ICPSR

SHARING DATA TO ADVANCE SCIENCE

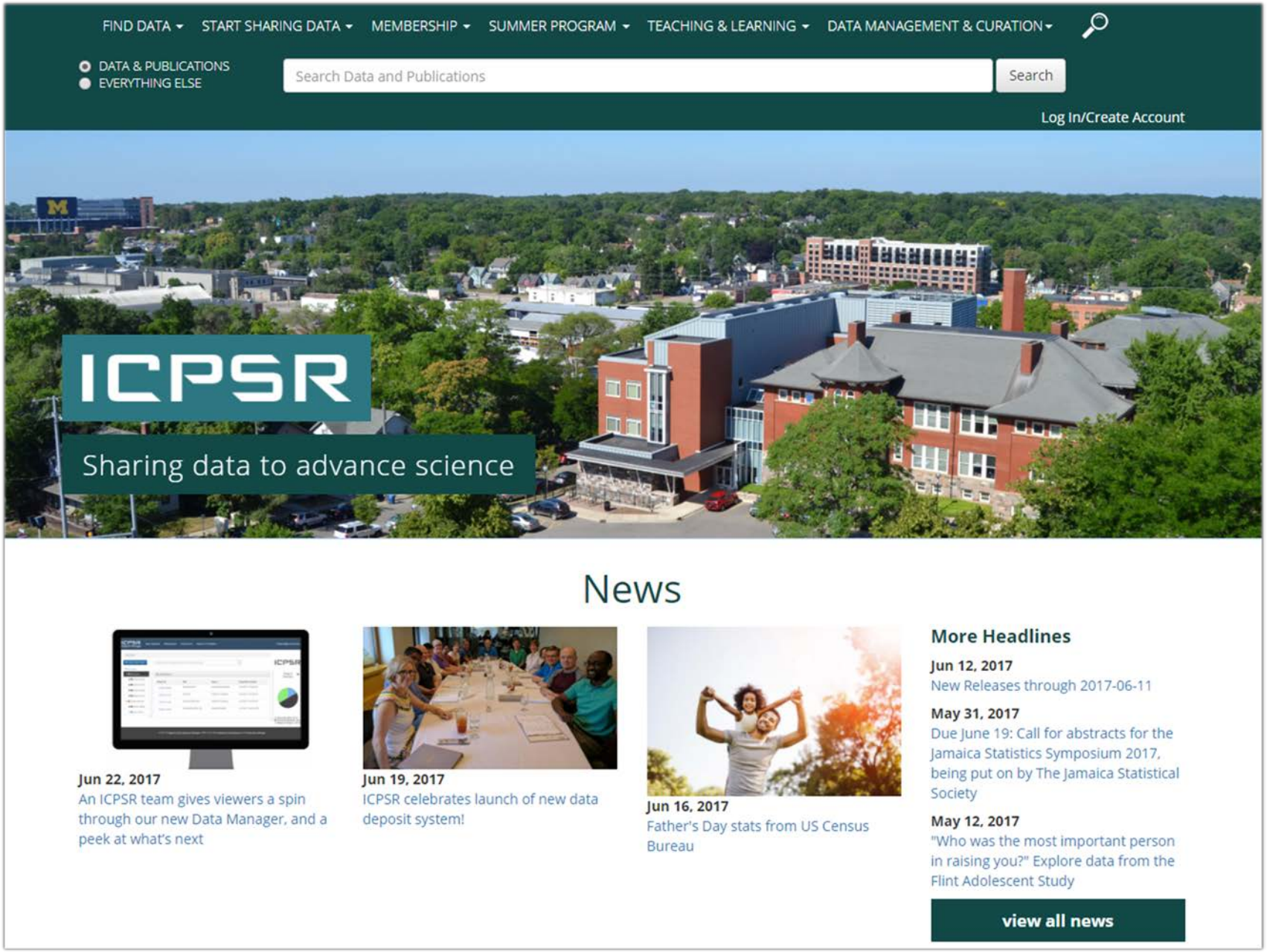
# Informal Data Citation: Its Impact on Tracking Shared Data Reuse

Elizabeth Moss & Jared Lyle
IASSIST
Sydney, NSW, Australia
May 29, 2019

ICPSR INSTITUTE FOR SOCIAL RESEARCH
UNIVERSITY OF MICHIGAN

# Overview

- Why data citations matter
- Challenges of tracking informal data citation
- Specific types of informal data citation
- Suggested improvements




http://www.icpsr.umich.edu

# ICPSR

Philip Converse, Warren Miller, and Angus Campbell

- Established 1962
- Originally 22 Members, now consortium of 776 world-wide
- Originally Political Science, now all social and behavioral sciences

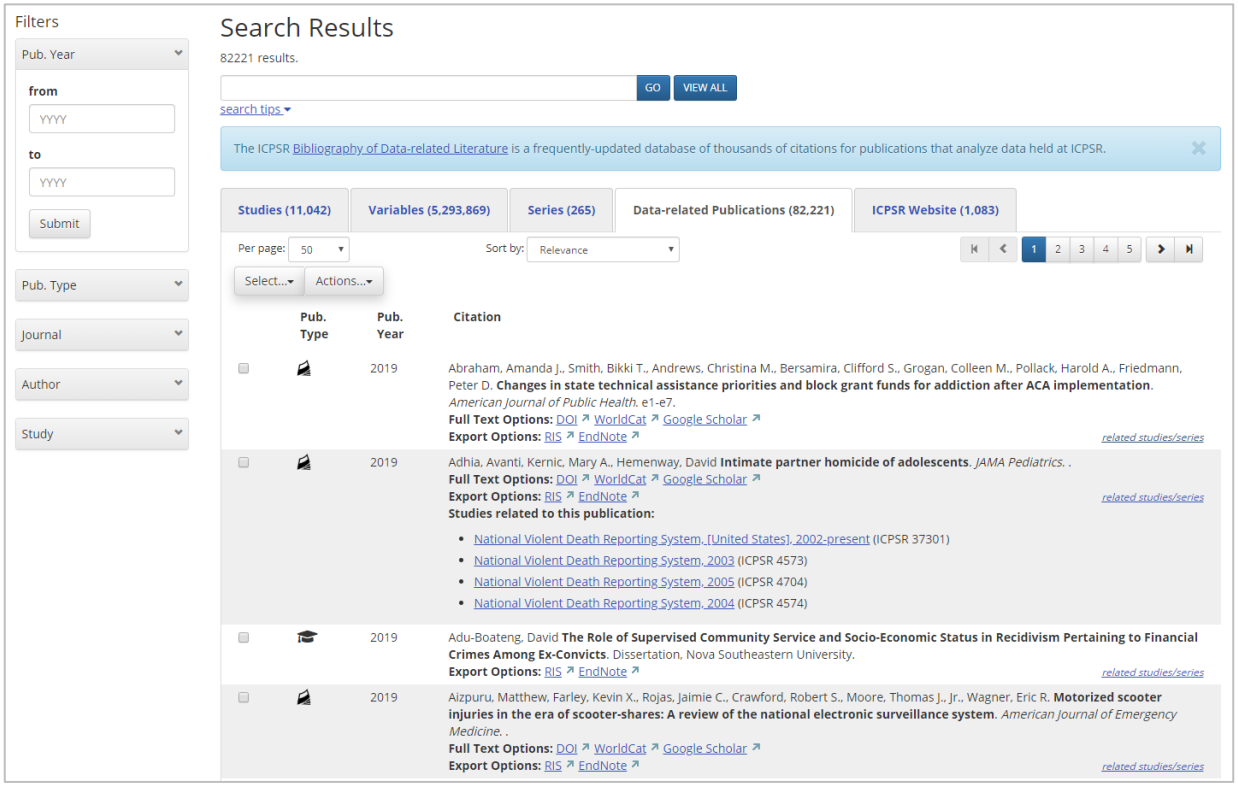
Source: http://www.icpsr.umich.edu/icpsrweb/content/membership/history/timeline.html

