AI, LLMs, and the Responsible Use of Public Research Data

A Journal (Data) Editor's Perspective

Lars Vilhuber

2026-11-10

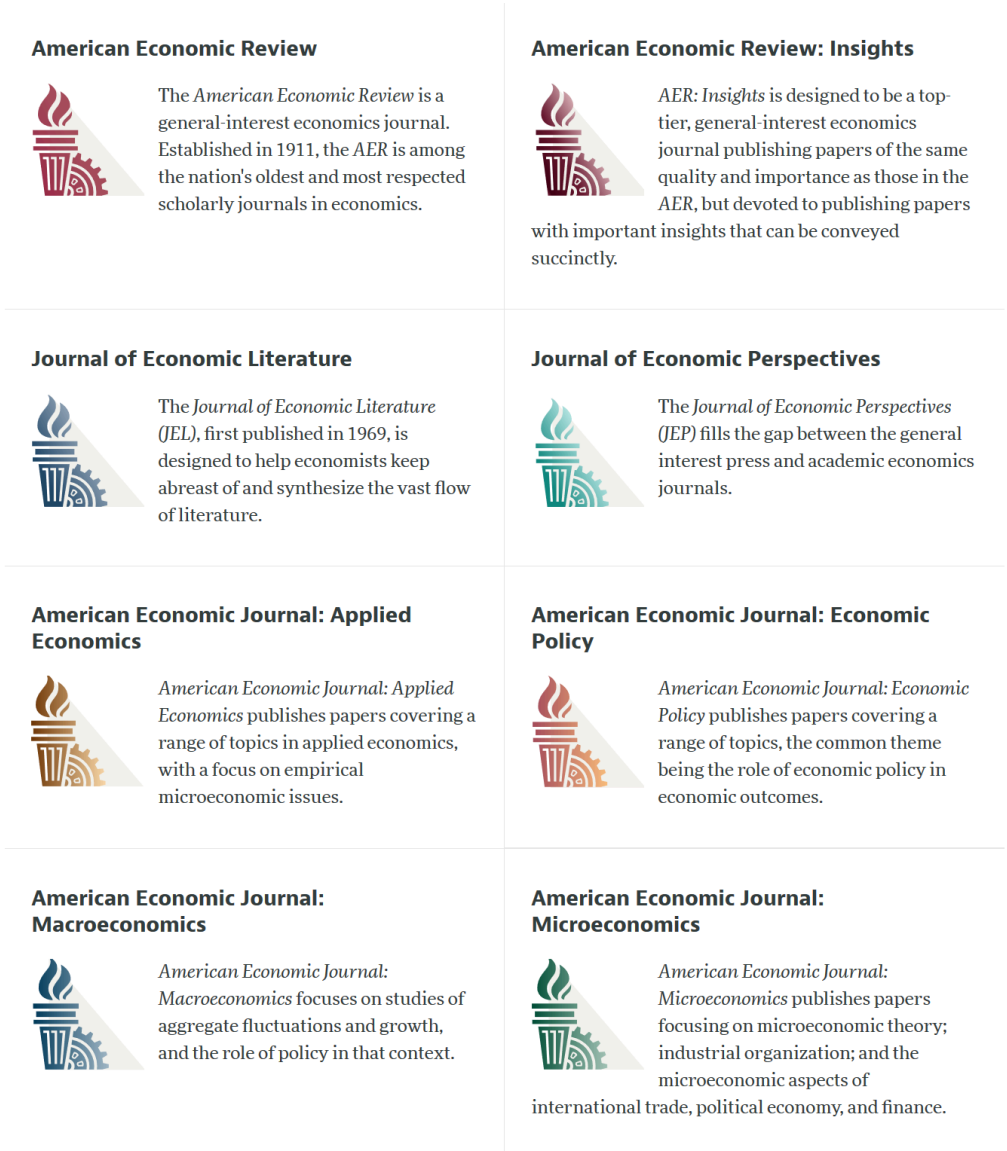
larsvilhuber.github.io/nasem-ai-public-data/



Who am I?

Lars Vilhuber

Economist at Cornell University, and the American Economic Association's Data Editor.



American Economic Review

The *American Economic Review* is a general-interest economics journal. Established in 1911, the *AER* is among the nation's oldest and most respected scholarly journals in economics.

American Economic Review: Insights

AER: Insights is designed to be a top-tier, general-interest economics journal publishing papers of the same quality and importance as those in the *AER*, but devoted to publishing papers with important insights that can be conveyed succinctly.

Journal of Economic Literature

The *Journal of Economic Literature (JEL)*, first published in 1969, is designed to help economists keep abreast of and synthesize the vast flow of literature.

Journal of Economic Perspectives

The *Journal of Economic Perspectives (JEP)* fills the gap between the general interest press and academic economics journals.

American Economic Journal: Applied Economics

American Economic Journal: Applied Economics publishes papers covering a range of topics in applied economics, with a focus on empirical microeconomic issues.

American Economic Journal: Economic Policy

American Economic Journal: Economic Policy publishes papers covering a range of topics, the common theme being the role of economic policy in economic outcomes.

American Economic Journal: Macroeconomics

American Economic Journal: Macroeconomics focuses on studies of aggregate fluctuations and growth, and the role of policy in that context.

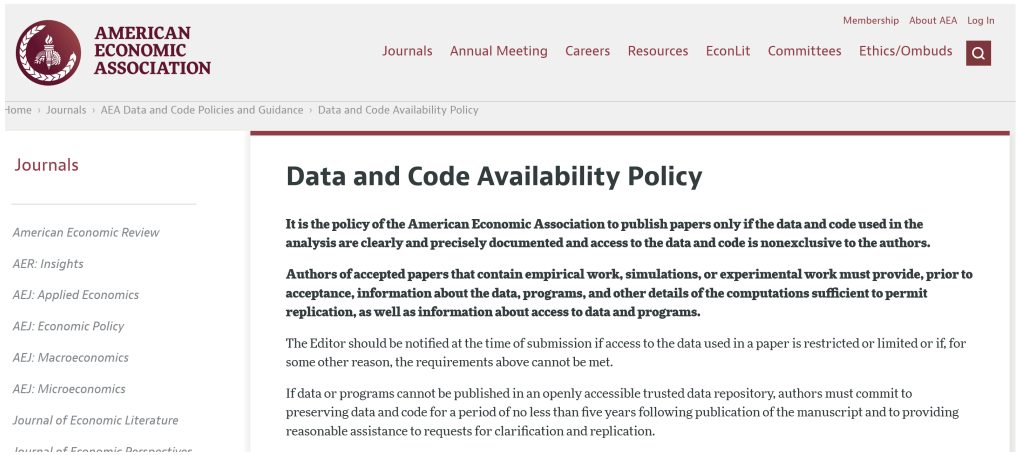
American Economic Journal: Microeconomics

American Economic Journal: Microeconomics publishes papers focusing on microeconomic theory; industrial organization; and the microeconomic aspects of international trade, political economy, and finance.

