

Curating for Reproducibility

Limor Peer, PhD

Institution for Social and Policy Studies, Yale University
& CURE (Curating for Reproducibility) consortium

Joining Forces to Promote Research Transparency

IASSIST | Montreal, Canada | May 31, 2018

#otherpeoplesdata

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.

Details



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

8 10



Lu Hugerth @luhugerth

Follow

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

4:28 PM - 11 Dec 2013

2 RETWEETS 1 FAVORITE

8 10



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

Details

8 10



Christie Bahlai @cbahlai

Follow

MERGED CELLS IN EXCEL SPREADSHEET
NOOOOOOOO! #otherpeoplesdata

11:10 AM - 6 Dec 2013

4 RETWEETS 5 FAVORITES

8 10



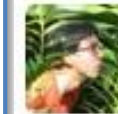
Jonathan Carroll @carroll_j... 8d
I have a
"data_merged_final_slightlyimproved_on_Thursday.csv" related headache. #otherpeoplesdata

8 10

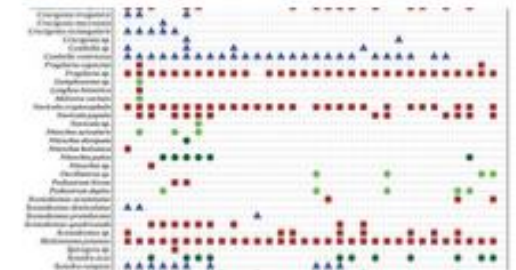


Christie Bahlai @cbahlai 21h
Aaaannnd: unexplained missing data. Must consult with data creator before proceeding. Will write stats code while I wait.
#otherpeoplesdata

Details



Rhymes With @squirrelbert 33d
Dear authors, thx for sharing your data in such a visually pleasing but difficult to reuse format #otherpeoplesdata
pic.twitter.com/o6RiDAMwZJ



8 10

Yale Institution for Social and Policy Studies Data Archive (ISPS)
Curating for reproducibility: Why? What? How?
Curating for Reproducibility (CURE)
Curation Tool: Yale Application for Research Data (YARD)

Yale Institution for Social and Policy Studies Data Archive (ISPS)

Curating for reproducibility: Why? What? How?

Curating for Reproducibility (CURE)

Curation Tool: Yale Application for Research Data (YARD)

ISPS was founded in 1968 as an interdisciplinary center to support social science and public policy research at Yale University



ISPS Data Archive

| | | | |
|---------------------|-----------------------|----------------------------|--------------------------|
| Author - Any - | Discipline - Any - | Keywords - Any - | Area of Study - Any - |
| Location - Any - | Year -Year | Research design - Any - | SEARCH |

| TITLE | AUTHOR(S) | YEAR ARCHIVED |
|---|---|------------------|
| Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment | Alexander Coppock | 2017 |
| Trading Barriers: Firms, Immigration and the Remaking of Globalization | Margaret E. Peters | 2017 |
| The Majority-Minority Divide in Attitudes Toward Internal Migration: Evidence from Mumbai | Nikhar Gaikwad and Gareth Nellis | 2017 |
| Chocolate Scents and Product Sales: A Randomized Controlled Trial in a Canadian Bookstore and Café | Mary C. McGrath, Peter M. Aronow, Vivien Shotwell | 2017 |

Since 2011
Specialized community
Open access
Website integration

90 studies
1,400 files
15 GB

<https://isps.yale.edu/research/data>

An open access digital collection of social science experimental data, metadata, code, and associated files produced by ISPS researchers, for the purpose of replication of research findings, further analysis, and teaching.

Peer, L., & Green, A. (2012). Building an Open Data Repository for a Specialized Research Community: Process, Challenges, and Lessons. *International Journal of Digital Curation* 7(1), 151–162.
<http://dx.doi.org/10.2218/ijdc.v7i1.222>

Replication presents data sharing (and preservation) with a concrete purpose. This essentially prescribes that steps be taken to ensure that the right materials are shared and used in the right way. These steps should be taken by the entity that assumes responsibility over the data (e.g., a repository, a journal, funder website, etc.), and they are an essential part of data curation.

