



*Measuring the Impacts of
Curatorial Actions on
Research Data Reuse at
ICPSR*



HOST



Rob Grim

Economics (Data) Librarian

Erasmus University, the Netherlands

Chair of LIBER's Research Data Management
Working Group

<https://libereurope.eu/working-group/research-data-management/>

SPEAKER



Sara Lafia

Postdoctoral Research Fellow

Inter-university Consortium for Political and
Social Research (ICPSR)

University of Michigan

slafia@umich.edu@lafia_s



NOTES

- The webinar is being recorded. All participants will receive a link to the recording later today.
- Slides are on Zenodo: See the chat box for the link.
- Questions? Put them in the chat box. We'll put questions to the speakers at the end of the webinar.

Measuring the Impacts of Curatorial Actions on Research Data Reuse

December 3, 2020

Sara Lafia, Ph.D.

Research Fellow, ICPSR, University of Michigan

What is ICPSR?



Founded in 1962 by 22 universities, now consortium of 800 institutions world-wide



Focus on social and behavioral science data, broadly defined



10,000 studies, 250,000 files



1500 are restricted studies, almost always to protect confidentiality

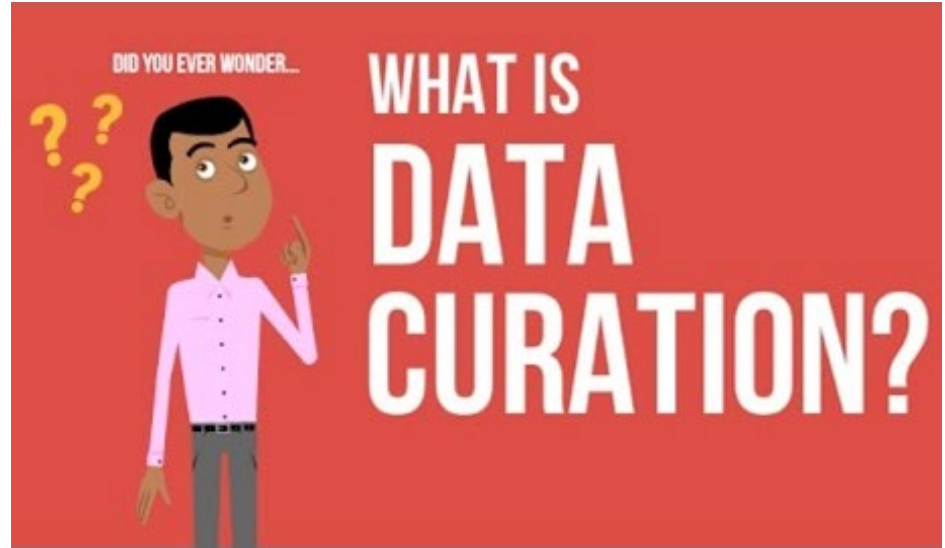


Bibliography of Data-related Literature with 94,000 citations



Approximately 60,000 active MyData (“shopping cart”) accounts

Data Curation at ICPSR



Making sure people can
find and **use** data, now
and in the future

Motivation



What are the **values** and **costs** associated with sharing research data?

Introducing **MICA**: Measuring the Impacts of Curatorial Actions

*Develop curatorial metrics to evaluate the impact and efficacy of specific **data curation** activities on **data reuse**.*

Funders



IMLS Award #LG-37-19-0134-19



NSF Award #1930645

Team



Andrea Thomer
UMSI



Dharma Akmon
ICPSR



Amy Pienta
ICPSR



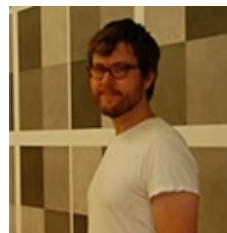
Elizabeth Moss
ICPSR



Libby Hemphill
ICPSR, UMSI



Elizabeth Yakel
UMSI



David Bleckley
ICPSR



Sara Lafia
ICPSR

Research Questions

What impacts do specific curatorial actions have on research data's impact or reuse?

How should we prioritize curatorial actions to achieve impact and return on investment?