Journals

Data Editor of the AEA

~3000 Manuscripts and 4500 Reports, approx. 4800 authors reached.



DCAP

The question

> How can journals ensure scientific **integrity, transparency, and reproducibility** while also **protecting confidentiality** in the context of SI/AI-assisted research?
>
> What can journals **reasonably require** of researchers to promote responsible SI/AI-assisted research?

Three parts

1. What is **the same**

2. What is **different**

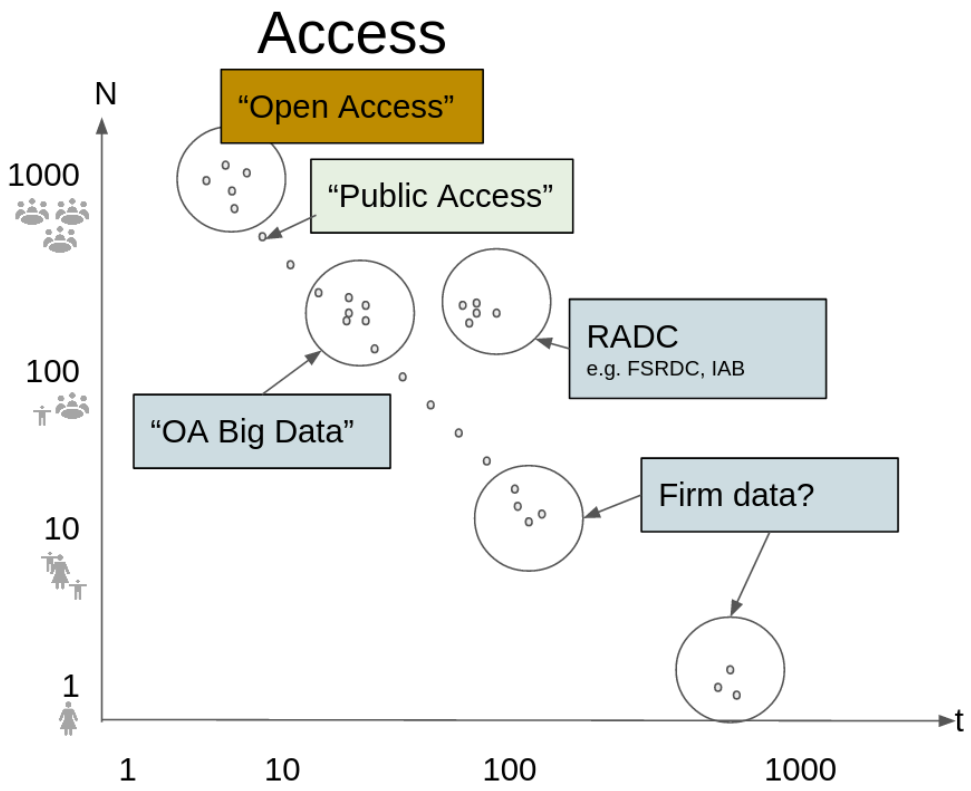
3. Possible **solutions**

Before we start

What are journals interested in?

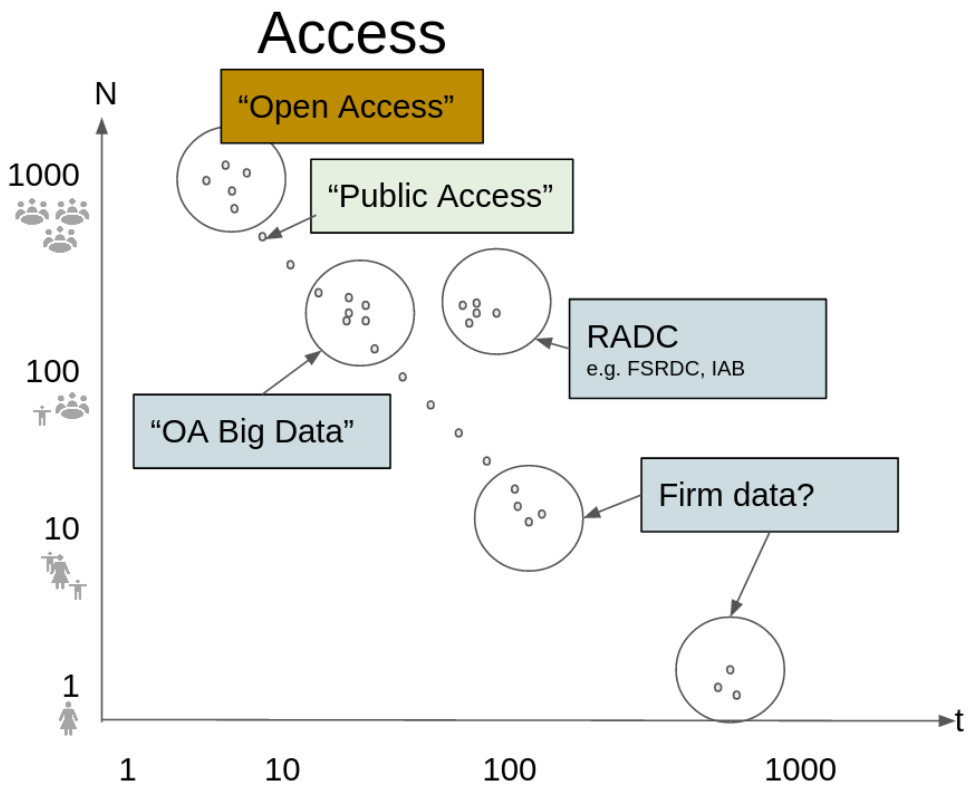
- **Primary**: Transparency for all
  - Data access details
  - All code and processing decisions
- **Secondary**: Reproducibility by many
  - Not all can run code, but many should be able to
  - Does not need to happen before publication

