Reproducibility and Open Science in Economics

Lars Vilhuber[1]

[1]Cornell University

March 18, 2025

This is a draft version for review purposes only. Please do not cite or distribute without author's permission.

1 Introduction

As a graduate student in economics at Université de Montréal, reading the economics literature was easy. While the main university library had all the relevant subscriptions, our department librarian, Fethy Mili, would populate the library of the economics department with multi-hued rows of working papers. Mili was also one of the key creators of what was initially known as Working Papers in Economics (WoPEc) and BibEc (for printed working papers) (Krichel 1997; Cruz and Krichel 2000; Krichel and Zimmermann 2009), populating the latter since 1993 (Bátiz-Lazo and Krichel 2012, p. 450). The overall network, known as Research Papers in Economics (RePEc), was born contemporaneously with the more widely known arXiv (Ginsparg 2011) and the more centralized Social Science Research Network (SSRN) (*Social Science Research Network* 2025). While electronic working papers were mostly free in those days (there was no way to pay for them), Mili's work consisted of sending out postage-paid envelopes to all of the various economics departments that were publishing the working papers, and then cataloging the incoming printed materials electronically, for public consumption. In the end, information about the existence of the working papers was freely available, but access to (printed) working papers still required a small fee to cover the cost of shipping.

I also experienced the openness of code sharing, with code samples by prominent authors being available to graduate students, though discovery was much more difficult at the time. The Statistical Software Components (SSC), primarily but not exclusively for STATA packages, appeared in 1998 (Cox 2010; Cox and Jenkins 2022), providing a convenient and open way to catalog, distribute, and provide open access to additional Stata functionality.[1]

Data sharing was harder, of course, with lots of floppies[2] being exchanged, but also the use of departmental FTP[3] servers (David Card's collection of data, or the NBER's), or even the replication archive of the Journal of Applied Econometrics, instantiated at Queens University in 1994 under the long-running guidance of James McKinnon.[4]. On the other hand, administrative data, such as the French administrative data used in AKM required travel to Paris, sitting in a room without windows at an assigned time, and typing the code into the system that had access. For non-local authors, this meant traveling for extended time periods as a Ph.D. student (Margolis) or spending a sabbatical in Paris (Abowd), neither of which is a cheap endeavor.[5]

When data were available, computing was straightforward: You logged on to the university's big computer, running some variant of Unix, and used whatever software was available. Software licenses were paid by the university, as was the computing hardware itself. Laptops powerful (and light!) enough to do actual work were only then emerging.

In 2025, there are concerned discussions about the cost of publishing academic articles, of accessing those same academic articles, of the ever increasing use of administrative data (Card et al. 2010b, 2010a; Chetty 2012; Einav and Levin 2014) that would appear to be hidden behind insurmountable access restrictions, the use of "proprietary software", and the increasing use of large computing infrastructure, all of which would seem to be restricting access to the basic elements of conducting research in economics. In this article, I will draw on my experience working on many aspects of increasing access to data and materials of all kinds, in particular my recent experience as the inaugural data editor of the American Economic Association (AEA) (Duflo and Hoynes 2018), to paint a picture of economics in an era of open science. How different are matters in practice now compared to that early view of the field of economics, back when I was a graduate student? In this article, I will discuss the current state of open science in economics as facilitated by and related to reproducibility. I will touch on the tension between accessibility, sharing, and preservation, and some of the approaches that are being implemented, sometimes tentatively, in economics, and sometimes elsewhere. My view will be biased - I am an active participant in this space, primarily via my current appointment as

[1] This kind of functionality was inspired by similar functionality available for other software, like CPAN for Perl and CRAN for R, but not for most statistical software used by economists.

[2] https://en.wikipedia.org/wiki/Floppy_disk

[3] https://en.wikipedia.org/wiki/File_Transfer_Protocol

[4] The JAE archive was migrated to the ZBW's archives in 2022 and can now be found at https://journaldata.zbw.eu/journals/jae, but legacy files are still visible as of 2024 at http://qed.econ.queensu.ca/jae/legacy.html.

[5] Both were, and still are, enjoyable, though.

data (and reproducibility) editor of the American Economic Association, but also as a past participant in networks that have and foster access, and a researcher and editor in the space of disclosure limitation.

The guiding theme will be the **accessibility** of the key ingredients for scholarship: manuscripts (or more generally, documents), data, software, and the necessary technology to combine the latter two in order to produce knowledge as published in manuscripts. My focus will be on the latter three, though I will provide some observations about scholarly publishing in the last section. In the conclusion, I will identify a few areas where there is (continued) movement towards greater openness.

2 Concepts

In order to write about "Open Science," a definition is needed. Open science is a surprisingly difficult term to define precisely. UNESCO (2022) sees four components to open **science**: *open scientific knowledge* (publications, data, code, and teaching materials "openly available, accessible and reusable for everyone"), open science *infrastructures* (which encompasses both physical infrastructure such as instruments and laboratories, as well as virtual components such as open access publication platforms), science *communication* (knowledge translation), and broad *engagement* beyond the boundaries of the academy. It also recognizes the limitations of such access in a caveat:

> ... human rights, security, personal privacy, ... In such cases, it may still be possible to share the existence of such information or share it among certain users who meet defined access criteria.

The Open Knowledge Foundation (Open Knowledge Foundation 2024, OKF) defines "open" as (my emphasis)

> ... anyone can freely access, use, modify, and share for any purpose *(subject, at most, to requirements that preserve provenance and openness)*.

Less broadly, Vicente-Saez and Martinez-Fuentes (2018) identify a consensus that defines open science as "transparent and accessible knowledge that is shared and developed through collaborative networks." In this article, I will focus on the what UNESCO 2022 calls open science "knowledge" and will briefly discuss "infrastructures." I will highlight how some elements have been quite widespread in economics for some time. I will try to identify limits to fully open accessibility, some of which are intrinsic to the nature of the research conducted in economics, and describe how widespread such limitations may be. In particular, I will highlight how those access restrictions are not, as many think, an impediment to **open** science, in the sense that aforementioned "collaborative networks" can still access these resources.