# ICPSR

- Current holdings
  - 10,000+ studies, quarter million files
  - 1500+ are *restricted studies*, almost always to protect confidentiality
- Approximately 60,000 active MyData (“shopping cart”) accounts
- Thematic collections of data about addiction and HIV, aging, arts and culture, child care and early education, criminal justice, demography, health and medical care, and minorities

# ICPSR Bibliography

- 82,000 citations of published and unpublished works resulting from analyses of data held in the ICPSR archive.



## About the Bibliography

The *ICPSR Bibliography of Data-related Literature* is a searchable database that as of 2019 contains 80,000 citations of published and unpublished works resulting from analyses of data held in the ICPSR archive. Developed with support from the National Science Foundation (SES-9977984), the *Bibliography* represents over 50 years of scholarship in the quantitative social sciences, extending from the inception of ICPSR in 1962 to the present. (more)

# Data-Related Publications

The *ICPSR Bibliography of Data-related Literature* is a continuously-updated database of thousands of citations of works using data held in the ICPSR archive. The works include journal articles, books, book chapters, government and agency reports, working papers, dissertations, conference papers, meeting presentations, unpublished manuscripts, magazine and newspaper articles, and audiovisual materials.

women india GO

- View all citations
- Browse by author
- Browse by journal
- Browse by study

## Search Tips

- Our citation search indexes only the literal citations, not the full text of the publications.
- Instead of entering a research question, use only one or two search words, and use the filters to narrow your results.
- Use quotes for phrase searches; use the minus sign to remove items from results: `"drug abuse" -adolescent`
- A Boolean "and" is invoked with multiple search terms.
- Stemming is automatic; do not use an asterisk.
- Search is not case-sensitive.

## Submit Citations

ICPSR encourages its users to submit bibliographic citations to data we disseminate. To add a new citation to our *Bibliography*, please use our citation form or email us at bibliography@icpsr.umich.edu.

3:37 AM Wed May 29

71%

icpsr.umich.edu

Filters

Pub. Year

Pub. Type

Journal

Author

Study

India Human Development Survey (IHDS), 2005 (32)

India Human Development Survey-II (IHDS-II), 2011-12 (22)

National Youth Survey [United States]: Wave III, 1978 (1)

Rural Development in Deccan Maharashtra, India: Village Panel Study, 1942-1977

# Search Results

44 results.

Women India GO VIEW ALL

search tips

Studies (1,195) | Variables (97) | Series (113) | Data-related Publications (44) | ICPSR Website (2)

Per page: 50

Sort by: Relevance

Select... Actions...

| | Pub. Type | Pub. Year | Citation |
|---|---|---|---|
| ☐ | | 2017 | Vernekar, Nisha, Singhal, Karan **Women's Agency Freedoms and Education Levels in the Post-marital Household: Evidence from Rural India**. *3rd World Conference on Women's Studies*. Colombo, Sri Lanka. **Full Text Options:** Abstract PDF DOI **Export Options:** RIS EndNote **Studies related to this publication:** • India Human Development Survey-II (IHDS-II), 2011-12 (ICPSR 36151) *related studies/series* |
| ☐ | | 2016 | Upadhyaya, Snigdha, Ramakrishna, Vasudha **The Dimensions of Women's Empowerment in India**. . **Full Text Options:** PDF **Export Options:** RIS EndNote **Studies related to this publication:** • India Human Development Survey (IHDS), 2005 (ICPSR 22626) • India Human Development Survey-II (IHDS-II), 2011-12 (ICPSR 36151) *related studies/series* |
| ☐ | | 2017 | Gorski, Hannah P., Kamimura, Akiko, Nourian, Maziar M., Assasnik, Nushean, Wright, Lindsey, Franckek-Roa, Kathy **Gender roles and women's health in India**. *Public Policy and Administration Research*. 7, (7), 12-16. |

# India Human Development Survey (IHDS), 2005 (ICPSR 22626)

**Version Date:** Aug 8, 2018 Cite this study | Share this page

**Principal Investigator(s):**
Sonalde Desai, University of Maryland; Reeve Vanneman, University of Maryland; National Council of Applied Economic Research, New Delhi

**Series:**

- India Human Development Survey (IHDS) Series

https://doi.org/10.3886/ICPSR22626.v12

Version V12 (see more versions)

Download | Analyze Online

At A Glance | Data & Documentation | Variables | Data-related Publications | Export Metadata | Staff

More published analyses of the data in this study/series may exist. If you know of any not listed here, please suggest a new citation.

Showing 1 to 50 of 458 entries.

**Sort by:** Pub Date (newest) GO more options

| Type | Year | Citation |
|---|---|---|
| | 2019 | Akter, Sonia, Chindarkar, Namrata **An empirical examination of sustainability of women's empowerment using Panel Data from India**. *Journal of Development Studies*. 1-17. **Full Text Options:** DOI WorldCat Google Scholar **Export Options:** RIS EndNote *related studies/series* **Studies related to this publication:** • India Human Development Survey (IHDS), 2005 (ICPSR 22626) • India Human Development Survey-II (IHDS-II), 2011-12 (ICPSR 36151) |
| | 2019 | Mishra, Aswini K., Satapathy, Sadhana, Patra, Biswabas, Patro, Rajendra P. **Distributional change, income mobility and pro-poor growth: Evidence from India**. *Journal of the Asia Pacific Economy*. 1-18. **Full Text Options:** DOI WorldCat Google Scholar **Export Options:** RIS EndNote *related studies/series* |

**178,755** Downloads * Usage Report
* past three years

**458** Data-related Publications

## Notes

- The public-use data files in this collection are available for access by the general public. Access does not require affiliation with an ICPSR member institution.