Access is a scale



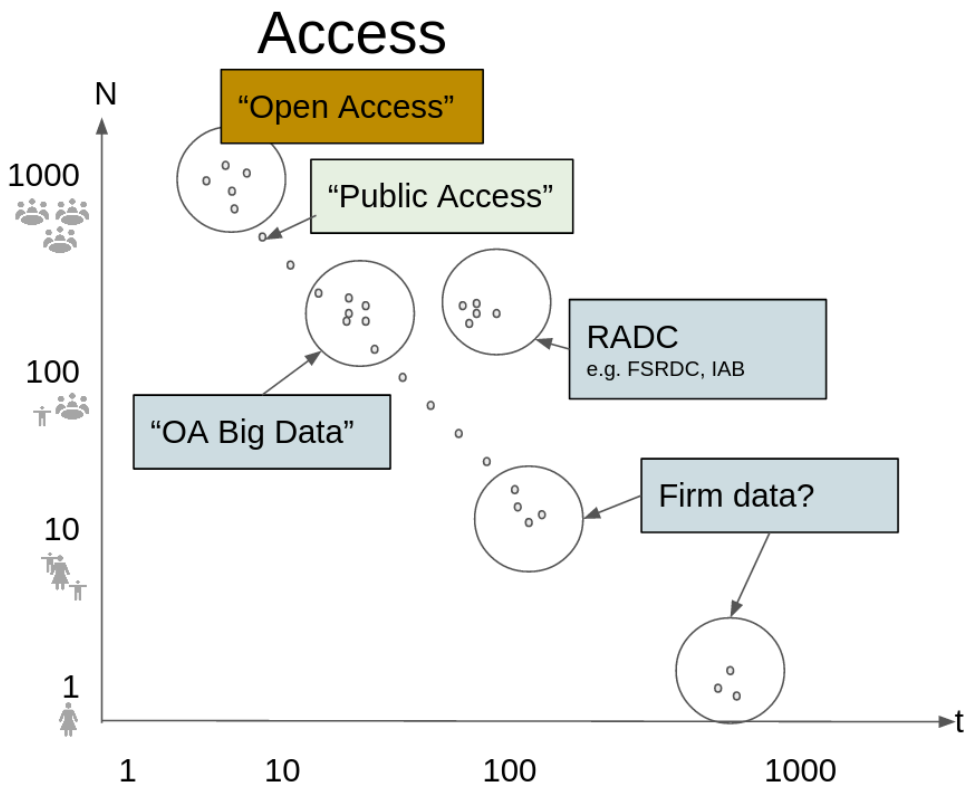


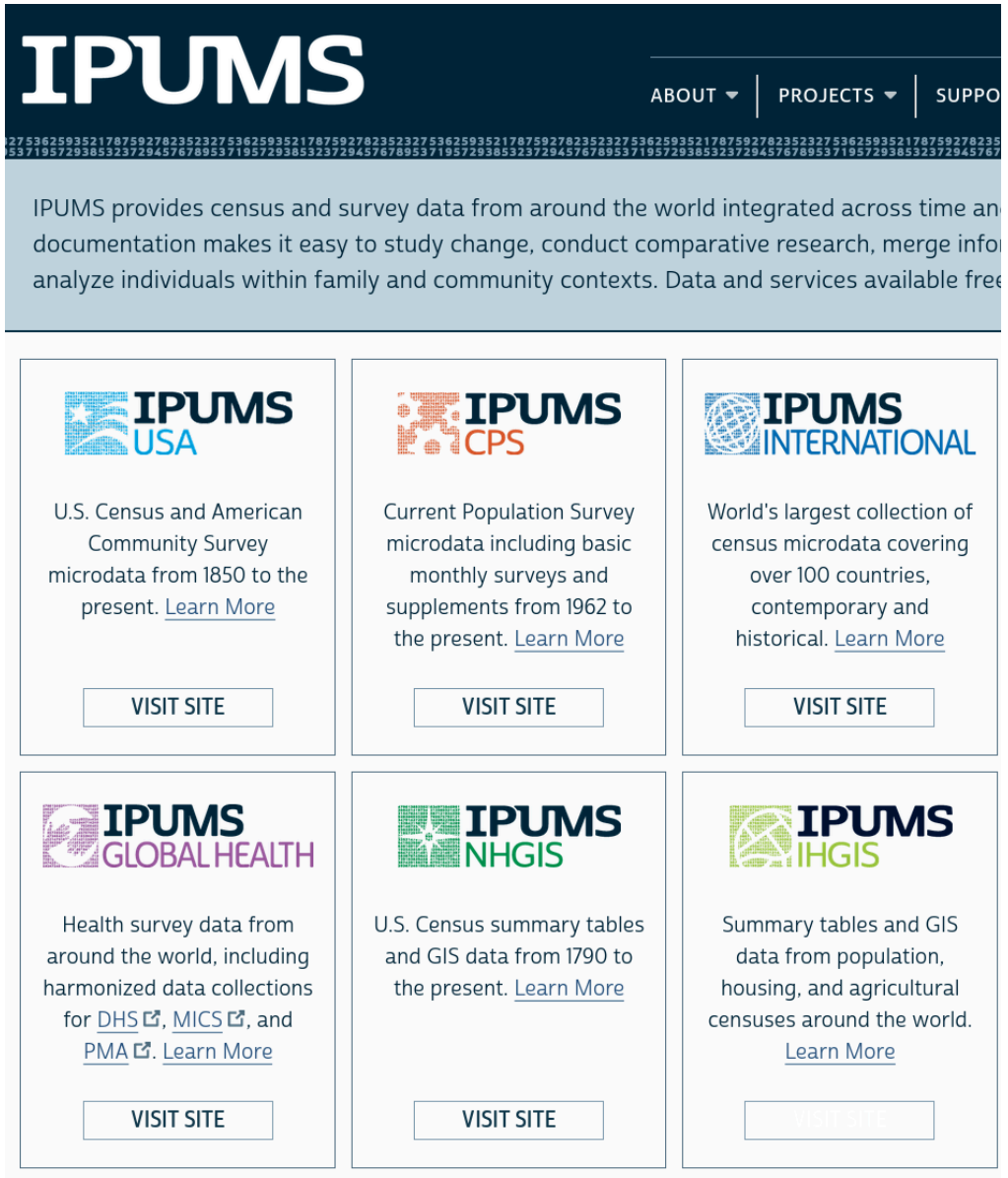
Access is a scale





An example of access at scale





What is the same

Straw-man arguments

> “AI changes everything about reproducibility.”

"LLM output is non-deterministic, so it cannot be reproduced"

SAME AS

Simulations, sampling, imputation, probabilistic record linkage

SOLUTION

Preserve the **draws**; report the variability

"LLM output is non-deterministic, so it cannot be reproduced"

SAME AS

Simulations, sampling, imputation, probabilistic record linkage

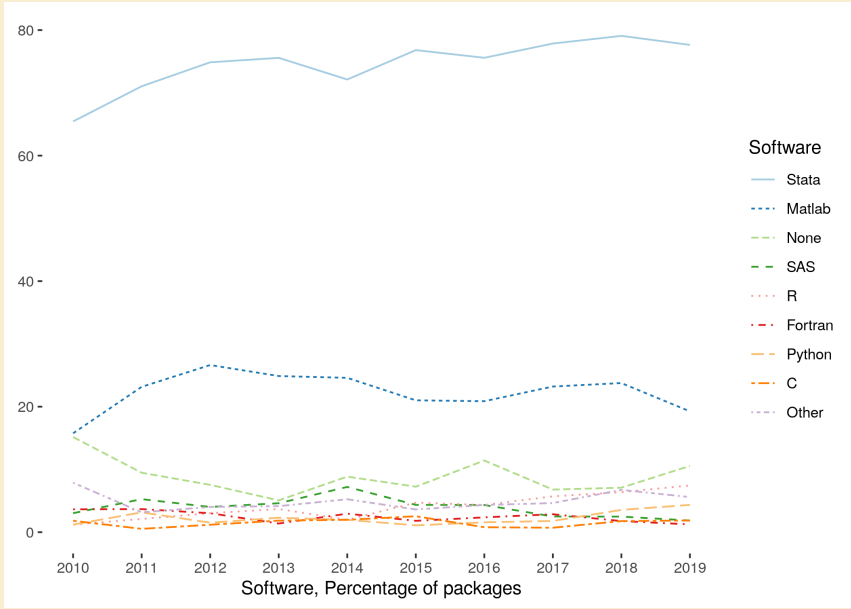
BETTER SOLUTION

Incorporate **variability** into your analysis (MI, robustness checks, etc.)