Project Activities

Activity

Method

Understanding priorities and
defining curatorial actions

semi-structured interviews,
content analysis

Measuring reuse and impact

multivariate regression,
structural equation modeling

Generating curatorial metrics

structural equation modeling,
path analysis

Methods



Code stakeholder
interviews

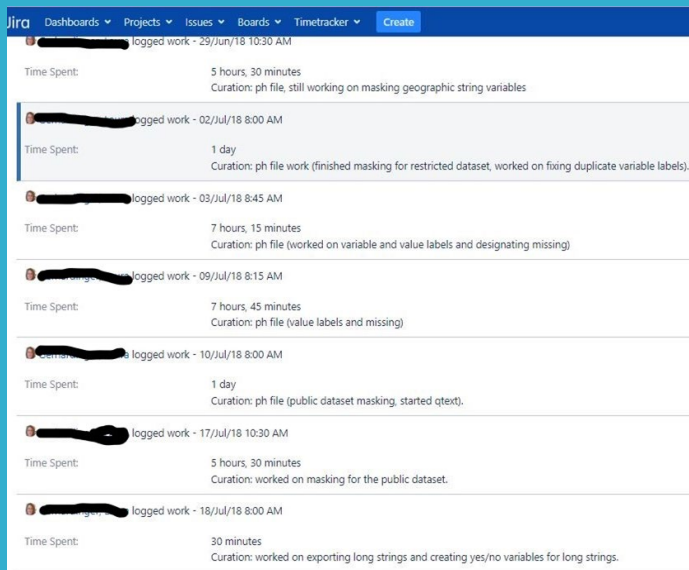


Annotate and **mine**
curation logs



Discover data
citations

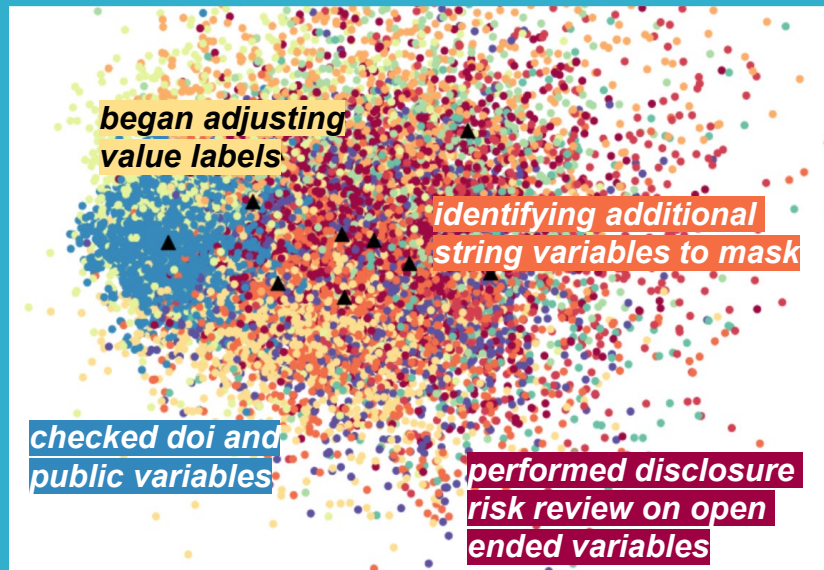
Defining Curatorial Actions



A screenshot of a Jira interface showing a list of work logs for a specific user. The logs are organized chronologically, showing the time spent and the duration of various tasks related to data curation and masking.

Logged work - 29/jun/18 10:30 AM
Time Spent: 5 hours, 30 minutes Curation: ph file, still working on masking geographic string variables
Logged work - 02/jul/18 8:00 AM
Time Spent: 1 day Curation: ph file work (finished masking for restricted dataset, worked on fixing duplicate variable labels).
Logged work - 03/jul/18 8:45 AM
Time Spent: 7 hours, 15 minutes Curation: ph file (worked on variable and value labels and designating missing)
Logged work - 09/jul/18 8:15 AM
Time Spent: 7 hours, 45 minutes Curation: ph file (value labels and missing)
Logged work - 10/jul/18 8:00 AM
Time Spent: 1 day Curation: ph file (public dataset masking, started qtext).
Logged work - 17/jul/18 10:30 AM
Time Spent: 5 hours, 30 minutes Curation: worked on masking for the public dataset.
Logged work - 18/jul/18 8:00 AM
Time Spent: 30 minutes Curation: worked on exporting long strings and creating yes/no variables for long strings.

Curator work logs



Classification of curation actions

Measuring Reuse and Impact



Secondary impact: how many times studies that reuse a particular dataset are cited



Diversity: breadth of disciplines that use the data

Tracking Data Citations

ICPSR Find & Analyze Data [Log In/Create Account](#)

[FIND DATA](#) [SEARCH/COMPARE VARIABLES](#) [DATA-RELATED PUBLICATIONS](#) [THEMATIC DATA COLLECTIONS](#) [HELP](#)

Filters

Pub. Year

from

to

Pub. Type

Journal

Author

Study

Search Results

Showing 1 - 50 of 94,024 results.