DSDR Data Sharing for Demographic Research

This study was originally processed, archived, and disseminated by Data Sharing for Demographic Research (DSDR), a project funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD).

# Benefits data depositors, users, and funders

- **Enables the discovery and re-use of data . . .** via the data-related literature, which can be less intimidating and more accessible than the raw data and documentation.

- **Helps users decide what data will fit their needs . . .** by reading how others used the data, identifying cross-disciplinary implications and uses of the data, and avoiding duplicating analysis that has already been done.

- **Points to key research areas . . .** allowing ICPSR to point to trending topics being addressed with data in the literature.

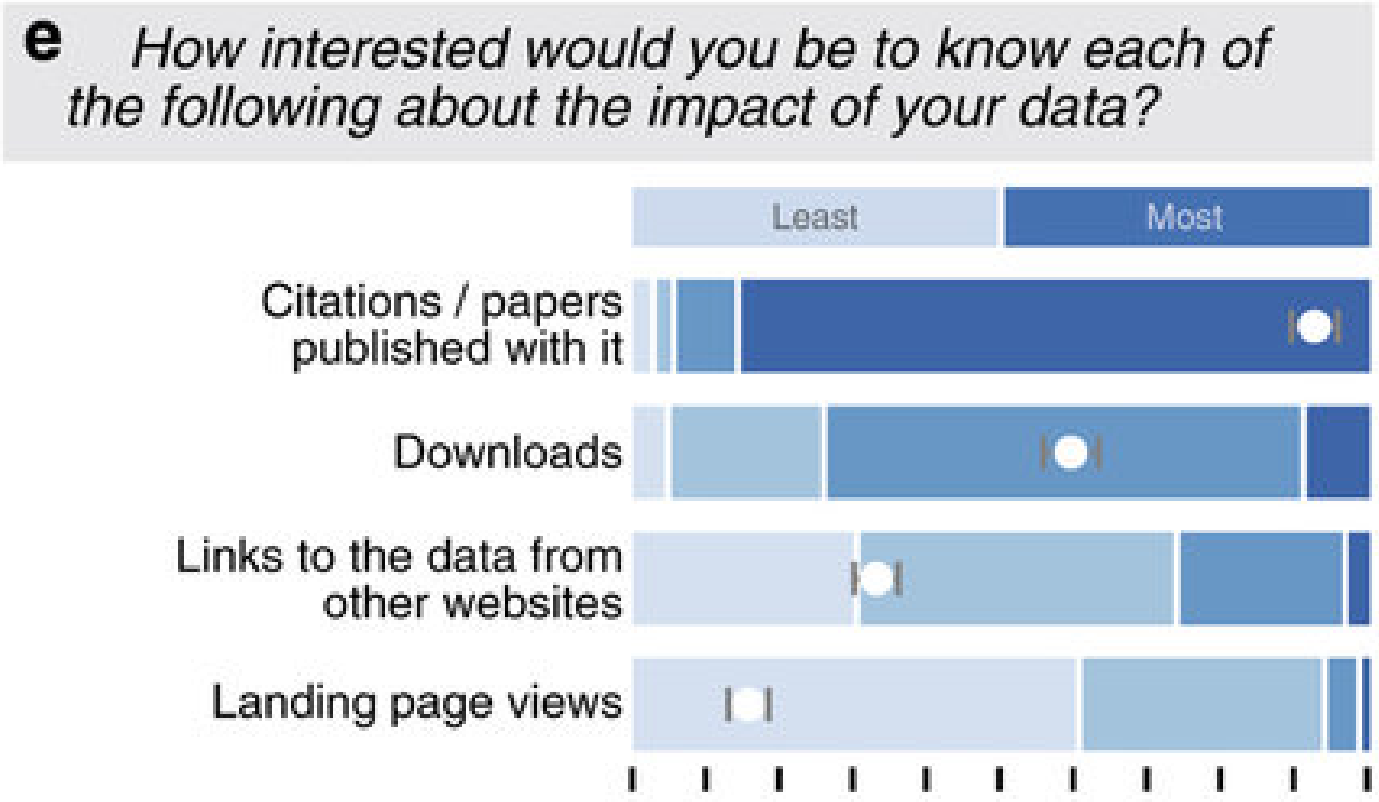
- **Encourages best practice . . .** by attempting to make visible all published data use, both cited with a DOI and informally cited, to ensure credit is given and to promote transparency.

Dodd, S. A. (1979), Bibliographic references for numeric social science data files: Suggested guidelines. J. Am. Soc. Inf. Sci., 30: 77–82. doi: 10.1002/asi.4630300203

"Conclusion:

Standardized procedures for bibliographic citations are designed to provide accurate and complete references, which in turn will be helpful to investigators and readers alike. It is hoped that guidelines or examples similar to the ones presented in this article will soon appear in the "authors' guide" section of the social science journals and will eventually be included in such works as the Chicago A Manual of Style and Kate L. Turabian's A Manual for Writers, etc. The ultimate goal would be to pave the way for social science data files to be included in printed bibliographies, end-of-work references, and indexing and abstracting works such as the Social Science Citation Index."

# Researchers (n=247) were asked: "How interested would you be to know each of the following about the impact of your data?




*White dots show the mean on a scale of one-to-four.

J Kratz and C Strasser. 2015. Making data count. Nature Scientific Data 2: 150039. dx.doi.org/10.1038/sdata.2015.39