“The model is a black box”

SAME AS

Commercial software, proprietary compilers



“The model is a black box”

SAME AS

Commercial software, proprietary compilers

SOLUTION

Document **name, version, provider, date**

“The model is a black box”

SAME AS

Commercial software, proprietary compilers

BETTER SOLUTION

Show that the proprietary components are not **critical** for the outcome.

“Black box does not matter”

BETTER SOLUTION

Ex: Optimum found by quick proprietary solver is also optimum **using open source solver**

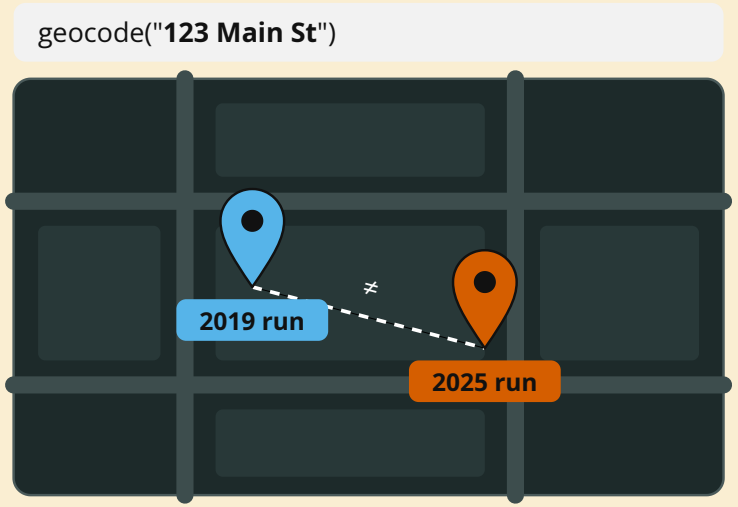
(BETTER) SOLUTION

Ex: **Open-weight LLM** obtains qualitative similar results as proprietary LLM

“The model will change or disappear”

SAME AS

External APIs fail/change, data, websites disappear, deprecated software



“The model will change or disappear”

SAME AS

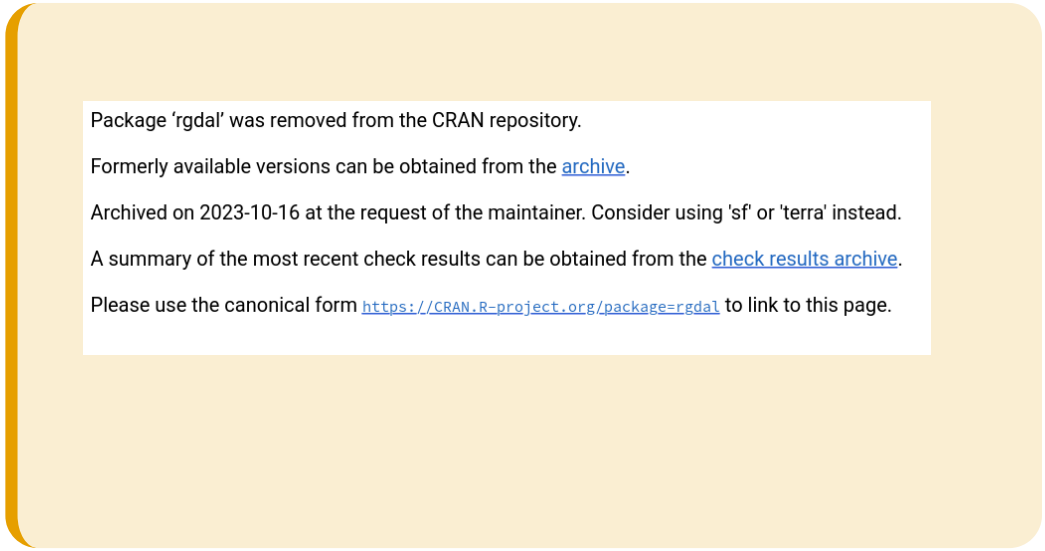
External APIs fail, **data, websites** disappear, deprecated software



“The model will change or disappear”

SAME AS

External APIs fail, data, websites disappear, **deprecated software**



Package ‘rgdal’ was removed from the CRAN repository.

Formerly available versions can be obtained from the archive.

Archived on 2023-10-16 at the request of the maintainer. Consider using 'sf' or 'terra' instead.

A summary of the most recent check results can be obtained from the check results archive.

Please use the canonical form https://CRAN.R-project.org/package=rgdal to link to this page.

“The model will change or disappear”

SAME AS

External APIs, websites, deprecated software

SOLUTION

Preserve the **outputs** you obtained; document **when** you accessed the model

“The model will change or disappear”

SAME AS

External APIs, websites, deprecated software

BETTER SOLUTION

Preserve the **software** and its environment

“The model will change or disappear”

SAME AS

External APIs, websites, deprecated software

(BETTER) SOLUTION

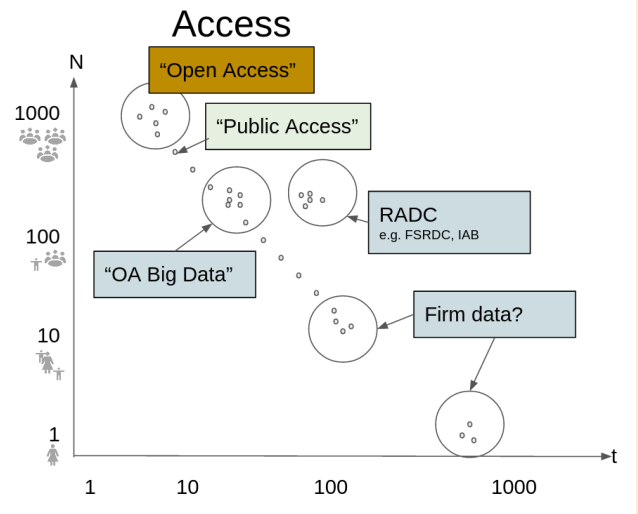
Preserve the **open weight LLM**

“Training data are confidential or proprietary”