The ICPSR *Bibliography of Data-related Literature* is a frequently-updated database of thousands of citations for publications that analyze data held at ICPSR.

Studies (15,478) **Variables (5,645,916)** **Series (274)** **Data-related Publications (94,024)** **ICPSR Website (862)**

Related Studies/Series: ☒ Visible Sort by: Pub Date (newest)

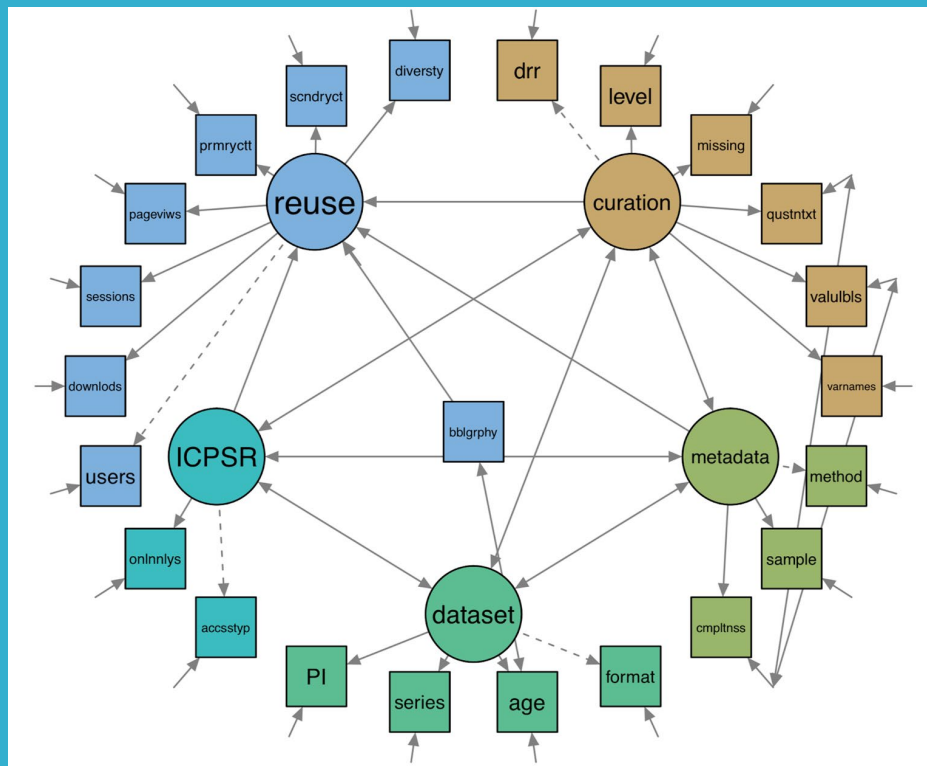
	Pub. Type	Pub. Year	Citation
<input type="checkbox"/>		2021	Essau, Cecilia A., de la Torre-Luque, Alejandro Parent's psychopathological profiles and adolescent offspring's substance use disorders. <i>Addictive Behaviors</i> 112. Full Text Options: DOI WorldCat Google Scholar Export Options: BIS FullText EndNote Studies related to this publication: <ul style="list-style-type: none">National Comorbidity Survey Adolescent Supplement (NCS-A) 2001-2004 (ICPSR 28581)
<input type="checkbox"/>		2021	Haas, Brian W., Hoeff, Fumiko, Omura, Kazufumi The role of culture on the link between worldviews on nature and psychological health during the COVID-19 pandemic. <i>Personality and Individual Differences</i> 170, (110336). Full Text Options: DOI WorldCat Google Scholar Export Options: BIS FullText EndNote Studies related to this publication: <ul style="list-style-type: none">Midlife in the United States (MIDUS 2) 2004-2006 (ICPSR 4652)Survey of Midlife in Japan (MIDJA 2) May-October 2012 (ICPSR 36427)
<input type="checkbox"/>		2021	Jenkins, Jade M., Duer, Jennifer K., Connors, Maia Who participates in quality rating and improvement systems? <i>Early Childhood Research Quarterly</i> 54, (1st Quarter), 219-227. Full Text Options: DOI WorldCat Google Scholar

```
{
  "data_set_id": 549,
  "unique_identifier": "10.3886/ICPSR02940",
  "title": "Monitoring the Future: A Continuing Study of American Youth (8th- and 10th-Grade Surveys), 1999",
  "name": "Monitoring the Future: A Continuing Study of American Youth (8th- and 10th-Grade Surveys), 1999",
  "description": "These surveys of 8th- and 10th-grade students are part of a series that explores changes in import",
  "date": "2005-11-04 00:00:00+00:00",
  "coverages": "",
  "subjects": "adolescents,attitudes,demographic characteristics,drug use,elementary school students,family life,jun",
  "methodology": "",
  "citation": "",
  "additional_keywords": "dimensions_oa",
  "family_identifier": "313",
  "mention_list": [
    "MTF",
    "MTF data",
    "MTF study",
    "MTF survey",
    "MTF survey is supported by the National Institute on Drug Abuse",
    "MTF surveys",
    "MTFS",
    "Monitoring the Future",
    "Monitoring the Future (MTF)",
    "Monitoring the Future (MTF) Project",
    "Monitoring the Future (MTF) study",
    "Monitoring the Future (MTF) survey",
    "Monitoring the Future (MTF) survey, funded by the National Institute on Drug Abuse",
    "Monitoring the Future Survey",
    "Monitoring the Future panels",
    "Monitoring the Future project",
    "Monitoring the Future study",
    "Monitoring the Future study (MTF)",
    "Monitoring the Future survey",
    "annual Monitoring the Future survey"
  ],
  "identifier_list": [
    {
      "name": "ICPSR Data ID (dataId)",
      "identifier": "10.3886/ICPSR02940"
    }
  ]
}
```