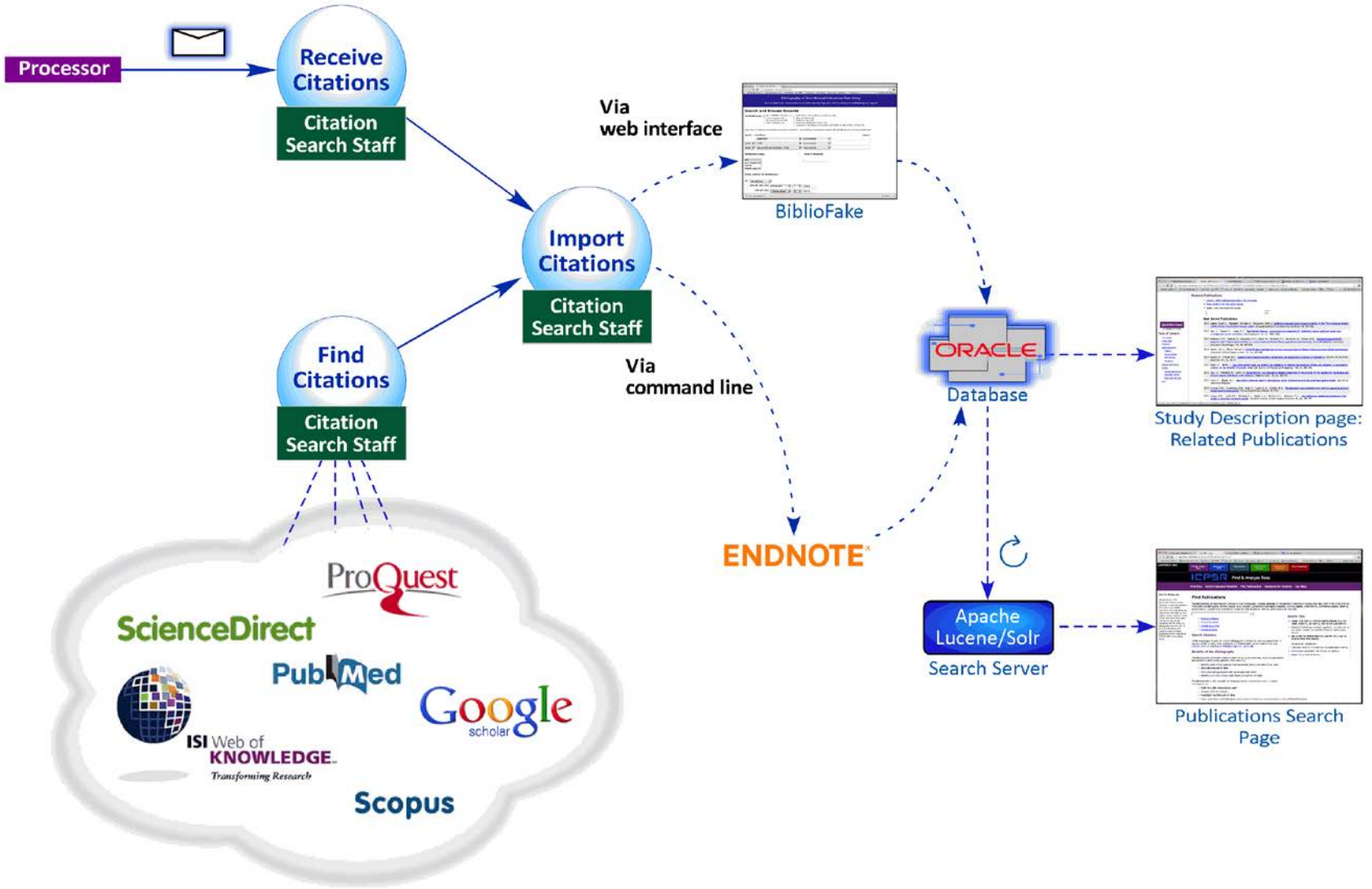
•**Downloads:** Download counts, on the other hand, are both highly valuable and practical to collect. Downloads were a resounding second-choice metric for researchers and 85% of repositories already track them.

•**Citations:** Citations are the coin of the academic realm. They were by far the most interesting metric to both researchers and data managers. Unfortunately, citations are much more difficult than download counts to work with, and relatively few repositories track them. Beyond technical complexity, the biggest challenge is cultural: data citation practices are inconsistent at best, and formal data citation is rare. Despite the difficulty, the value of citations is too high to ignore, even in the short term.

https://datapub.cdlib.org/2015/08/04/2334/

# Challenges of tracking informal data citation

# Populating the ICPSR Bibliography



## References


Bachman, J. G., Johnston, L. D., & O'Malley, P. M. (1998). Explaining recent increases in students' marijuana use: Impacts of perceived risks and disapproval, 1976 through 1996. *American Journal of Public Health, 88,* 887–892. doi:10.2105/AJPH.88.6.887

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design.* Cambridge, MA: Harvard University Press.

Brown, T. L., Parks, G. S., Zimmerman, R. S., & Phillips, C. M. (2001). The role of religion in predicting adolescent alcohol use and problem drinking. *Journal of Studies on Alcohol, 62,* 696–705. Retrieved from http://www.csa.com

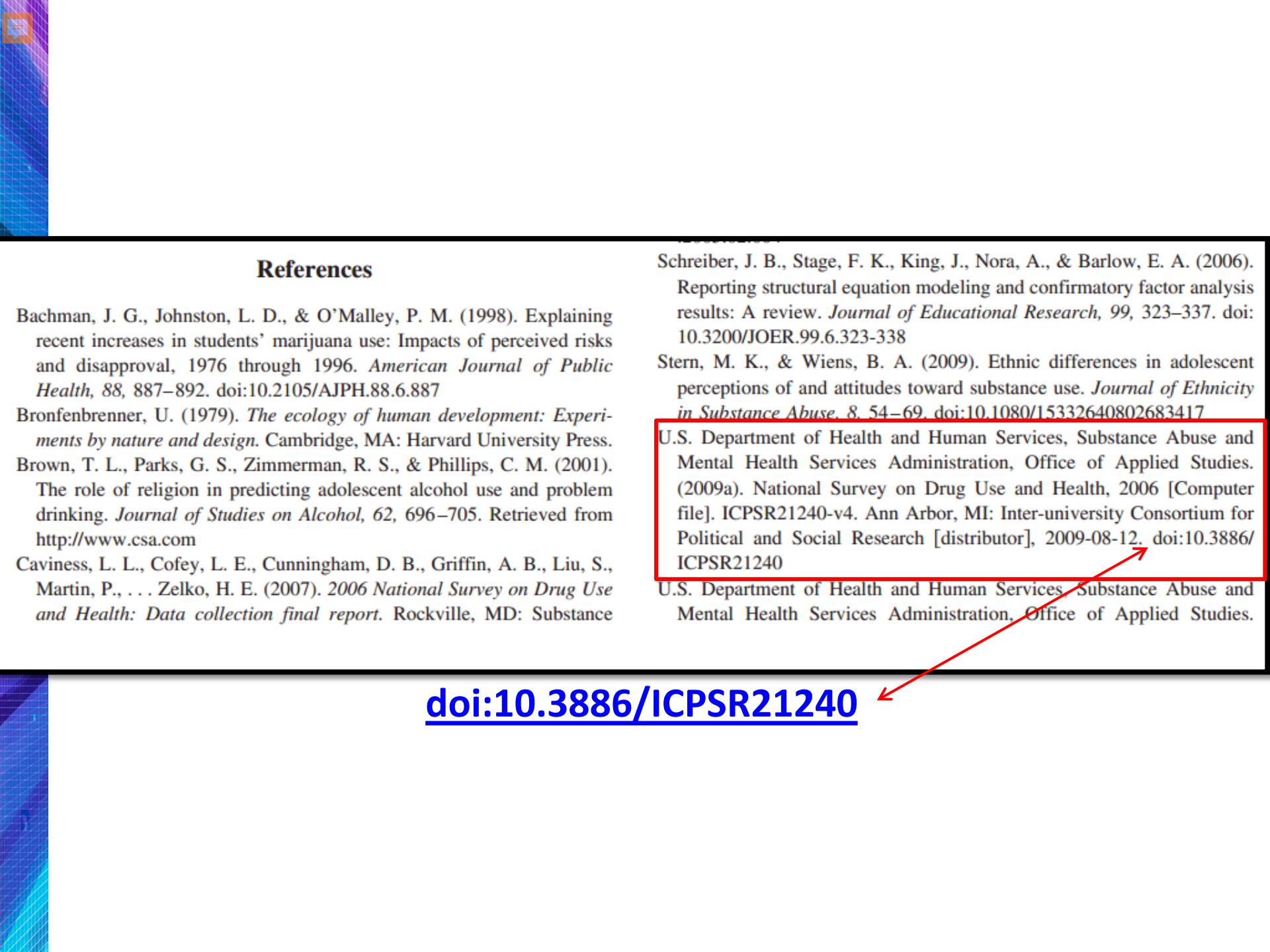
Caviness, L. L., Cofey, L. E., Cunningham, D. B., Griffin, A. B., Liu, S., Martin, P., . . . Zelko, H. E. (2007). *2006 National Survey on Drug Use and Health: Data collection final report.* Rockville, MD: Substance