SAME AS

62% of papers already use restricted-access data (Vilhuber 2026a)

SAME AS

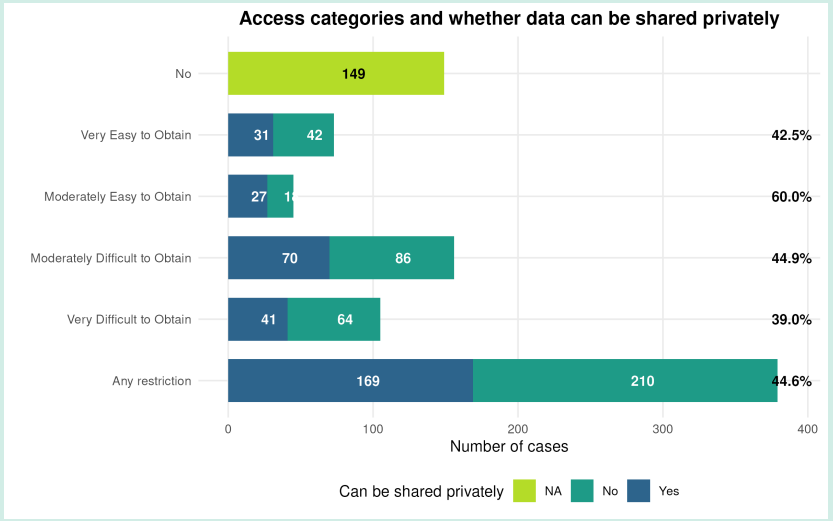


“Training data are confidential or proprietary”

SAME AS

62% of papers already use restricted-access data (Vilhuber 2026a)

SOLUTION



Data Editors may verify anyway (**45%** of those)

"Models (and their data) are too big to preserve"

SAME AS

Large data, **enclave** data, **sensitive** data

SOLUTION

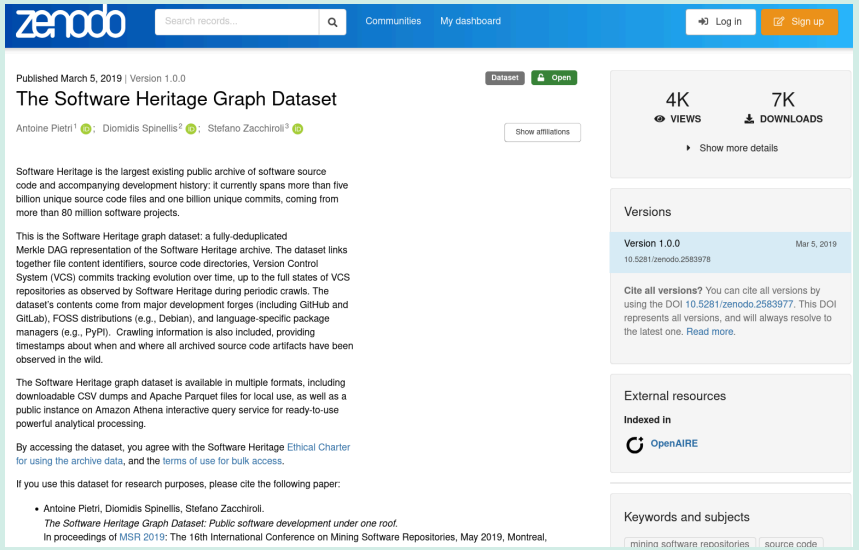
… still looking for easy solutions…

"Models (and their data) are too big to preserve"

SAME AS

Large data, enclave data, sensitive data

SOLUTION



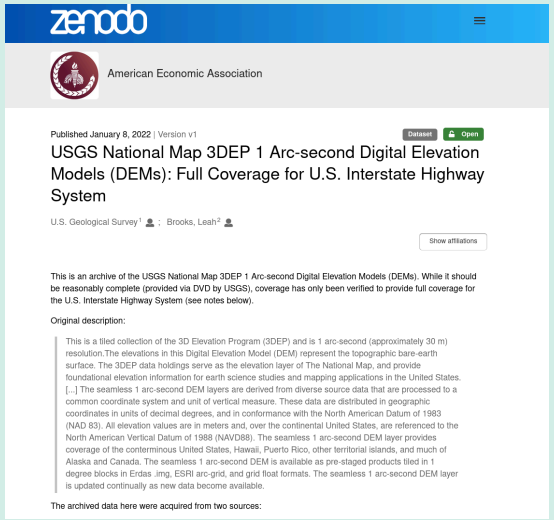
(Pietri et al. 2019) 480GB

"Models (and their data) are too big to preserve"

SAME AS

Large data, enclave data, sensitive data

SOLUTION



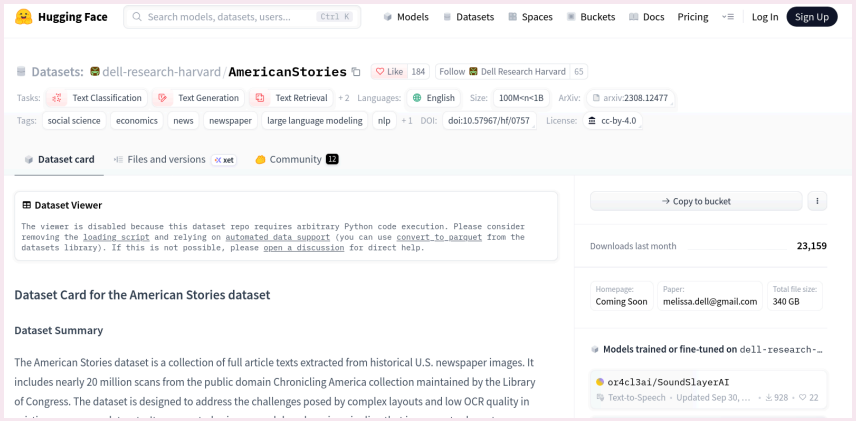
(Survey and Brooks 2022) 83GB, 3500 files

"Models (and their data) are too big to preserve"

SAME AS

Large data, enclave data, sensitive data

(BETTER) SOLUTION



(Dell 2023) 340GB

Preserve the **generated data**

"Models (and their data) are too big to preserve"

SAME AS

Large data, **enclave data**, sensitive data

SOLUTION

Hope that the enclave has a **preservation policy**.

"Models (and their data) are too big to preserve"

SAME AS

Large data, enclave data, **sensitive** data

SOLUTION

"If you collected the data, **preserve it for 5 years**." is not really a solution. There is no easy solution for self-collected sensitive data.

“Prompts are something new”

SAME AS

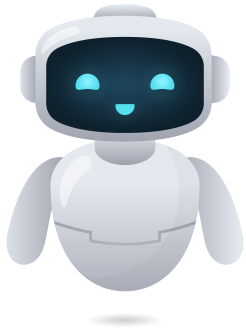
Prompts are **code** (or **metadata**)

SOLUTION

Include them, with the responses

So...