A key ambiguity will arise in how big such networks need to be in order to be considered "open." Clearly, $n = 2$ is not considered a network. The National Bureau of Economic Research (NBER) defines its affiliated scholars as a network: $n = 1804$ as of January 2025, primarily in North America (National Bureau of Economic Research 2025). However, in 2024, a total of 2,966 authors published 1,223 NBER working papers. J-PAL has approximately $n = 1725$ affiliates at 120 universities on all populated continents (Abdul Latif Jameel Poverty Action Lab 2025). Between 2001 and June 2024, there had been $n = 2023$ researchers on projects that used confidential U.S. Census Bureau in the Federal Statistical Research Data Centers (FSRDC) (U.S. Census Bureau 2024). In an average year (2013-2023), $n = 1238$ students graduate from a U.S. university with a Ph.D. in economics (National Science Foundation 2024, Table 1-5). In the approximately 30 years since inception of RePEc, $n = 526$ authors from 52 institutions in **Norway** have published a paper listed on RePEc (presumably in economics) (IDEAS//RePEc 2025). All of these are measured across different spatio-temporal dimensions. Are they large? Context and purpose matter. Some may intersect. How many U.S. graduate students in the past 10 years have published an NBER working paper and are now at a Norwegian institution?

Journals play a key role in this space, and will be an important background to my discussion and experiences, possibly also my biases. Most top economics journals have a data (and possibly code) availability policy.[6] The AEA's policy was first implemented in 2005 (Bernanke 2004; American Economic Association 2005). While the focus of early availability policies was on the data, the code often came along for the ride, albeit not always in its most complete form. I am the AEA's inaugural data editor, appointed in 2018 (Duflo and Hoynes 2018). The AEA implemented a new policy in 2019 (American Economic Association 2019, 2019). Many other economics journals appointed data editors around that time, and multiple journals coordinated on a common policy core called Data and Code Availability Standard (DCAS) (Koren et al. 2022), and revised their policies to align with DCAS, see American Economic Association (2024) for the AEA.[7] A key part of these newer policies was increased pre-publication monitoring of the content of replication packages (Duflo and Hoynes 2018; Christian et al. 2018).

One way to start to move away from a binary perspective of access is to consider **time** as summary metric that captures what is needed in order to access generic resources, whether data, manuscripts, or computing resources. Time might be needed to write an application to access data, or time might be needed in order to obtain access to large-scale computing resources. Time might be needed in order to obtain grant funding that allows to purchase such resources. I choose time, rather than money, as the metric, since it might appear to be slightly more egalitarian, given that much of science has (theoretical) access to subsidies and grants. In the other dimension, the number of people who have some probability of accessing the resource (the **size of the network**) can be taken as an approximate measure of openness. Figure 1, taken from Vilhuber (2023), serves to illustrate this idea, for access to data, with various institutions that facilitate that access mapped out into the space of time vs. size of network. I will return to this throughout the discussion.

3 Data Access

One subcomponent of open science, and locus of much attention throughout the literature in the social sciences, are "open data." This in principle easy - why should the data used in research not be open? However, the various caveats that policies and principles include are important to recognize. (Open Knowledge Foundation 2024) mentions "requirements that preserve provenance and openness," which does not take into account privacy. UNESCO (2022) does note "human rights, security, personal privacy." On the other hand, even much data that is available to almost anybody on the Internet may not actually be "open. " Consider the S&P 500, viewed in newspaper and many websites (e.g. S&P Dow Jones Indices LLC 2025), is not "open data" because it does not allow for free re-use. OKF defines "open data" as requiring machine readability, absence of licensing charges, and free re-use, but does not mandate availability via download on the internet, absence of all fees, nor absence of any technical measures, such as a requirement to register and agree to abide by these rules (Open Knowledge Foundation 2024).

I will discuss two sub-areas within this space: Secondary data use, and primary data generation. Much of economic research uses data collected by others, such as survey organizations and national statistical offices, but also private company data and various administrative data sources ("organic data", Groves 2011a, 2011b). Primary data generation is more frequent in behavioral and development economics.

Secondary data: The data produced by the United States government are in the public domain (i.e., without any restrictions on usage or attribution), which makes it "open" in the above sense (Copyright Act of 1976, Wikipedia 2025). Many countries have switched their government data to default to open data (Statistics Canada 2012; UK Government 2014). However, many well known "public-use" data are not always "open": IPUMS has a redistribution restriction (encapsulated in "Terms of Use", not a license), and some geographic data by international statistical offices remain under more stringent licensing requirements in other countries (e.g., United Kingdom).

Many well-known survey organizations impose redistribution and usage restrictions that are not consistent with the OKD definition. Many such redistribution restrictions apply to datasets with more detailed personal

[6] For a review of the history of data and code availability policies in economics, see Vlaeminck (2021).

[7] These initiatives are not restricted to economics, of course. Political science (*Data Access & Research Transparency (DA-RT)* 2014; Jacoby 2015; Basile, Blair, and Buckley 2023), sociology (Sociological Science 2018; Weeden 2023), and general initiatives like the Transparency and Openness Promotion (TOP) guidelines (Nosek et al. 2015).