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research, 99,* 323–337. doi: 10.3200/JOER.99.6.323-338

Stern, M. K., & Wiens, B. A. (2009). Ethnic differences in adolescent perceptions of and attitudes toward substance use. *Journal of Ethnicity in Substance Abuse, 8,* 54–69. doi:10.1080/15332640802683417

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies. (2009a). National Survey on Drug Use and Health, 2006 [Computer file]. ICPSR21240-v4. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2009-08-12. doi:10.3886/ICPSR21240

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

# In ScienceDirect

# But . . . tracking data use remains primarily a manual process

- Most authors do not include formal data citations with machine-readable, persistent identifiers.
- A 2018 study of biomedical literature by Park, You, and Wolfram found that “because only 62 of 513 observed instances of initial data reuse were included in the reference section of the examined articles, fewer than 20% of these citations are likely to be indexed in citation databases.”
- We risk not counting or linking 80% of data use if we do not track informal data citation.

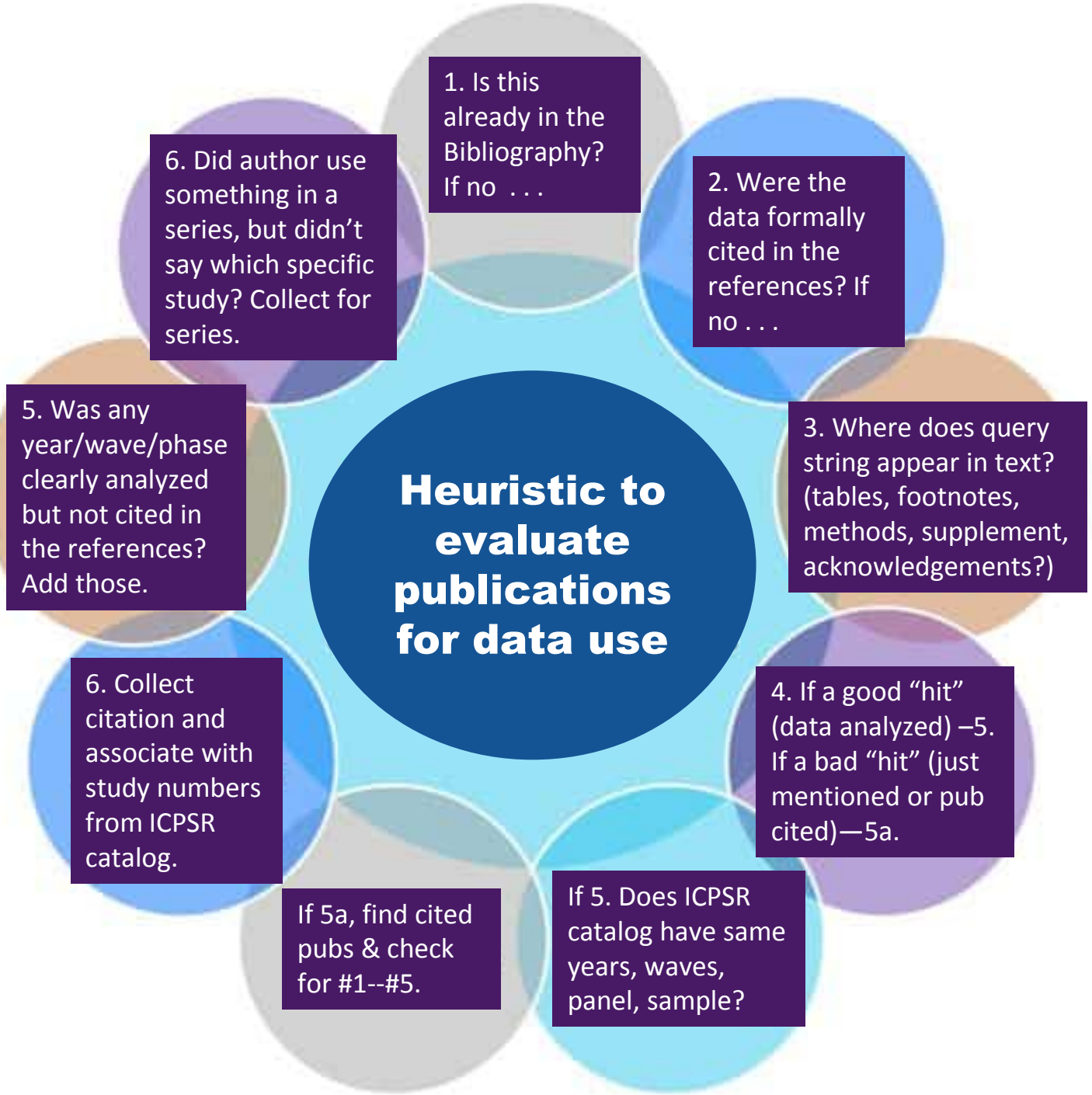
# How ICPSR finds and collects informally cited data-related publications

# Challenge 1 (Humans): Authors don't "cite data right"

- **We must create queries using the study name and other metadata.** Limiting our collection to only those publications in which ICPSR DOIs were cited would mean most data use would go unlinked and uncounted.

- **Imprecise by nature, these queries can bring back what we consider to be bad hits**. We do not collect data mentioned in passing. ICPSR only collects publications in which the data were analyzed, so distinguishing this must be done with human judgment.

- **Bad hits=time wasted.** Each hit needs to be evaluated, in multiple locations, e.g., the abstract, methods, acknowledgements, footnotes, references, etc.