> None of this is AI-specific, but the difficulties may be magnified when using LLMs in research.



What is new

Four new challenges

1. **Scale**
2. **Access mechanisms** lagging behind
3. **Past practices** coming to light
4. Consequence:Shrinking pool of **publicly accessible** data

1. Scale

Bots everywhere

- Access requests are now easily scaled up
- Previously “freely” accessible resources are struggling with the bandwidth and access requests

Large data sets are increasing

WHY IT MATTERS

Traditional preservation methods struggle with **large data sets**

WHY IT MATTERS

… but more researchers are generating more data.

2. Access mechanisms

Open access

Publicly accessible =

- **public domain**, or
- **CC-BY**

WHY IT MATTERS

- In many cases, no formal **recording** of agreement/compliance
- In many cases, lack of understanding by **economists** of what those terms mean.

Controlled but free access: PSID

- Free, but requires **registration**
- Conditions of use: **no redistribution**

- Compliance expected
- Often ignored
- Costly compliance checks (**takedown requests**) [link]

… and AI raises new questions

- Free, but requires **registration**
- Conditions of use: **no redistribution**

WHY IT MATTERS

- Is uploading data to an AI service **redistribution**?
- Do terms of use **permit** AI-assisted processing?

PSID

- Free, but requires **registration**
- Conditions of use: **no redistribution**
- **Conditions of use: no use of AI**

⚠️

- Reduction in access and availability
- Increase in administrative burden
- Less use

3. Past practices coming to light

Compliance checks

- Enforcement of **terms of use**

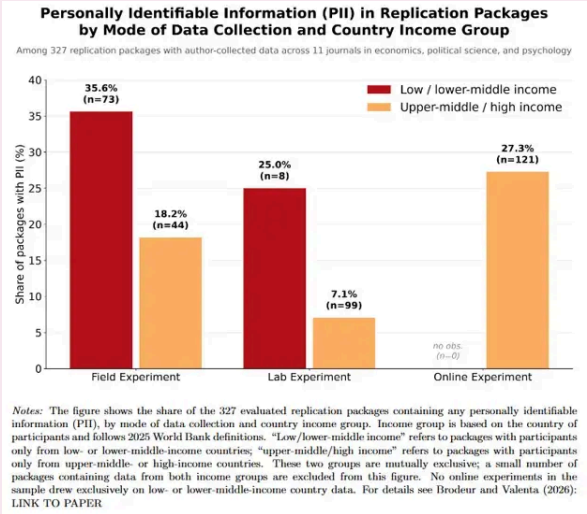
CURRENT

PSID sent requests to remove **PSID data** that had been included in **replication packages** (AEA: 24 cases in 2024)

AI makes scanning cheap

- Existing replication packages can be **scanned at scale**
- Detect **data that should not be there**

WHAT IS NEW

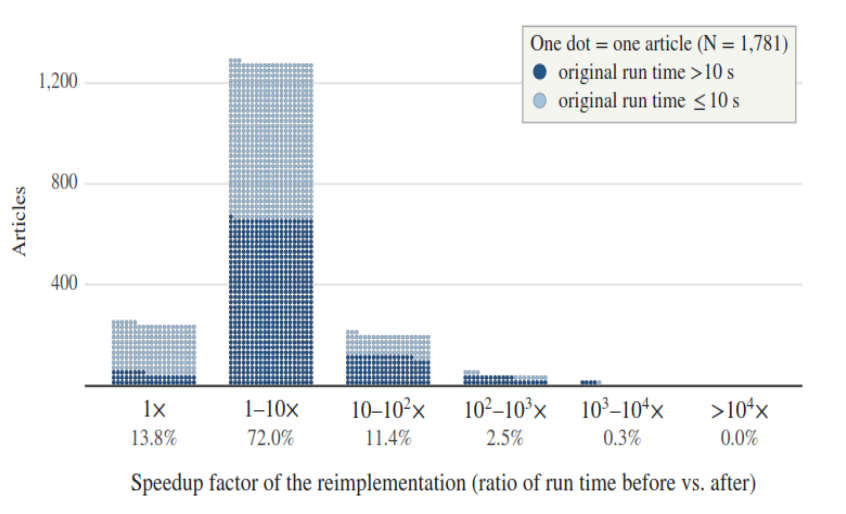


(Brodeur and Valenta 2026), 327 papers

AI makes scanning cheap

- Existing replication packages can be **scanned at scale**
- Detect **data that should not be there**
- Identify **poor programming practices**

WHAT IS NEW



(Schwartz et al. 2026) 1,781 papers

Caution: median run-time of 10s not plausible

4. Consequences

Consequences

ACTION ITEM

- Researchers have new tools to improve their practices
- Quality can be improved

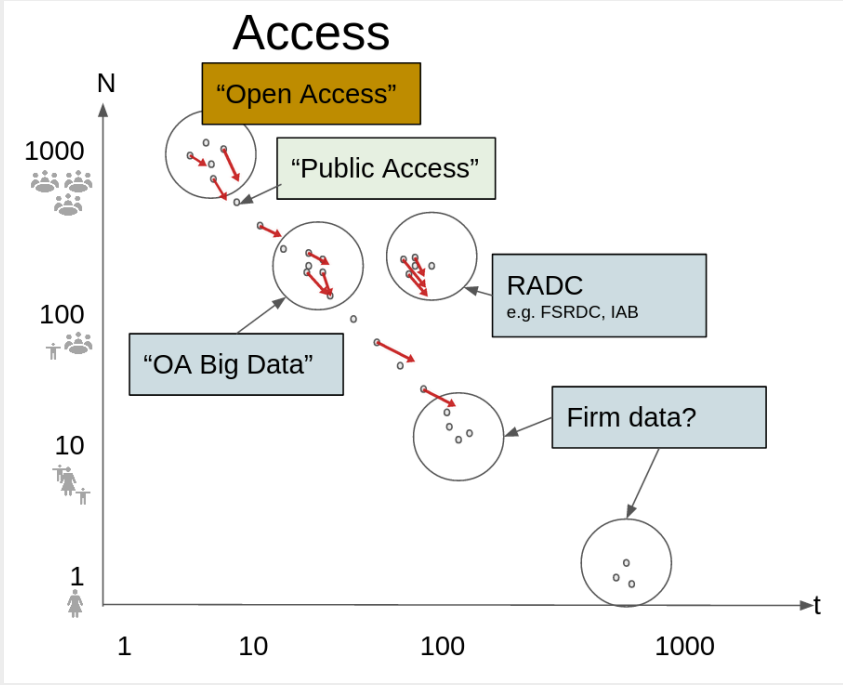


- Researchers increasingly reticent to share data
- Data custodians increasingly hesitant to share data

Risks

⚠️

- Researchers increasingly reticent to share data
- Data custodians increasingly hesitant to share data



Emerging solutions

Building blocks

CURRENT

- Federated university authentication systems
- NIH GA4GH Passports, ICPSR Researcher Passports (Levenstein et al. 2018)
- `eduroam`!