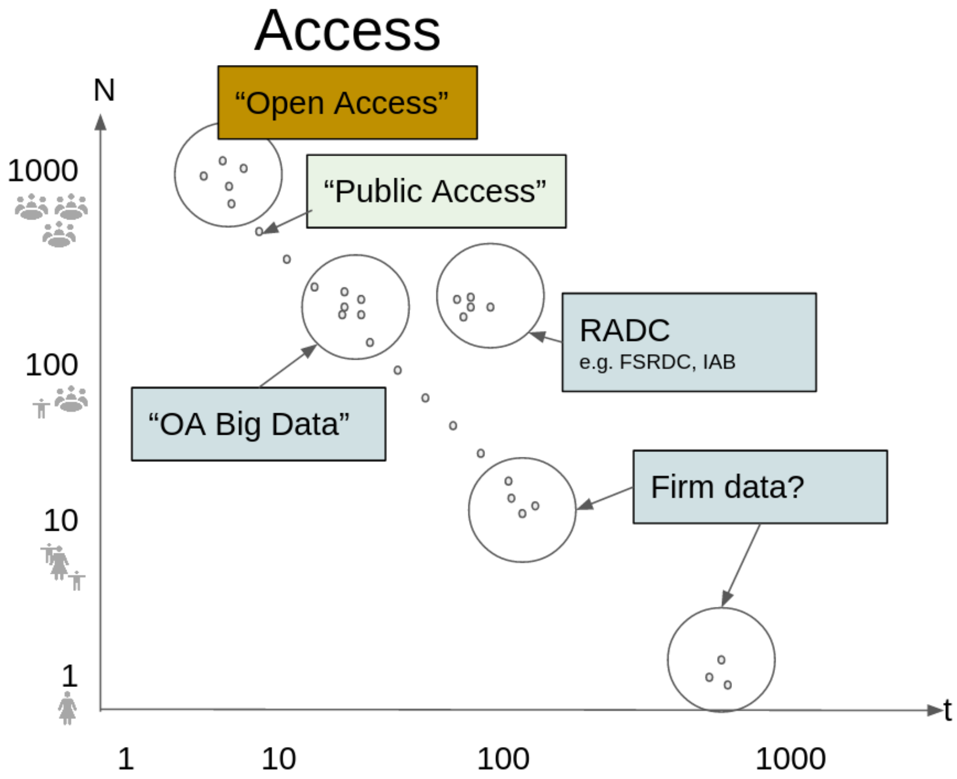

Figure 1: Conceptual trade-off between number of individuals accessing data, and time required to do so. Figure first published in Vilhuber (2023).

information collected through surveys, and are meant to ensure continued compliance with ethical rules of behavior, often with informed consent agreements by survey participants and local privacy laws. Notable examples include PSID (Institute for Social Research 2024), World Value Survey (Haerpfer et al. 2024), Demographic and Health Surveys (DHS) Program (DHS Program 2024), German Socio-Economic Panel (Goebel et al. 2019; Goebel et al. 2024), all of which have broad (cost-free) usage, conditional on registration and compliance with usage restrictions.[8] Table 1 shows a few examples.

Table 1: Example restrictions

Dataset	Variant	Usage.restrictions.and.justification
DHS		No redistribution and access by unauthorized individuals. To ensure that provider meets host-country agreements, such as compliance with usage by real people with legitimate research purposes.
WVS	Except for joint WVS-EVS files	The data can be used freely for non-commercial purposes such as research, publication, teaching. Data redistribution is prohibited: publication of the original WVS datasets at other online platforms is against the WVSA Constitution.
PSID	Public data	Use the data in the PSID data sets only for scientific research and aggregate statistical reporting; Make no attempts to identify study participants;
PSID	Restricted data	In order to safeguard the confidentiality of respondents at the highest level, some data are provided only under conditions of a restricted use contract, including human subjects review and data security plan.
SOEP	European version	No redistribution and access by unauthorized individuals. Compliance with German Federal Data Protection Act.
SOEP	International version	Same as European version, but 95% sample, and institutional agreement required.

The steps needed to obtain access range from click-through agreements to acknowledge compliance with CC-BY licenses (consistent with open access definitions) to the need to write a paragraph about the purpose on how the data is to be used (maybe consistent), to various commitments to not redistribute the data, all while being cost-free. The first three rows in Table 1 apply some variant of the latter: While anybody can obtain access, the restriction on redistribution is not consistent with definitions of open access. Yet there are no impediments to actually using the data in research.

The last three rows of Table 1, however, go further. Researchers using SOEP data must satisfy certain geographical requirements, such as presence of the researcher in the countries covered by General Data Protection Regulation (GDPR), otherwise they can only access a modified version of the data. Similar restrictions apply to the restricted-use data of many other surveys, for instance, the US-based National Longitudinal Survey of Youth (NLSY). Users of restricted PSID data must comply with even more stringent rules: secure approval from their institution's ethics board, and use of secure computing environments. Clearly, these restrictions will inhibit broader use of the data per se. Yet these restrictions are not globally very restrictive: There are in Economics alone 1238 new US-based researchers being given the ability to request access to geo-restricted NLSY data every year: newly minted Ph.Ds, as noted earlier. A similar number in the EU likely obtain the right to access the geo-restricted SOEP data as well. Despite the restrictions on the use of SOEP data, there are currently 6,443 papers that in some fashion have used the data.[9] The PSID bibliography[10] lists over 1,300 dissertations and over 5,300 articles. In the space depicted

[8] I come back to the case of the DHS Program in Conclusion.

[9] Source: https://www.diw.de/en/diw_01.c.789503.en/publications_based_on_soep_data__soeplit.html accessed on 2025-02-08.

[10] Source: https://psidonline.isr.umich.edu/publications/Bibliography/search.aspx accessed on 2025-02-08

in Figure 1, access is very much towards the left of the figure.[11].

Organic data: However, most economists, when asked about data subject to restrictions, will think of "proprietary data," a term often applied to any data that may be subject to restrictions of use and re-use. Access to most administrative data is typically not "open" in the sense of OKD, but are they open enough, given the privacy concerns that are attached to these data? Many have argued that access is not broad enough (Card et al. 2010b; Einav and Levin 2014), while acknowledging the difficulty of addressing privacy and security of the data at scale. Abowd and Schmutte (2018) discuss the challenge of making the choice of between accuracy of (public) statistics and data, and the privacy loss inherent in doing so. Nagaraj, Shears, and Vaan (2020) argue that more openness improves scientific progress, and Nagaraj and Tranchero (2023) study this in the context of the US system for providing access (FSRDC). They note that 4% of US-based empirical authors have had some access to the FSRDC system. I am not aware of similar studies for other countries, such as France (the equivalent system is the Centre d'accès sécurisé aux données (CASD), Gadouche 2019) or Canada (Currie and Fortin 2015). In absolute terms, these networks host several hundred researchers every year. As per (CENSUS), 2,084 researchers have had access to projects involving Census Bureau data since (YEAR).[12] Nagaraj and Tranchero (2023) mention 861 papers. Similar numbers can be obtained for the French (xxx researchers and 417 publications) and Canadian (2201 active researchers as of January 2025, and 3245 papers published between 2000 and May 2024)[13] networks.[14] The number of publications from access to these networks is smaller in absolute terms then those from PSID and SOEP, though likely higher in impact (Nagaraj and Tranchero 2023).