Heuristic to evaluate publications for data use
1. Is this already in the Bibliography? If no . . .
2. Were the data formally cited in the references? If no . . .
3. Where does query string appear in text? (tables, footnotes, methods, supplement, acknowledgements?)
4. If a good "hit" (data analyzed) –5. If a bad "hit" (just mentioned or pub cited)—5a.
If 5. Does ICPSR catalog have same years, waves, panel, sample?
If 5a, find cited pubs & check for #1--#5.
6. Collect citation and associate with study numbers from ICPSR catalog.
5. Was any year/wave/phase clearly analyzed but not cited in the references? Add those.
6. Did author use something in a series, but didn't say which specific study? Collect for series.

# Challenge 2 (Automation): No Adequate Tools Exist

- **Data repositories do not have the resources.** Manual searching is burdensome for any large and growing collection. In ICPSR's case, we grow by over a hundred new studies a year.

- **New automated linking tools are for PIDs only.** Publishers and CrossRef/DataCite collaborate only on connecting PIDs together, not informally cited objects.

- **No good solution for automation at scale.** Currently APIs can make comparisons, collection to collection, but that is not enough to be efficient, i.e., high precision with high recall.

# Challenge 2, cont'd: No Adequate Tools Exist

- **You can automate targeted searching at scale, but not evaluation at scale.** ICPSR could use APIs to search at scale already. Potential matches based upon narrow identifier criteria (DOI, grant number, ICPSR study number), or a broad textual query (combinations of study title, investigator name, time period, geography, etc.) would act as a filter to avoid, for instance, evaluating every article in ScienceDirect against every ICPSR study**. But this would just leave a lot more hits to be evaluated.**

# Specific types of informal data citation in current use

# Almost complete

Data are formally cited in the references, but the PID is not included, which would have enabled machine-actionable detection and linking.

US Dept. of Justice, Federal Bureau of Investigation. (2003). Uniform Crime Reporting program data [United States]: Arrests by age, sex, and race, 2003 [Computer file]. ICPSR04285-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 2007-03-21.

# Indirect

A title and year of the data are mentioned in the methods, so it is clear data were used and even named, but no formal reference is made to the data or where they may be accessed. Often, the reference provided is to another publication and not directly to the data.

"Our analysis draws upon the India Human Development Survey 2005 (IHDS) collected by the University of Maryland and the National Council of Applied Economic Research (NCAER) in 2005."[46]

[46] Desai S, Dubey A, Joshi BL, Sen M, Shariff A, Vanneman R. Human Development in India: Challenges for a Society in Transition, New Delhi, India: Oxford University Press, 2010.

# Mismatched

When the data are archived, the title may differ from the title used in publications prior to archiving.

*You see it in the literature this way:*

"Data analysed from WHO World Health Survey (WHS)."

*Title used when archived:*

WHO Study on Global AGEing and Adult Health (SAGE): Wave 1, 2007-2010 (ICPSR 31381)

# Barely there

Authors may mention the name of a data collection, and may even provide the specific years analyzed. But they do not include the PID provided by the archive, let alone the version number. Not enough information is provided to know exactly where the author got the data.

"This study investigated the association between victim reporting and the police response to past victimizations with data from the National Crime Victimization Survey from 1998–2000."

# Vaguely described by necessity

The investigator writes about her own data well before having a citation and registered PID.

Or, sometimes, in sensitive areas of research, authors tend to say as little as possible about identifying characteristics of the data, even in broad terms.

"Data were collected from 1,342 men in First Offender programs in California (n = 996), Oregon (n = 77), and Nevada (n = 269). Men voluntarily participated in these deferred adjudication programs following their arrests for attempting to hire a prostituted woman on the street."

# Deducible with inside information

In some cases, even though no formal citation is used, an informed reader may know where the data were mandated to be deposited.

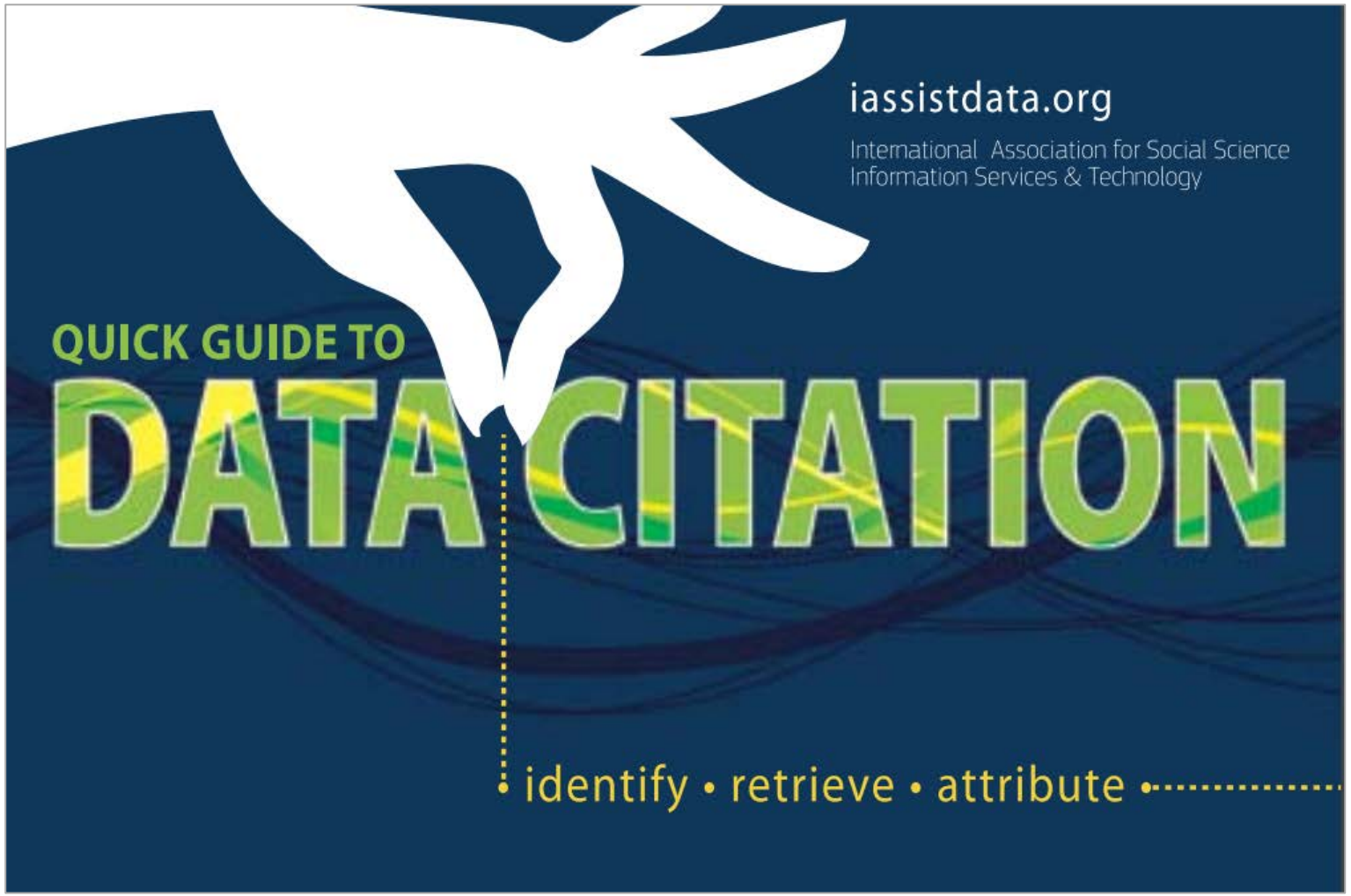
In other cases, authors acknowledge assistance from a repository. Or, only a staff member with knowledge of some aspect of the data would be aware of an identifying characteristic.