Building blocks

CURRENT

- APIs are … a thing
- tying APIs to **access rights** and **researcher identity**

Unified Access in Social Sciences

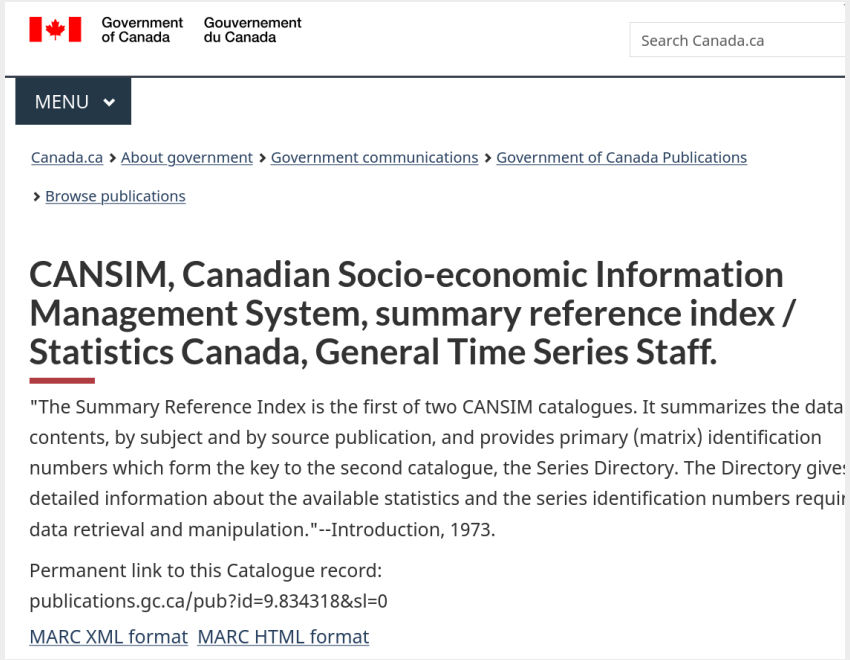
EMERGING SOLUTION

- Coordinate on a unified **researcher identity** across institutions and data repositories
- Use **existing** mechanisms (not “yet another standard”)
- tie it to **access rights** and possibly **flow control**

Caution

⚠️

- Not **too** many barriers!
- Not **exclusive** to academic researchers
  - must allow for other access (hard!)



CANSIM

Access mechanisms lagging

Access is still manual

CURRENT

- Web forms, click-through, manual download
- **No authenticated APIs** for most registered-access data
- Most economists don’t use APIs even when they exist

Access is still manual

WHY IT MATTERS

- Researchers **copy and redistribute** rather than re-acquire
- Increasingly bots **scrape**

Exceptions: IPUMS

Solution.

- IPUMS API: authenticated (**API key**)
- Extract definitions are **code** ⇒ reproducible acquisition (*mostly*)
- **API key** is contingent on accepting **terms of use**

Existing NIH practices

EMERGING SOLUTION

- **NIH** has various cloud-native platforms
- **Petabyte**-scale data, **standardized access protocols**

Solution.

- **Unified access**: GA4GH (Global Alliance for Genomics and Health)
- **Distributed computing**: commercial and academic clouds

Risk: the pool shrinks

⚠️

- Data providers do not act as archives
- Centralization creates single-points of failure



Other verification methods

Process transparency

EMERGING SOLUTION

TRACE = Transparency Certified



A framework that allows inquiry into the reproducibility workflow **at any stage** — without requiring re-running the code.

Key insight

EMERGING SOLUTION

Document the *process*, not just the *outputs*

- **File arrangements** (manifests with checksums)
- **Processing steps** (software, timing, method, isolation)
- **Cryptographic signatures** by certifying organizations

Why

WHY IT MATTERS

Transient resources, such as **deprecated commercial LLM APIS**

(BETTER) SOLUTION

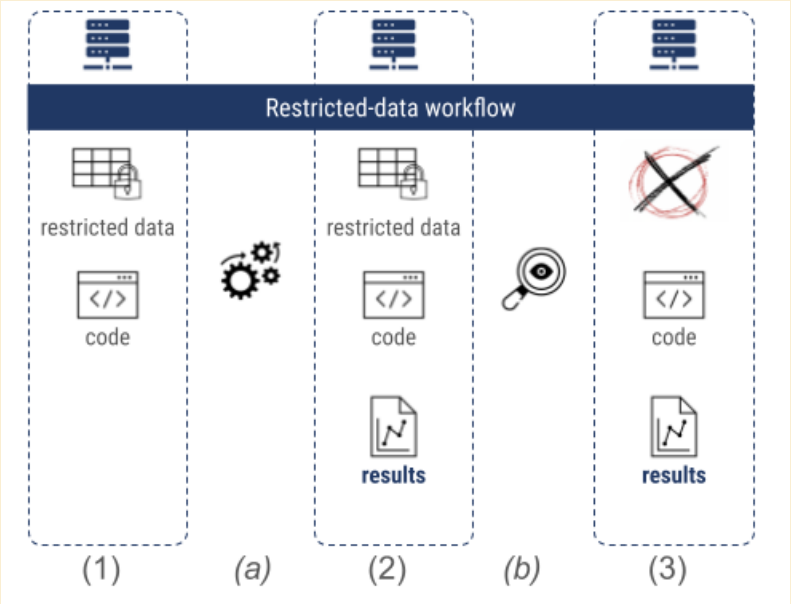
When data or services used in computation may no longer be available, TRACE provides a record of what existed at the time of execution.

