility

REPLICABILITY

empirical reproducibility

methodological reproducibility

REPEATABILITY

conceptual replication

statistical reproducibility

VALIDATION
direct replication

PREPRODUCIBILITY

http://cure.web.unc.edu/defining-reproducibility/
An experiment or analysis is **preproducible** if it has been **described in adequate detail for others to undertake it**. Preproducibility is a prerequisite for reproducibility, and the idea makes sense across disciplines.

A matter of quality

A process:

Curating for reproducibility =
Prereproducibility +
Computational reproducibility

ISPS Data Archive at Yale University

https://isps.yale.edu/research/data
“Research curators as “first re-users”

“We are missing labels for the following variables: _n1, _n0, V1 and V0.”

Archive staff

"Here are the labels: 
_n1 is the number of observations in the treated strata before matching
_n0 is the number of observations in the comparison strata before matching
v1 = turnout for treated observations
v0 = turnout for comparison observations

... this reminds me that I needed to include the .ado code in the Matching Code folder. I just did that and updated the readme file. Boy, the things your forget about after not thinking about something for two years!"
DATA REVIEW

☑ Check for undocumented variable and value information
☑ Examine data for inconsistencies and errors
  ✓ Discrepancies in number of observations
  ✓ Out-of-range or wild codes
  ✓ Undefined null values
☑ Review data for confidentiality issues
☑ Review data for copyright or licensing issues

Perform Data Quality Review

Data Quality Review Framework

Data Quality Review Framework

CODE REVIEW

- Convert absolute file paths to relative file paths
- Check code for presence of non-executable comments that document analysis processes
- Identify packages required to execute code
- Execute code to ensure code is error-free
- Compare code output to findings presented in paper

Perform Data Quality Review

Managing Data Quality Review

1. Assign staff to study and files
2. Move original files to Archive space
3. Make copies of processed files and move to collaborative space
4. Identify related publications and projects
5. Rename all copied files for public dissemination according to naming conventions
6. Check variable and value labels
7. Compare variable information with other materials
8. Check for additional variables and recoded variables
9. Check and complete variable-level metadata for each data file
10. Check all files for confidential and other sensitive information
11. Create DDI-XML file with variable-level information
12. Run the statistical code
13. Check output against published results
14. Re-write statistical code in R and check replication
15. Communicate with PI as needed
16. Create additional files by converting to readable formats (e.g., ASCII, PDF)
17. Update study- and file-level metadata record
18. Update tracking documents: process record / general study database / status document

... and now share the data!
A workflow tool that allows Depositors, Curators, and Administrators to submit, review, process, and publish data within one system, which structures and records data curation and code review workflow actions.
YARD: A curation workflow tool

Main features

Data curation and code review

• Integration
• Tracking; Management
• Open source; Modular
• DDI metadata production
• Push out information to specified destinations

• Users

DEPOSITOR
CURATOR
ADMIN
Curation Tool: YARD

Data Deposit

Create New Catalog Record

General

Title *

Authors *

Description *

Number

Keywords *

Separate multiple outcome measures with a comma.

Citation

Funding *

Access

Access Statement *

Not Selected
Curation Tool: YARD

Data Deposit

Research Study 2

General Methods Files Notes Submit for Curation

Methods

Research Design * Natural experiment

Mode of Data Collection * Interview: Face to Face, Interview: email

Field Dates * 2017-05-19

Study Time Period * 2017-05-19 - 2013-05-08

Location * Africa

Sample Size * 5000

Inclusion/Exclusion Criteria * None

Randomization Procedure * None

Depositor
Curator
Admin

#OSSC19 @l_peer
Curation Tool: YARD

Data Deposit

Research Study 2

Files

<table>
<thead>
<tr>
<th>File</th>
<th>Version</th>
<th>Type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>02_PerformAnalysis.do</td>
<td>1</td>
<td>Program</td>
<td>Accepted</td>
</tr>
<tr>
<td>POQ_PublicReplicationDatafile.dta</td>
<td>1</td>
<td>Data</td>
<td>Accepted</td>
</tr>
<tr>
<td>testdata.txt.doc</td>
<td>1</td>
<td>Supplementary Materials</td>
<td>Accepted</td>
</tr>
<tr>
<td>README.txt</td>
<td>1</td>
<td>Supplementary Materials</td>
<td>Accepted</td>
</tr>
</tbody>
</table>

Add or Update Files  Download All Files
Curation Tool: YARD

Data Deposit

Submit for Curation
Your record is ready for curation. Click the button below to submit it.