"These findings emerge from a two-year study that focused on Muslim-Americans in four communities: Seattle, Washington; Houston, Texas; Buffalo, New York; and Raleigh/Durham, North Carolina."

https://tuffpuppy.fandom.com/wiki/Mind_Reading_Helmet

# Suggested improvements

# Educate authors




https://iassistdata.org/community/data-citation-ig/data-citation-resources

- Most authors do not include formal data citations with machine-readable, persistent identifiers.
- A 2018 study of biomedical literature by Park, You, and Wolfram found that “because only 62 of 513 observed instances of initial data reuse were included in the reference section of the examined articles, fewer than 20% of these citations are likely to be indexed in citation databases.”
- We risk not counting or linking 80% of data use if we do not track informal data citation.

# Rethink education about data citation

**Many authors don't know they are supposed to cite data.** Failing to give credit to data creators has not been considered plagiarism, nor is there any ethical standard that is uniformly handed down across the social sciences in codified pedagogy.

**Many authors still don't know how to cite data.** Until recently, there has been no universally accepted standard way to do so, little support for it in the major style guides, and often journals' author instructions do not require that data be given attribution, let alone in the references.

# Rethink education about data citation

**The author was taught to acknowledge data by citing a report** or other written primary works, not the dataset, itself.

"…authors (and journals) are confused by being asked to cite their own data in their references. You don't cite your figures or tables, so why would you cite your data? Unless publishers and journals can re-educate the research community into always citing their own datasets, this approach seems unlikely to succeed."

https://scholarlykitchen.sspnet.org/2018/05/28/whats-up-with-data-citations/

# Emphasize use of DOIs

**The author did cite the data in the references section, but** either was prevented from using a DOI by the journal's style guide, or did not realize there was one to use.

# Change publisher practices

# Publishers invest in citation tagging

Devote extra resources at the typesetting phase to get the data citations right.

- Publishers pushing for the inclusion of data citations in the references, and tagging them appropriately at typesetting stage.

- In-text and data availability statements references to dataset DOIs being tagged as well, so that linkages between articles and their datasets are visible to Crossref, and authors can receive credit for the deposition of their data

https://scholarlykitchen.sspnet.org/2018/05/28/whats-up-with-data-citations/

# Incentivize publishers to require data citations

"Citations to a publisher's journals boost Impact Factors, and hence eventual revenue, so having typesetters carefully curate article citations has a commercial incentive…. no such incentive exists for open data – having excellent connections between datasets and articles doesn't have a clear path to future revenue. "

https://scholarlykitchen.sspnet.org/2018/05/28/whats-up-with-data-citations/

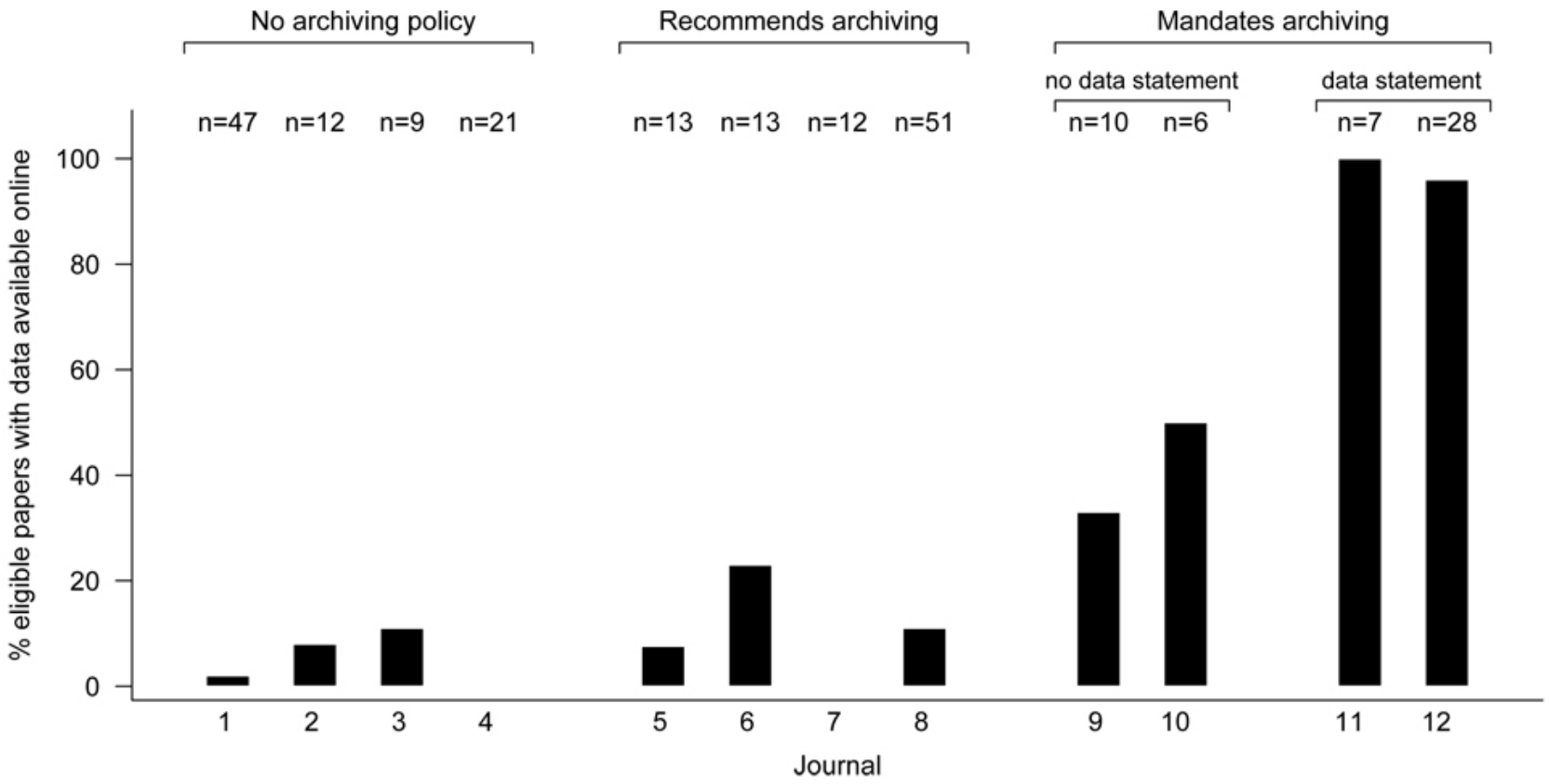
*The FASEB Journal* • Life Sciences Forum