Peer, L. (2011). Building an Open Data Repository: Lessons and Challenges. *SSRN*.
<http://dx.doi.org/10.2139/ssrn.1931048>

See also,

Peer, L. (2013). The Role of Data Repositories in Reproducible Research. *ISPS Blog*.
<https://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research>

Peer, L. (2013). The Repository as Data (Re) User: Hand Curating for Replication. *Yale Day of Data*.
<https://elischolar.library.yale.edu/cgi/viewcontent.cgi?article=1017&context=dayofdata>

Yale Institution for Social and Policy Studies Data Archive (ISPS)

Curating for reproducibility: Why? What? How?

Curating for Reproducibility (CURE)

Curation Tool: Yale Application for Research Data (YARD)

"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.



Stark, P. (2018). Before Reproducibility Must Come Preproducibility. *Nature* 557, 613.
<https://www.nature.com/articles/d41586-018-05256-0>

The *replication standard* holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author.

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, 28(3), 444–452.
<http://doi.org/10.2307/420301>

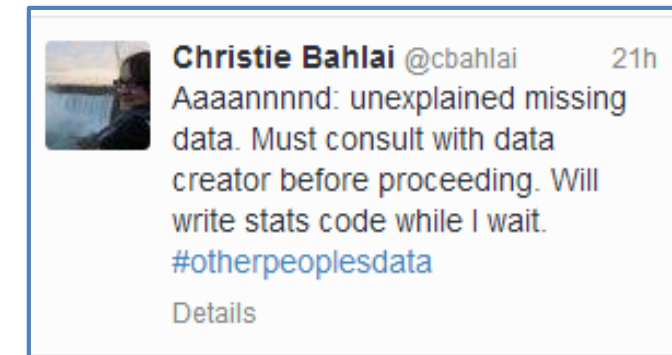
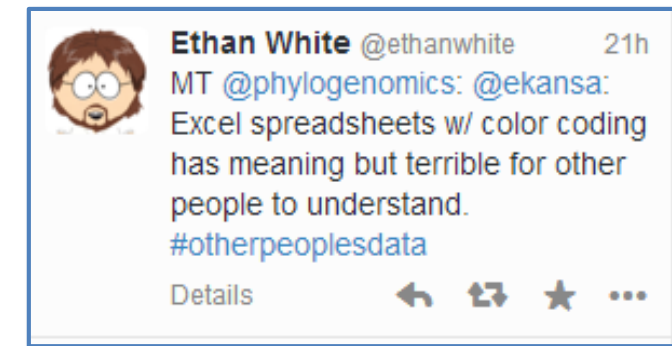
Reproducibility: Calculation of quantitative scientific results by independent scientists using the original datasets and methods.

Stodden, V. (Ed.), Leisch, F. (Ed.), Peng, R.D. (Ed.). (2014). *Implementing Reproducible Research*. New York: Chapman and Hall/CRC.

How to curate for reproducibility?

The most commonly reported problems associated with [replication] attempts were the lack of... data and code, *followed by insufficient documentation.*

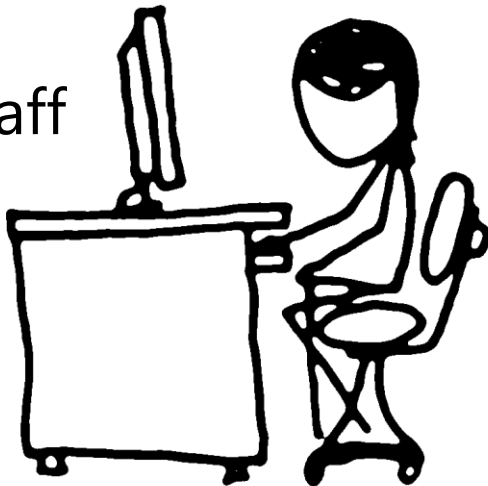
Janz, N., Werfel, S., Wykstra S. (2014). Replication in political science graduate courses: an untapped resource? *Monkey Cage*.
<https://www.washingtonpost.com/news/monkey-cage>



ISPS Data Archive: First Re-User

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

Archive staff



<http://xkcd.com/662/> Creative Commons Attribution-Noncommercial

"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 you forget about after not thinking about something for two years!"