Primary data collection: The discussion of choices made by survey organizations should in principle be applicable when smaller teams of economists, not entire survey organizations, do the primary data collection. Similar to survey organizations, such teams have to balance the privacy of their respondents with the benefits of open science, in particular the broader knowledge to be gained from open access to the data. Many, so it would seem, provide much of the data in replication packages, subject to de-identification (see Kopper, Sautmann, and Turitto 2020; Bjarkefur et al. 2021, for examples), though typically not with stronger disclosure avoidance measures similar to those employed by statistical agencies and larger survey institutions (for a brief discussion of the issues and one possible solution, see Mukherjee et al. 2023). For research teams, ethics boards and institutional review boards (IRBs) have a role to play (Grant and Bouskill 2019), with some arguing very strongly that greater availability to others (though not blind publication of all data) is required in order to maximize the societal benefits that are the *quid pro quo* for the respondents' consent to their privacy being invaded (Meyer 2018; Grant and Bouskill 2019). Making such data as broadly available, while respecting the privacy of respondents, is precisely what open access to such data promises, *modulo* appropriate access restrictions or data use agreements similar to those outlined in Table 1. In general, however, primary data collections do not have access to robust third-party systems that would allow for access similar to the access required by PSID and similar organizations, situated between no access and fully public access. Thus, while access may be requested in ad-hoc fashion via the original authors, this is known to be fraught with problems (Watson 2022; Gabelica, Bojčić, and Puljak 2022). An ideal scenario would see researchers deposit the data they collected in third-party repositories, which then handle issues such as verifying ethics approval and secure access mechanisms. Some full-service repositories, such as Inter-university Consortium for Political and Social Research (ICPSR) or various national archives, offer such deposits, though they are rarely used by individual researchers in economics. One example are the data

[11] In fact, authors sometimes forget to abide by the rules for these datasets. As AEA Data Editor, I get notified via "take-down requests" from data providers 12-15 times per year, including in 2024 from the PSID, and have posted information on how to achieve compliance, at least in some cases, at https://aeadataeditor.github.io/posts/2024-11-01-psid-requests. Most of the cases affect papers published prior to my tenure, because I do alert authors to data use agreement violations that I am aware of. Ultimately, however, it is the authors' obligation to remain compliant with such data use agreements.

[12] https://labordynamicsinstitute.github.io/fsrdc-external-census-projects/

[13] Provided by Grant Gibson, CRDCN, on February 10, 2025.

[14] The Nagaraj and Tranchero (2023) number only includes papers published by economists. All other numbers are counts of researchers and publications in all disciplines, and include non-economists.

collected and used by Ahrsjö, Niknami, and Palme (2024b). The data were collected from public information from the Stockholm District Court. However, in combining individuals' information into a database, the result was subject to GDPR, and could not be included in the replication package (Ahrsjö, Niknami, and Palme 2024a). However, the authors were able to deposit their pseudoymous data at the Svensk nationell datatjänst (Swedish National Data Service), as a restricted dataset (Ahrsjö, Niknami, and Palme 2023). The SNDS will verify compliance with Swedish law (e.g., GDPR), without requiring future involvement of the original collectors of the data.

How onerous are access restrictions in general? In my work as Data Editor, it matters less than it would seem at first blush. The AEA asks that authors, when submitting, provide some information on how restrictive the data used in the article are, and whether the authors are able to provide the data editor with a not-for-publication copy, for the purpose of verification.[15] Table 2 lists the four levels of restrictions, with some guidance on how to classify specific data. Access via research data centers, described earlier, generally fit within the 'moderately difficult to obtain' category. Many private-sector datasets, relying on personal interactions with individuals within the company, fall into the 'very difficult to obtain', because others may never be able to obtain them. Also in these lists are access-restricted data that have been discontinued, and are thus no longer available to anybody, such as Zillow's ZTRAX program, terminated in 2021 (Zillow 2021). Authors can list multiple categories, and we review these, possibly adjusting them before we record them in our internal database.

Table 2: Restriction categories

Category	Explanation
Very Easy to Obtain	Request takes just a few minutes with no associated costs and the expected response is within a few days
Moderately Easy to Obtain	Request takes less than an hour with minimal cost
Moderately Difficult to Obtain	Request requires a multipage application; request needs university approvals; request involves significant cost; there is some uncertainty as to whether the proposal will be granted
Very Difficult to Obtain	Request must be made in person and/or access is provided only in person; request requires substantial funding; data and/or access mechanism may no longer exist

Figure 2 lists the distribution of the four levels of restrictions plus the no-restriction category, for the 96 AER articles in 2024, interacted with whether the authors offered to share the files with the data editor.[16] Past studies have found that about between 40% and 60% of articles rely on data that are not freely available (Herbert et al. 2024; Hamermesh 2013; Chetty 2012). In 2024, slightly more than 50% of all articles had no data restrictions at all (included the data in the replicatin package), with another 19% being 'very easy to obtain'. The combined percentage of 70% is higher than the earlier long-run average. More importantly, out of the 39 data sources that are difficult to obtain, about a third could be shared with the data editor. It is also worthwhile highlighting that the 'moderately difficult to obtain' tends to be data from research data center networks that have fairly robust mechanisms of access, so that in principle and in practice, many researchers do have access to the same data used in these articles, even when the data editor may not have the time to access these data, and when it is not a legal option for the authors to share the data with the data editor. Arguably, the only data sources that are of immediate concern are data where access is uniquely attributable to the authors of the paper, or where access is no longer possible.