Deposit Agreement
ISPS Data Archive: Data Deposit Agreement
Curation Tool: YARD

Data Curation

Research Study

Task | Status | Completed
--- | --- | ---
Collection | Created | deposit.demo@example.edu on 5/12/2017
| Accepted | deposit.demo@example.edu on 5/12/2017
Processing | Review Observation Count | 
| Check Missing Labels | 
| Compare Questionnaire, Codebook, and Data in Data File | 
| Check for Personally-Identifiable Information (PI) in Data File | 
| Identify Potential Errors in Data File | 
| Confirm Code Executes | 
| Create Code Report/Documentation | 

Depositor
Curator
Admin

#OSSC19 @l_peer
Curation Tool: YARD

Data Curation

POQ_PublicReplicationDatafile.dta

weight
case weight

yal130
voting private matter

Frequencies

<table>
<thead>
<tr>
<th>Value</th>
<th>Label</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>democracy works best when people treat their vote choices as personal matters</td>
<td>$56</td>
</tr>
</tbody>
</table>
Curation Tool: YARD

Data Curation

Request Publication

When curation is complete, you can request publication.

Please complete the following fields before requesting publication.

- Incomplete task for 02_PerformAnalysis.dox: Confirm Code Executes
- Incomplete task for testdattext.docx: Create Preservation Format
- Incomplete task for POQ_PublicReplicationDatafile.dta: Check Missing Labels
- Incomplete task for POQ_PublicReplicationDatafile.dta: Identity Potential Errors in Data File
- Incomplete task for POQ_PublicReplicationDatafile.dta: Check for Personally-Identifiable Information (PI) in Data File
- Incomplete task for 02_PerformAnalysis.dox: Confirm Code Replicates Reported Results
- Incomplete task for POQ_PublicReplicationDatafile.dta: Compare Questionnaire, Codebook, and Data in Data File
- 02_PerformAnalysis.dox: Software
testdattext.docx: Software
POQ_PublicReplicationDatafile.dta: Data Type
testdattext.pdf: Software
README.txt: Software
Curation Tool: YARD

Data Curation

File History

Edit a File
- Data Type changed from blank to Survey:Interview (e.g., ANES)
- Source changed from blank to blank
- Test October 22
  - Butler_Dynes_AJPS_2015a_StateLegislators.dta

Accept task: Identify Potential Errors in Data File
- No obvious data errors to flag
- Test October 22
  - Butler_Dynes_AJPS_2015a_StateLegislators.dta

Accept task: Check for Personally-Identifiable Information (PII) in Data File
- No PII
Curation Tool: YARD

Data Curation

Research Study 2

Public and Archive

- Approve Publication
- Reject Publication

Depositor
Curator
Admin

#OSSC19 @l_peer
YARD: A curation workflow tool
Implementing and enforcing journal data replication policies

Providing data curation and Results Reproduction ($R^2$) services

Aligning data curation workflows and systems with Data Quality Review

http://cure.web.unc.edu/
Reproducibility and Quality Data: Presenting YARD

Limor Peer, PhD
Associate Director for Research

Institution for Social and Policy Studies
Yale University
limor.peer@yale.edu
https://isps.yale.edu/team/lmor-peer
@l_peer

Acknowledgments:
Innovations for Poverty Action
Colectica
Yale University Library
Yale Digital Scholarship
Yale Sites
Digital Lifecycle Research & Consulting

#OSSC19 @l_peer
About YARD:
https://yard.yale.edu/

Technical information:
https://github.com/Colectica/curation & https://docs.colectica.com/curation/

For general questions about the YARD open source code and how it could be adapted to feed into other archives and repositories and enhanced to integrate with other applications:
https://www.colectica.com/support.

For questions about the ISPS Data Archive:
isps@yale.edu.

For questions about YARD implementation at Yale:
digitalscholarship.services@yale.edu.