# Mandated data archiving greatly improves access to research data

**Timothy H. Vines,[*,†,1] Rose L. Andrew,[*] Dan G. Bock,[*] Michelle T. Franklin,[*,‡] Kimberly J. Gilbert,[*] Nolan C. Kane,[*,§] Jean-Sébastien Moore,[*] Brook T. Moyers,[*] Sébastien Renaut,[*] Diana J. Rennison,[*] Thor Veen,[*] and Sam Yeaman[*]**

[*]Department of Biodiversity, University of British Columbia, Vancouver, British Columbia, Canada; [†]Editorial Office, *Molecular Ecology*, Vancouver, British Columbia, Canada; [‡]Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada; and [§]Ecology and Evolutionary Biology Department, University of Colorado at Boulder, Boulder, Colorado, USA

https://doi.org/10.1096/fj.12-218164




https://doi.org/10.1096/fj.12-218164

# Automation: Machine learning

What we need is a tool that can evaluate (via scoring) whether a given article is likely citing (formally or informally) ICPSR study(s).

**An evaluative algorithm would be:**

- **modular in nature** (something we would install on our systems)
- **able to accept** dataset metadata as one input, and the full text of the article as the other input, and return some sort of numeric score that reflects the accuracy of the match.

## An evaluative algorithm would be, cont'd:

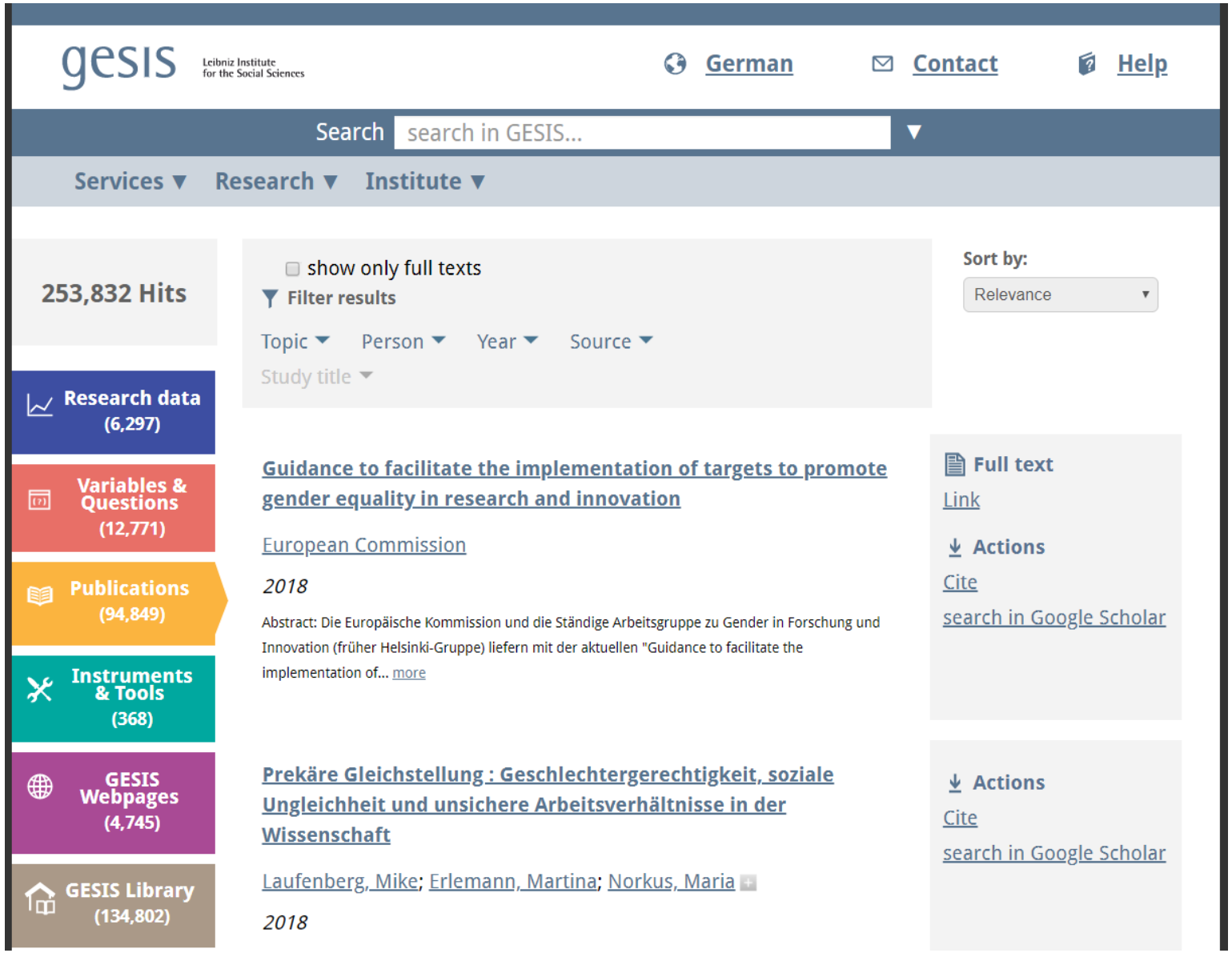
- **agnostic,** i.e., a repository's developers could write crawlers that comb through other collections of articles (subscribed to by their institutions) and point those matches at the evaluative algorithm.
- **flexible, enabling a repository to set and modify thresholds for inclusion/exclusion (**to control the volume of hits inherent in at-scale queries), e.g.:

  >90% automatically retain;

  60-90% flag for evaluation;

  <59% ignore entirely, though retain record so it doesn't pop in future searches

# Collaboration



# Don’t worry…change takes time

Interfaces
Vol. 38, No. 2, March–April 2008, pp. 125–139
ISSN 0092-2102 | EISSN 1526-551X | 08 | 3802 | 0125

informs

DOI 10.1287/inte.1070.0317
© 2008 INFORMS

## The Ombudsman: Verification of Citations: Fawlty Towers of Knowledge?

Malcolm Wright
Ehrenberg-Bass Institute, University of South Australia, Adelaide, South Australia, malcolm.wright@marketingscience.info
J. Scott Armstrong
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, armstrong@wharton.upenn.edu

https://pubsonline.informs.org/doi/pdf/10.1287/inte.1070.0317

# Suggestions?

Cite Data Right
https://www.icpsr.umich.edu/datacitation

# Thank you!

eammoss@umich.edu
lyle@umich.edu