[15] Data and Code Availability Form

[16] Not all offers to share data are taken up, depending on available resources and time. Appendix Table **??** shows the same numbers.

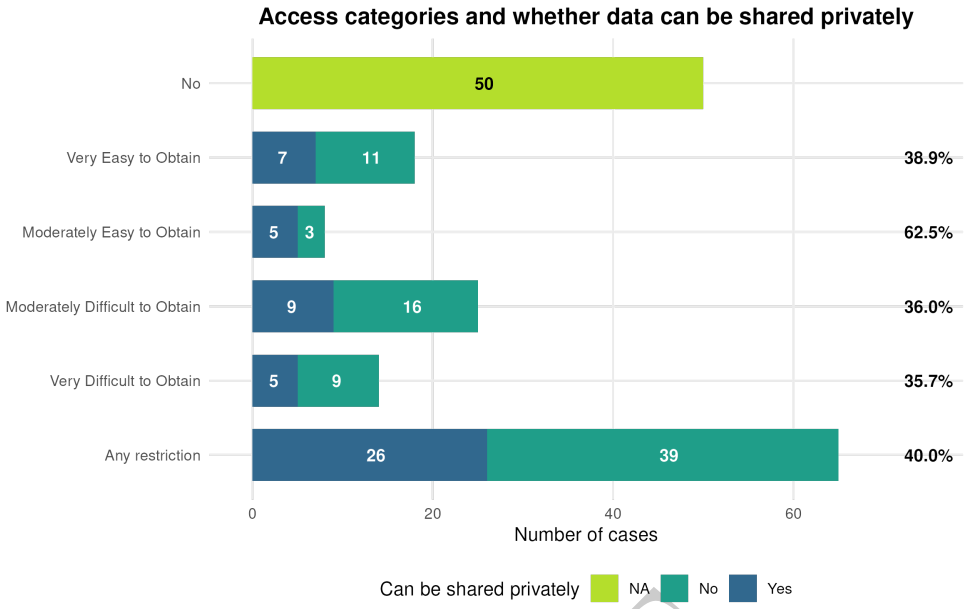

Figure 2: Access categories and whether data can be shared privately

4 Access to Software

Turning to software, I will again need to define more precisely what I mean by that. I distinguish two key categories of software: **high-level interpreters** (or more rarely compilers), and **instructions**. The former will in turn comprise software used in two key but distinct features of the scientific production process: **data collection** and **analysis**.

High-level software for **analysis** are the flagship software that receive the most focus in the social science literature. These are software products such as Stata, MATLAB, R, Julia (see Figure 3). Most of these are interpreted languages at the user level, though user-contributed packages may be compiled (R, Julia). Increasingly present is Python, which may be both compiled and interpreted; less common are purely compiled languages such as C or Fortran. We also include here dedicated geographical information system (GIS) software, mostly used to create maps, though some analytical tasks can also be performed. It arguably may also comprise custom plugins to other software, such as Dynare (Adjemian et al. 2024; Cherrier, Saïdi, and Sergi 2023).

While many economists use data collected by others, some are primary data generators or collectors, for instance through laboratory or field experiments, as well as surveys. **Data collection** software in this context are the survey software tools (Qualtrics, LimeSurvey, SurveyCTO, KoboToolbox) used to conduct surveys (including as part of lab or field experiments), as well as customized experimental software, such as oTree (Chen, Schonger, and Wickens 2016) and z-tree (Fischbacher, Bendrick, and Schmid 2021).

Software instructions (code): The core of a paper's analysis, however, resides in how the more general software package is manipulated, in combination with the data, via **instructions**. I do not call these 'source code', as that terms tends to be used in conjunction with complex compiled software, such as those listed as **high-level software**. Rather, these instructions are mostly interpreted code, or compiled just-in-time, in the language used by the high-level software, or, as is still sometimes the case, in the form of instructions to

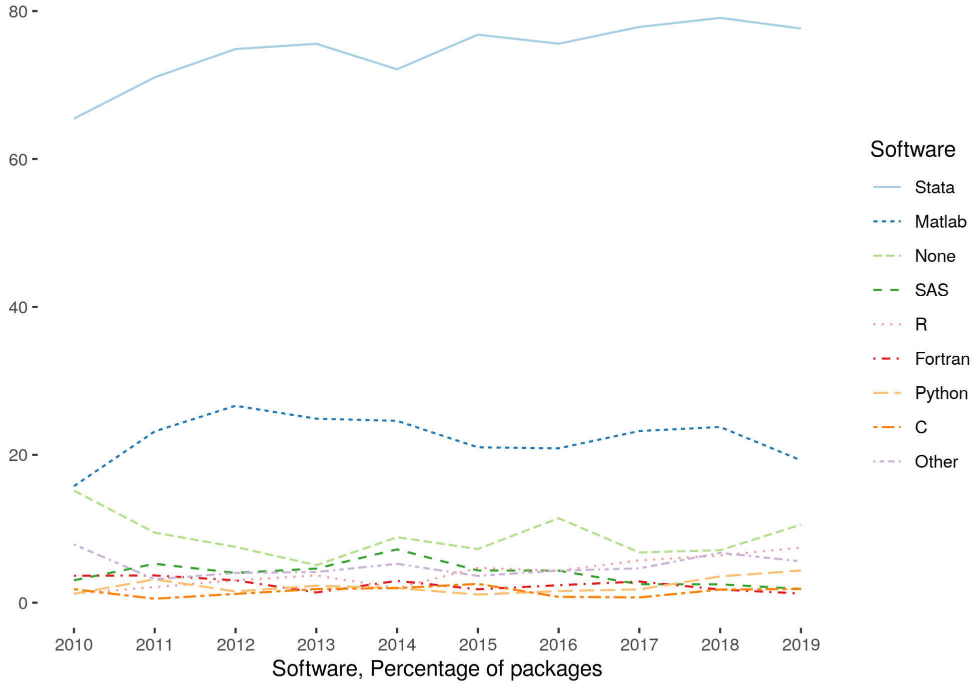

Figure 3: Software usage in AEA journals, 2010-2019, percentage of replication packages. Data do not sum to 100% as a replication package may use several different software. Originally published in Vilhuber, Turitto, and Welch (2020), corrected version see Vilhuber (2020a).

humans on how to manipulate a non-programmed user interface for software.[17]

Open access, therefore, can work through several channels. Software can have a cost. In economics, it is extremely rare to sell access to the **instructions**. However, it is quite normal for **high-level statistical software** and **data collection software** to be accessible only through purchase. As Figure 3 shows, the top two statistical software products used in replication packages are commercial closed-source statistical software products: Stata and Matlab. In order to be able to re-use the instructions (code) from academic papers, a software license is required, and must be purchased.