Data-related bibliography

Detecting data citations

Generating Curatorial Metrics



Pilot: What Impacts Data Downloads?

Year of Release	# Studies Released	Percent (n=2,225)
2006	173	7.78
2007	202	9.08
2008	270	12.13
2009	210	9.44
2010	196	8.81
2011	217	9.75
2012	177	7.96
2013	275	12.36
2014	224	10.07
2015	281	12.63

Pilot: What Impacts Data Downloads?

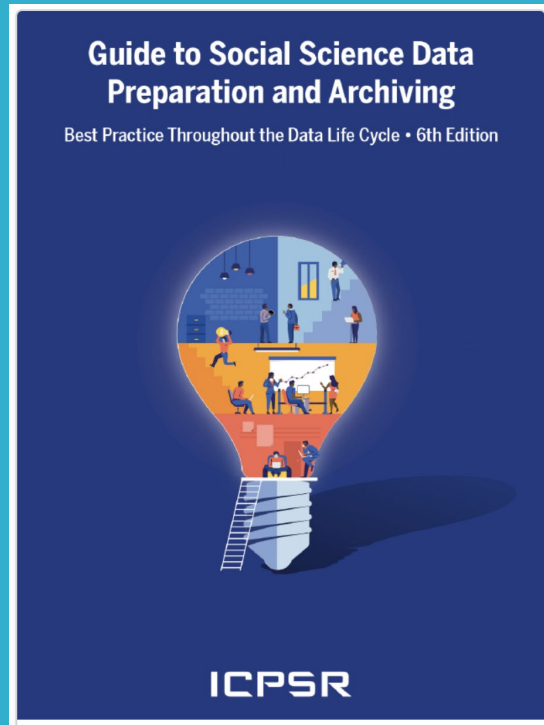
	Model 1	Model 2	Model 3
Study Properties			
Series (0 = No; 1 = Yes)	-0.55***	-0.52***	-0.70***
Variables (0 = 0-199; 1 = 200+)		0.93***	0.61***
Single PI (0 = No; 1 = Yes)			-0.15*
Curation Activity			
Sponsored (0 = No; 1 = Yes)			0.87***
Intense curation (0 = No; 1 = Yes)			0.60***
N metadata terms			0.01*

Pilot: What Impacts Data Downloads?

Predictions	Result
Curation activity will increase downloads.	Support
Which curation activities will have the most impact?	Sponsorship and intense curation
Bigger studies will receive more downloads.	Supported
Series will receive more downloads	Not supported

What does this mean for data depositors?

Curation, especially metadata terms and well-documented variables, increases reuse.



Thank you!

Main Goal

Develop curatorial metrics to evaluate the impact and efficacy of specific data curation processes

Award Information

- IMLS Award #LG-37-19-0134-19
- NSF Award #1930645

Contact

Sara Lafia

slafia@umich.edu

[@lafia_s](#)



THANKS!

Questions?

Please put them in the chat box.

Slides and a recording will be sent to all registered delegates.