Researcher

- Insufficient documentation
- Missing variables
- Deviations in number of observations
- Unavailable software extensions
- Omitted code
- Incompatible datasets

Data Quality Review



Data Quality Review



- ✓ Assign persistent identifier
- ✓ Create study citation and study-level metadata record
- ✓ Record file size details
- ✓ Check for presence of all files
- ✓ Verify content of files matches expected format
- ✓ Create non-proprietary versions of files
- ✓ Implement migration strategy for file formats

Data Quality Review



- ✓ Confirm presence of comprehensive descriptive information necessary for informed reuse
 - Data definitions
 - Variable construction
 - Methodology
 - Sampling information
 - Original data source citation
 - Analysis software version
- ✓ Link to related research products

Data Quality 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

Data Quality 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 article

Yale Institution for Social and Policy Studies Data Archive (ISPS)

Curating for reproducibility: Why? What? How?

Curating for Reproducibility (CURE)

Curation Tool: Yale Application for Research Data (YARD)

Data curation and code review for the purpose of facilitating the digital preservation of the evidence base necessary for future understanding, evaluation, and replication of scientific claims.

Establish Standards

Share Practices

Promote Data Quality Review

THE ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

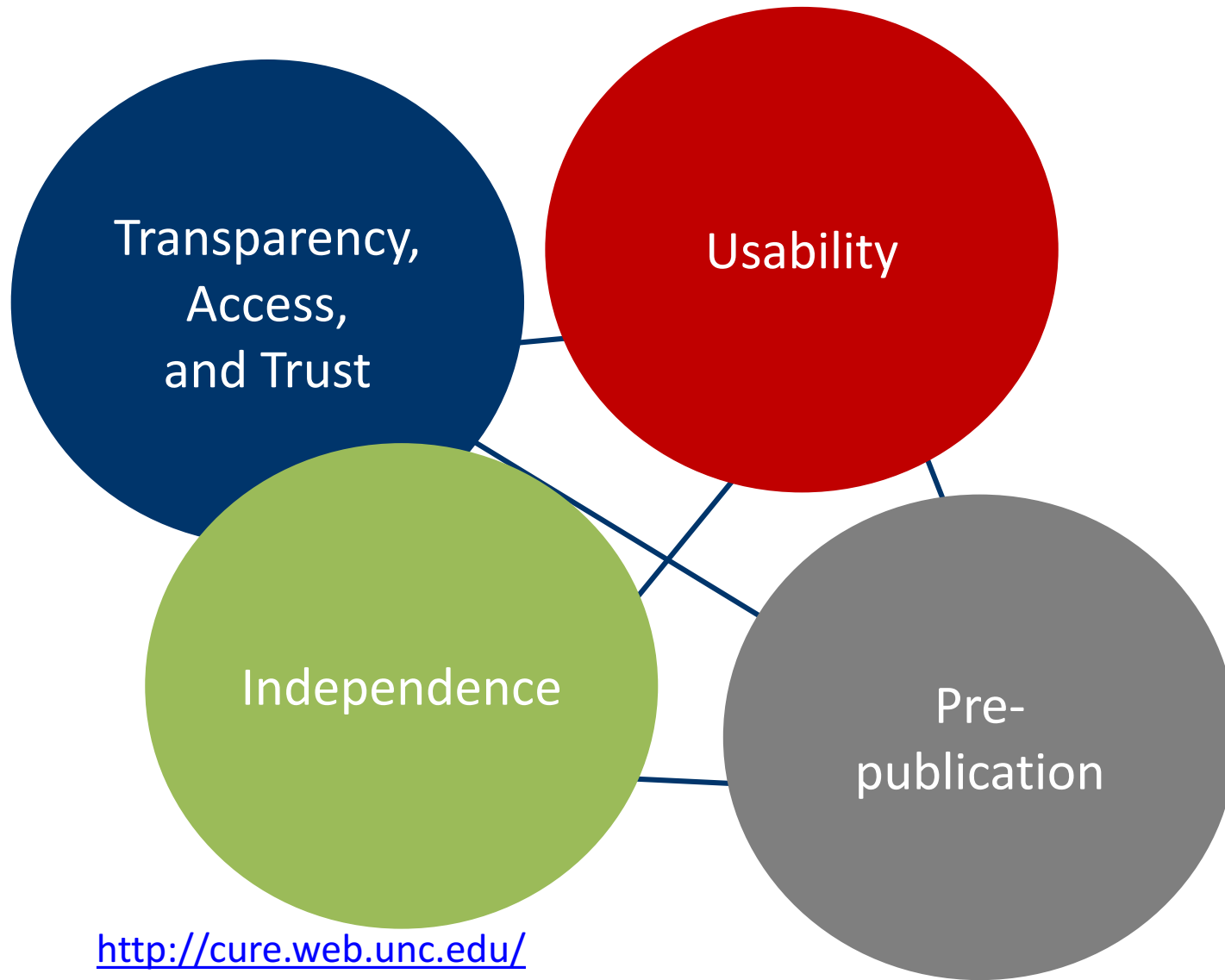
<http://cure.web.unc.edu/>

CISER

Yale
ISPS

Yale

CURE: Curating for Reproducibility



<http://cure.web.unc.edu/>

The CURE Consortium is committed to building a community of practice to support data curation for reproducibility.

We do this through establishing standards, sharing practices, and promoting the philosophy of Data Quality Review.



In 2017, CURE received a grant from the Institute of Museum and Library Services (IMLS) Laura Bush 21st Century Librarian Program to define the necessary skills for this type of work and to develop an evidence-based training program focused on data curation for reproducibility for academic librarians and archivists.

Models of practice:

Institution for Social and Policy Studies (ISPS)
Yale University

*** Aligning Data Curation Workflows with Data Quality Review**

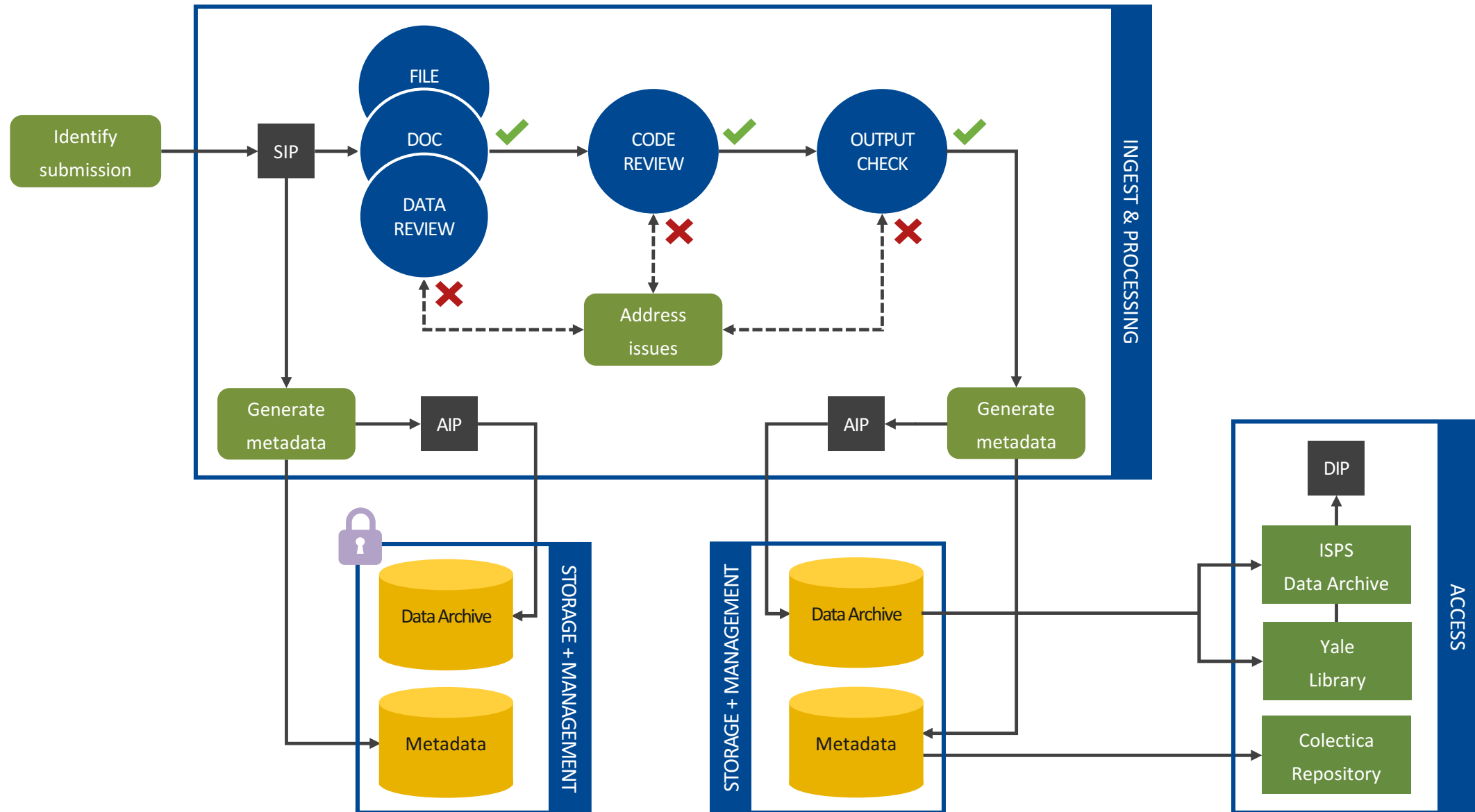
Cornell Institute for Social and Economic Research
Cornell University

*** Providing Data Curation and Reproduction of Results (R²)
Services**

Odum Institute for Research in Social Science
University of North Carolina, Chapel Hill

*** Enforcing Journal Data Replication Policies**

Curating for Reproducibility at ISPS



Yale Institution for Social and Policy Studies Data Archive (ISPS)
Curating for reproducibility: Why? What? How?
Curating for Reproducibility (CURE)
Curation Tool: Yale Application for Research Data (YARD)

YARD: Yale Application for Research Data

A new workflow tool that allows Depositors, Curators, and Administrators to submit, review, process, and publish data within one system. The software structures the curation and review workflow and all actions are recorded in the system. The tool integrates and captures DDI metadata production with data and code review and cleaning. Processed data packages are directed to pre-specified destinations.



Production and code release in 2017-2018

Curation tool: YARD

Log in



Log in to the ISPS Data Curation Tool with your username and password.

Don't have a ISPS Data Curation Tool account?

[Create an account.](#)

Email

Password

☐

Remember me

Log in

[Forgot your password?](#)

<https://docs.colectica.com/curation/>

Thank you!

limor.peer@yale.edu

[@l_peer](#)

<https://isps.yale.edu/team/limor-peer>

About Curating for Reproducibility

CURE: <http://cure.web.unc.edu/>

CISER: <https://ciser.cornell.edu/>

ISPS: <http://isps.yale.edu/>

Odum: <http://www.odum.unc.edu/>