How much of an impediment to open access is this? Without loss of generality, to illustrate, consider Stata. An academic single-user yearly license as of January 2025 was $690 for a US-based student. The price is the same for a student in (much poorer) Greece. For a student in Vietnam, the price drops to $220.[18] To put this into perspective, the average monthly incomes in each of these countries, which students are unlikely to earn, are $6,704, $1,883, and $343, respectively, and a reasonably powerful laptop a student is likely to have was around $1,000.

Table 3: Software licensing internationally

Country	Average.monthly.income	Stata.license	Percent
USA	6,704	690	10.3
Greece	1,883	690	36.6
Vietnam	343	220	64.1

Notes: Amounts in USD as of 2025. Source for Stata license prices: Stata website.

An informal survey of several economists working with colleagues and students in Latin America, Africa, and Asia uniformly suggested that access to commercial statistical software was often difficult for students, alleviated somewhat when students were fortunate to attend private universities. In fact, in a recent survey (CIDR), respondents were asked to name the top two factors that could support early-career African scholars' publication success. 46% of respondents chose 'Providing access to datasets and data management tools' as one of the two factors. University staff and students (across multiple disciplines) were also asked which aspect they most needed funding for, and 38% mentioned 'Data analysis software'. Within top universities, on the other hand, most researchers, including graduate students, will have access to these software products, and even many undergraduates may be able to leverage university computer labs to access these software. Outside of academia, however, it may be more constraining. Precise information is hard to come by.[19]

The landscape is even harder to assess for **data collection software**. Open source packages otree and z-tree are often used for lab experiments. otree is cited by between 1,400 (OpenAlex, as of 2025-01-29) and 2,400 articles (Google Scholar, as of 2025-01-29), z-Tree is cited between 9,000 (OpenAlex) and 12,000 times (Google Scholar), but it is not clear how to benchmark that, given poor software citation practices in economics. Lab as well as field experiments will often use online survey platforms to augment experiments, or as primary data collection tool. Qualtrics, one of the big commercial platforms for surveys, is mentioned (but not cited) over 200,000 times (Google Scholar). Open source alternative LimeSurvey is mentioned about 31,000 times.

The second channel is through **software instructions**. Conditional on having the right high-level software and the appropriate resources (see next section), most economics papers have openly accessible code. The advent of data and code availability policies most certainly has helped, but also reflected attitudes already present in the researcher community. As noted earlier, the AEA's earliest policy was announced in

[17] In economics, this appears to occur most often for GIS software, and sometimes for data extraction software used by data providers.

[18] See source text for data.

[19] To a smaller degree, the problem also arises for **compilers**. While there are open-source compilers for Fortran and C (GNU Fortran, LLVM Flang), many economists will use what they consider to be more efficient commercial compilers, which may require the payment of subscription fees (e.g., Intel, NAG Fortran, PGI (now Nvidia)). More recently, some of the commercial compilers (Intel, Nvidia) have become freely available. Some numerical (non-linear) optimizers (e.g., Gurobi, Knitro) may also require payment of fees, though some limited academic free use may be available.

2004 (Bernanke 2004), but the oldest replication package curated by the AEA accompanied an 1999 article (Frankel and Romer 1999a, 1999b) preceding the policy by several years.[20] The newer policies (American Economic Association 2019; Koren et al. 2022) required that more code be provided (starting with raw, rather than cleaned data), and the systematic provision of supplementary materials, such as the survey code to use with data collection software. These policies require that the code be made available in a reasonably liberal license, though no specific open source license is specified. Furthermore, most of the top journal's repositories of replication packages are at trusted repositories, and not behind journal paywalls. Thus, software instructions in economics, as associated with the scientific output in journals, is generally cost-free, and almost always available under an open access license.

5 Access to Other Resources

In the template README published by myself and several other data editors in economics (Vilhuber et al. 2022), we emphasize that a README should provide enough instructions for a reasonable person to re-implement the analysis described in the replication package. Authors need to take into account that cutting edge methods, including technology, may require more information and instruction than for standard methods. Authors likely do not need to describe how to run an analysis in Stata, given its ubiquitous use in economics, and the ease with which instructions can be found more generally, but they may need to provide detailed step-by-step information if the technology used is rare or bespoke. For instance, the emerging use of large language modelss (LLMs) and artificial intelligence (AI) methods is far from the economic mainstream as of the writing of this article. Recent articles are still identifying ways economists scan actually use these tools, both for personal productivity (Korinek 2023) and as part of the technical toolkit (Athey and Imbens 2019; Dell 2024).

The most recent modern toolkits are not the only resource constraints that might restrict broad access. While it might be argued that the use of proprietary software is restrictive, it is one of many resource constraints that can be binding for some researchers. Researchers in lower-ranked (and lower-funded) research institutions, including in low- and middle-income countriess (LMICs), may well not have the funds to purchase proprietary software, but access to computers may be equally constraining. The template README requests information on the type of computer that was used by the original researcher, to provide a benchmark to future re-users. Acquiring access to sufficient memory (random access memory (RAM)), storage, and use over time of those resources can be expensive, even when renting such resources in cloud environments (which very few researchers appear to be doing). Traditionally, that access may be embedded within a single purchased computer, which may have (in 2024) around 32GB of RAM, 1-2 TB of storage, and have 4-12 compute cores available exclusively to the owner. More complex analyses may require access to shared compute clusters (using hundreds or thousands of compute cores), very large storage arrays (measured in the two- to three-digit TB range), and may require up to 1024 GB of RAM. Cutting edge analyses may require specialized chips, such as one or more graphical processing units (GPUs), or even a cluster of GPUs. I have observed analyses that may run data cleaning or data acquisition processes for months at a time.