Generic Workflow

CURRENT

Consider a researcher using confidential data in a Restricted Access Data Center (RADC):

1. Researcher gets environment with confidential data, writes code
2. Code is executed (a) → output produced
3. Data custodian inspects, removes confidential data (b)
4. Researcher receives code + output only (arrangement 3)



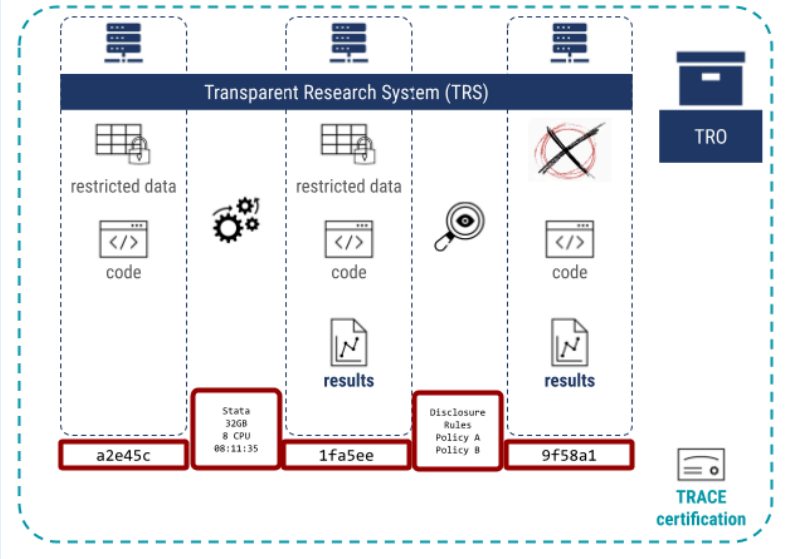
Generic workflow with confidential data

TRACE-Compliant

EMERGING SOLUTION

Add a few computationally easy steps:

1. Document each **file arrangement** (1, 2, 3) with manifests + checksums
2. Describe each **processing step** (software-driven or manual) with salient info



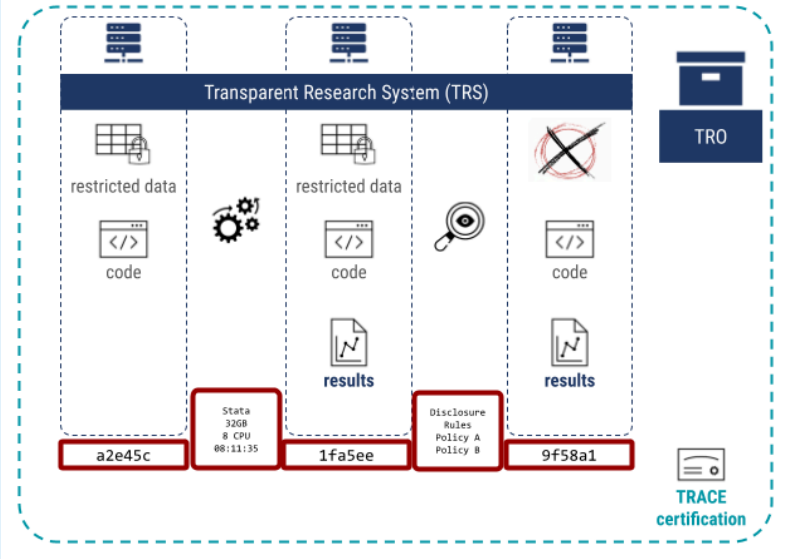
Example Trusted Research System

TRACE-Compliant

EMERGING SOLUTION

Add a few computationally easy steps:

3. Express in a **controlled vocabulary** (TROV)
4. Wrap everything into a package with a **cryptographic signature** → creates a **TRO**

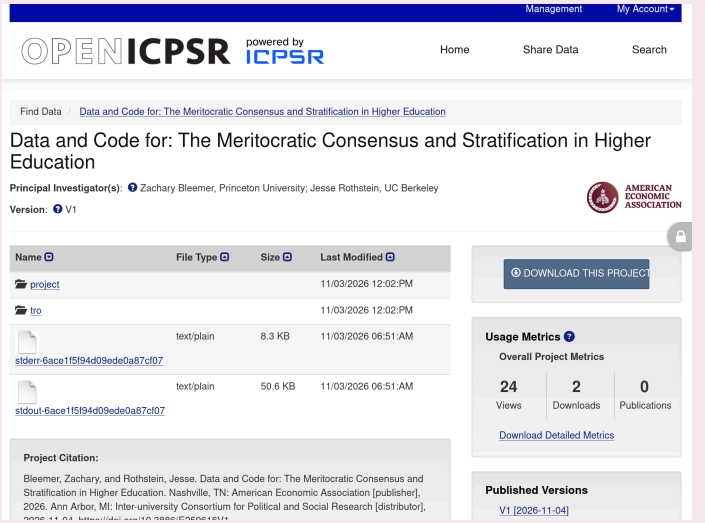


Example Trusted Research System

TRACE-Compliant

(BETTER) SOLUTION

- As of **2026-11-10**, there are 40+ TRACE-compliant packages published or in draft form in the AEA Data and Code Repository.
- Confidential data processes at some Restricted Access Data Centers (RADCs) are being adapted to comply with TRACE standards



(Bleemer and Rothstein 2026)

Journals and Transparency

Today

CURRENT

- Provenance and access conditions for **all** data (incl. data not included)
- **All code**, including survey code, experimental code
- **Generated data**, preserved in trusted repositories
- Documentation of how to **process** (technical, procedural)

Already includes

CURRENT

- prompts (**code**)
- responses (**API**, **generated data**)
- model identifiers and versions (**system documentation**)
- variability (**robustness**)

Concerns

- Ability to access resources **decreases** (expensive APIs)
- Less data available **under any access** condition

Conclusion

AI is not special

CURRENT

… for reproducibility

CURRENT

… for transparency

AI changes

WHY IT MATTERS

Scale, size of research



Risk of underprovision of archival resources

AI changes

WHY IT MATTERS

Programming tools

(BETTER) SOLUTION

- Allows for more complex **access** mechanisms (APIs at scale?)
- Allows for (in principle) **better** reproducibility

Changing Data Ecosystem

WHY IT MATTERS

- Less **open** data
- More **enforcement**

⚠️

- Lagging **access** infrastructure
- Reckoning with **past** practices

Other?

Questions

Useful links

- GitHub
-
- Presentation QR Code:

- Last Run on 2026-11-10

Content is licensed (cc) BY-NC.

References

Bleemer, Zachary, and Jesse Rothstein. 2026. *Data and Code for: The Meritocratic Consensus and Stratification in Higher Education*. ICPSR - Interuniversity Consortium for Political and Social Research. https://doi.org/10.3886/E259616V1.

Brodeur, Abel, and David Valenta. 2026. *On the Prevalence of Personally Identifiable Information (PII) in the Social Sciences*. Working {{Paper}} No. 320. I4R Discussion Paper Series. https://www.econstor.eu/handle/10419/343373.

Dell, Melissa. 2023. *AmericanStories (Revision 3484aca)*. Hugging Face. https://doi.org/10.57967/hf/0757.

Levenstein, Margaret C., Allison R. B. Tyler, and Johanna Davidson Bleckman. 2018. *The Researcher Passport: Improving Data Access and Confidentiality Protection*. May. https://hdl.handle.net/2027.42/143808.

Pietri, Antoine, Diomidis Spinellis, and Stefano Zacchiroli. 2019. "The Software Heritage Graph Dataset." Version 1.0.0. Zenodo, March. https://doi.org/10.5281/zenodo.2583978.

Schwartz, Matthew, Isaiah Andrews, and Jesse M. Shapiro. 2026. *An LLM Workflow*