The vast majority of articles published in economics journals usually require no more computing resources than a modern laptop provides, in all the dimensions enumerated in the previous paragraph. In fact, a formal quantitative measurement of resource usage in economics articles is surprisingly hard to obtain, as most researchers are not very good at reporting the resources they have used to conduct their research. In part, this is because measuring such usage is non-trivial, but to a larger extent, I postulate that this is because most research institutions provide such resources to their researchers in a "convenient way," and researchers conduct research within those constraints. More importantly, however, it suggests an important constraint on how "open" access can be for some if not all economics research.

Some newer research requires vastly different types of resources. Studies using raw satellite data may require more than 10TB of data storage (Khachiyan et al. 2022b; Khachiyan et al. 2022a), may need more

[20] The oldest replication package in economics may be Koenker (1988) in the JAE replication archive, see https://aeadataeditor.github.io/posts/2023-02-02-oldest-replication-package-jae, which actually contained data processing code, but not the data analysis code. That was only published in a later replication package, Koenker and Zeileis (2009).

than 20,000 compute hours on a cloud provider (Rudik 2020b, 2020a), or the use of one (Dell 2024, 2025) or dozens (Khachiyan et al. 2022a) GPUs.[21] Access to the code and data for the papers mentioned is open: The AEA-related replication packages for these articles are licensed under a standard Creative Commons Attribution (CC-BY) license. Some of the data not included in the Dell (2025) replication package is on Huggingface, also under a CC-BY license (Silcock et al. 2024). Open access to satellite data is one of the canonical examples of the benefits of open access (Nagaraj, Shears, and Vaan 2020). In these cases, the computational resources may restrict the benefits of the open access of data and code.

Are such computational constraints a problem? No consistent analysis exists that correlates resource requirements to academic outcomes such as citations, primarily because it is very hard to measure consistently the resource requirements of economic articles. The very small sample in the previous paragraph may serve to illustrate this, but without controls for scientific merit, is purely an indicator. Silcock et al. (2024) had been downloaded 98 times in December 2024, six months after the arXiv paper associated with it was published (Silcock et al. 2024). Rudik (2020a) has had 1432 views, 124 downloads for replication package, as the manuscript (Rudik 2020b) has 15 citations. The replication package Khachiyan et al. (2022a) has 2124 views, 175 downloads, while the manuscript (Khachiyan et al. 2022b) has 5 citations (all as of January 2025). For comparison, the average article in one of the AEA's journals has 908 views and 106 downloads (Vilhuber 2025, Table 4).

6 The Benefits of Open Science in Economics

To illustrate some of the benefits of open science, I describe two recent articles in economics that leverage the availability of code and data to improve econometric methods for future researchers. The first paper relies on the empirical recomputation of prior papers to assess a theoretically ambiguous potential bias in inference. The second paper also focuses on inference, and uses actual data from previous papers to simulate the relevance of the impact. Both then provide new software (R and Stata packages), under open source licenses, to "fix" the problem for future studies.

Roth (2022) uses 12 previously published papers to assess whether the usual tests for pre-existing differences in trends when using difference-in-differences methods are properly controlling for power, and the empirical impact on subsequent inference. The theoretical bias is ambiguous, so an empirical evaluation is necessary. The study both leverages the open availability of materials in economic journals, but also illustrates the limitations imposed by imperfect adherence to openness. To wit, Roth writes that he searched for

> "The phrase "event study" in papers published in the *American Economic Review, American Economic Journal: Applied Economics*, and *American Economic Journal: Economic Policy* between 2014 and June 2018 ... The search returned 70 total papers that include a figure that the authors describe as an event-study plot."

but continues to then be limited by lack of data in the majority of cases:

> "I exclude 43 papers for which data to replicate the main event-study plot were unavailable. (Roth 2022, pg. 307)"[22]

While it remains unclear whether the excluded papers are non-compliant with the AEA's policy at the time, or whether they have legitimate reasons not to provide the data (Roth does not provide the raw result of his search), the paper is able to make an important methodological point (723 / 415 citations as of January 2025, per Google Scholar/ OpenAlex) because it is able to fully recompute the results in previous papers, apply new tests and methodologies, and come to meaningful recommendations and tools — Roth provides an (open source) R package to implement his methodology.

[21] As of January 2025, the type of GPU used by Dell (2025), costs between USD 4500 and USD 7700, or between 2 and 7 times as much as a standard laptop.

[22] He also excludes another 15 papers for reasons not related to data availability.

Chaisemartin and Ramirez-Cuellar (2024) use open access information on RCTs (AEA registry) to find 15 RCTs of a particular type (clustered paired or small strata), of which 4 have publicly available data (and reproducible artifacts). They then provide results both on simulations using these data, and in particular, re-estimate the regressions used in those studies and apply their proposed solution, showing that the number of significant effects is reduced by one-third. In other work (Chaisemartin and D'Haultfœuille 2020, 2024), the results from various other papers are also recomputed to empirically demonstrate the relevance of the proposed methods, and software packages (e.g. Chaisemartin et al. 2025) are developed and made openly available.[23] Chaisemartin and D'Haultfœuille (2020) has been cited between 2,600 and 4,600 times (OA, GS).

In both of these examples, the ability to access prior data, code, and information is critical to improving future scientific progress, but is limited by both historical and unavoidable limitations on openness.

> “These studies were identified by a systematic search of papers in the AEA Data and Code Repository”

7 Open Infrastructure: Publications

The challenges of openly accessible written scholarship are manifold, with the current focus on “Plan S”, master publication agreements, and in the US, similar efforts under the moniker of the “Nelson memo” (Brainard and Kaiser 2022; Brainard 2024). I note that the economics profession has a very long history of making much of the written knowledge available at very low cost via working papers (Vilhuber 2020b), with the first working papers at the reputable NBER working paper series going back to 1973 (Welch 1973).

At the time of this writing, I have three editorial appointments. I am the Data Editor for the journals of the American Economic Association (AEA), the joint executive editor for the open access and multi-disciplinary Journal of Privacy and Confidentiality (JPC), and I am a column editor for the open access Harvard Data Science Review (HDSR). I will use each of these to highlight a particular pattern in broadening access to publications, without any claim to generality.

The AEA is a not-for-profit organization, as are many other learned societies. It self-publishes eight journals, plus the proceedings of the annual conference, without relying on a commercial publishing house. Depending on the measurement, three of these publications are in the top ten journals in economics (Mogstad et al. 2022). Its publication costs account for about half of its overall operating expenses, and are only partially offset by directly attributable subscription and membership fees (Cherry Bekaert, LLP 2024). In fact, 6 of the top 10 journals in economics (Mogstad et al. 2022) are published by societies (JEL, JEP, Econometrica, AER, Restud, JOLE), some of which have as sole or primary purpose the publication of the journal.[24] A further three journals are primarily associated with economics departments (QJE, JPE, RESTAT), which arguably may not be driven by pure profit. The sole outlier in the top ten is the Journal of Financial Economics (JFE), which is owned by Elsevier, a big commercial publisher. It should be noted that the European Economics Association (EEA) severed its relationship with Elsevier in 2003 for its official journal, creating a journal that is fully owned by the association, adding to the list of society-owned journals in economics (Tirole et al. 2003). Access to these journals is generally still on a subscription basis (JEP is the exception, being free to read), but given the primarily not-for-profit organization of its owners, personal subscriptions (often via society membership) are quite low, compared to journals in many other sciences. For instance, as of 2024, a personal subscription to the Review of Economic Studies is $156 or €141 per annum; a yearly membership to the AEA, providing access to the seven subscription journals and the proceedings is $25 for students and researchers in low-income countries, and $150 at the highest personal income tier. As outlined earlier for the AEA, these subscription fees cover only a small fraction of the production costs. Nevertheless, even these (arguably low) costs do not satisfy “Plan S” or “Nelson memo” requirements, which

[23] Note that as of January 2025, many of the packages do not have an explicit open source license — or any license — applied, a common feature of economists working in the open source world. My presumption is that they simply assume that everybody knows that the code is openly available.

[24] JOLE is a bit of an outlier, in that one becomes a member of the Society of Labor Economists by subscribing to the journal, rather than the other way around.

require no access cost to the end consumer, and in the case of “Plan S”, also require a liberal license allowing for re-use.[25]

Interestingly in the context of the previous sections, all of the society-owned journals in the previous paragraph have appointed data (reproducibility) editors.[26]

Since 2018, I have been the executive editor of the JPC, an open access multi-disciplinary journal, having taken over the journal from Stephen Fienberg (Vilhuber 2018).[27] As of 2024, the journal does not charge submission fees, and is free to read (what is called “diamond open access.” Articles default to a Creative Commons Attribution-NonCommercial-NoDerivatives (CC-BY-NC-ND) license, though authors are allowed to choose a more liberal license, for instance to comply with “Plan S” (which does not allow for the “no-derivatives” part). As executive editor, I have been responsible for all aspects of running the journal, not just finding referees for the articles that I am responsible for. The journal is made available through open-source software called Open Journal System, hosted by its creators at Simon Fraser University’s Public Knowledge Project, preserved via industry-standard mechanisms (CLOCKSS, a non-profit) in case the journal ever needs to shut down, indexed in a variety of academic indexes, including via assignment of DOI. Copy-editing is done through a mixture of professional copy-editors and volunteer work by editors and board members. All editors, including myself, are unpaid, and referees are, like in much of the publishing industry, unpaid volunteers. Yet I do pay bills, for each of the above components of a properly managed, indexed, and preserved academic journal — and professional copy-editors and university staff do not work for free. I am thus quite aware of the absolute minimum cost of running a (small) journal. Over the years, funding has come from a variety of chaired professorships at Carnegie Mellon (Fienberg), Cornell (Abowd), and Harvard (Dwork). In order to make such funding more robust, a non-profit society was created to better and more robustly structure the funding situation (Abowd et al. 2024). Time will tell if this will stabilize the funding situation, while maintaining the foundational commitment to open access. Others, in particular Sociological Science, have shown that it is feasible to sustainably publish high-quality research

8 Conclusion

I have described in this article how open data and code are in the academic literature in economics, and how access to software and hardware resources can be limiting factors. Almost all code is openly accessible in top journals. The vast majority of data is accessible with little to no effort, and a large proportion of the remaining restricted data can be accessed by networks that include thousands of researchers. I provide a few concrete examples where the openness of the data and code available allows others to directly build on prior results. Broader assessments of the benefits of open access are more difficult to measure, in part because the right controls are hard to construct, in part because the community is still only starting to learn how to technically leverage the openness in a large-scale fashion.

Nevertheless, access to networks of data access and financing remains one of the key worries. I am a regular participant in discussions within three research networks that provide access to restriced data (FSRDC, Canadian Research Data Center Network (CRDCN), and CASD). Core discussions center around equity and access, and how to balance those criteria while preserving the privacy of the respondents for which these networks act as curators.

One under-appreciated aspect of open access is that it enables persistence.

[25] “The author(s) [...] grant(s) [...] a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, [...] subject to proper attribution of authorship” (Max-Planck-Gesellschaft 2023).

[26] Two additional societies, not previously mentioned, also employ data editors: the Canadian Economics Association (CEA) and the Western Economics Association International (WEAI).

[27] Fienberg, together with Cynthia Dwork and Alan Karr, founded the journal in 2009 (Abowd, Nissim, and Skinner 2009). Fienberg passed away in 2016 (Slavković and Vilhuber 2018). In 2023, I recruited Rachel Cummings to jointly manage the journal.