Report for 2018 by the AEA Data Editor

By Lars Vilhuber*

Your abstract here.

The purpose of scientific publishing is the dissemination of robust research findings, exposing them to the scrutiny of peers. Key to this endeavor is documenting the provenance of those findings. For empirical articles, the foundations on which they reside are external to the article, and often to the journal, in which they are published. Our scientific community faces increasingly complex issues of privacy and confidentiality that prevent "open" access to those same sources. In consequence, there is a need to properly cite the digital inputs to our published output and to properly curate those inputs.

Many scientists, journals, learned societies, and funding agencies have called for greater transparency of research practices, and more assurance that published research is reproducible (Stodden et al., 2016; Fuentes, 2016; Moffitt, 2016; Camerer et al., 2016; Bollen et al., 2015; Joskow, 2015). This has lead to a focus on transparent access to research data and code (Coffman, Niederle and Wilson, 2017; Hoeffler, 2017; Duvendack, Palmer-Jones and Robert Reed, 2017; Hamermesh, 2017). This interim report describes the current conceptual and practical issues, outlines the American Economic Association (AEA)'s efforts, through the new position of Data Editor, to address these issues, and highlights some of the short- and medium-term changes that economists might expect.

I. The current environment

The AEA's data and code posting policy (American Economic Association, 2008), as well as that of other societies and journals, are intended to create a minimal framework from which to replicate empirical findings. Many partial solutions have been implemented. A few journals have implemented verification of submitted code and data during the editorial process,[1] highlight the verification on data archives (Open Science Framework, 2017), maintain lists of acceptable third-party repositories,[2] and interlink with collaborating repositories to highlight authors' (and repositories') contributions to the data component of a scholarly work.[3] Outside of journals, several projects are working to educate the community

* Vilhuber: Cornell University, lars.vilhuber@cornell.edu.

[1] The American Journal of Political Science outsources this activity to the Odum Institute for Research in Social Science (CITE). The Journal of the American Statistical Association performs a "broad evaluation of quality and potential for usability of the code and data" since 2016 (Stodden et al., 2016).

[2] Nature Scientific Data maintains a list for its journals (Nature Scientific Data, 2016), and other institutions (CoreTrustSeal, FAIRsharing) have as their primary purpose to perform this kind of vetting.

[3] Elsevier interlinks, for instance, with ICPSR, highlighting the use of a repository on the article's web page.

to incorporate principles of replicability and traceability into their workflow.[4] No journal currently does, in my opinion, an adequate job of providing information about restricted-access data, in part because most restricted-access data centers cannot provide structured information about existence, modalities of access, or even data landing pages.[5] None of these solutions are widespread, and standards are only now being developed.

In the summer of 2018, I had the privilege of contributing a white paper to the Committee on Reproducibility and Replicability in Science of the National Academies of Science on the history and state of reproducibility in economics. In many sciences, new preprint services have emerged within the last two years, e.g., PsyArXiv. These are considered to be part of the broader move to greater research transparency. While writing the white paper, I pointed out that this kind of pre-publication exchange has long been the norm in economics. The first National Bureau of Economic Research (NBER) working paper, one of the most prestigious working paper series in economics, was published (in paper form) in 1973 (Welch, 1973). By the early 1990s, there was a wide variety of such working paper series, typically provided by academic departments and research institutions. Since grey literature at the time was not cataloged or indexed by most bibliographic indexes, a distinct effort to identify both working papers and the novel electronic versions grew from modest beginnings in 1992 at Universit de Montral and elsewhere into what is today known as the Research Papers in Economics (RePEc) network, a collaborative effort by hundreds of volunteers in 99 countries (, n.d.; Krichel and Zimmermann, 2009; Bátiz-Lazo and Krichel, 2012). The initial index was split into electronic (WoPEc) (Krichel, 1997) and printed working papers (BibEc) (Krichel and Zimmermann, 2009; Cruz and Krichel, 2000), testimony to the prevalence of the exchange of scientific research in semi-organized ways. In 1997, BibEc counted 34,000 working papers from 368 working paper series (30). RePEc today has data from around 4,600 working paper series and claims about 2.5 million full-text (free) research items, provided in a decentralized fashion by about 2,000 archives (31). These items not only include traditional research papers, but also, since 1994, computer code (3234).

[4] Open Science Framework, Project TIER, BITSS are just a few of those active in that field (Gentzkow and Shapiro, 2014; Wilson et al., 2016).

[5] Restricted-access data hosted on ICPSR and possibly Harvard Dataverse are notable exceptions. On the journal side, Elsevier journals have experimented with ”Data Descriptions”, but while the form is machine-readable, it is essentially free-form text, and checking the box ”confidential data” essentially stops the process of filling in any information.

II. Definitions

A brief definition of the terms "replicable" and "reproducible" is in order. The terms are somewhat ambiguously defined in the scientific community. We will use the definitions used by Bollen et al. (2015). "Reproducibility" refers to the ability of researchers to achieve the same results as previous researchers, using the same data and same analysis methods, and "replicability" refers to the ability of researchers, using the same analysis methods, to duplicate results using "new" data. Other authors may use similar or conflicting terms (Clemens, 2017).

III. Implementing improved transparency of research

A modern data and code availability policy should support both reproducibility and replicability, by supporting accurate and transparent description of the provenance of the scientific results. In particular, a functional implementation of those concepts suggests that both data and code need to be subject to the Findable, Accessible, Interoperable, Re-usable (FAIR) principles FORCE11 (2016): findable, accessible, interoperable, and re-usable. In this context, we interpret the "interoperability" of code as "code that works, and the workings of which are comprehensible by a third party" (CITE??).

A. Goal 1: Improved findability of data used in research articles

Under this goal, we start by abolishing journal-specific "supplementary materials" as the primary repository of data and code that are part of the provenance chain of an article. As currently implemented at most journals, including the AEA's journals, they lack findability, proper citability as first-class objects, and are somewhat opaque (packaged as ZIP files). Historical materials will be migrated to a new curated archive at ICPSR. The AEA's "Data and Code Archive @ ICPSR" will display the full contents of the materials as deposited by authors in the past, without the need to download ZIP files. The materials will receive their own citable Digital Object Identifier (DOI). Through the DOI registrars, we automatically leverage the ability to link and associate the archives with their original articles.

On the AEA's journal websites, the links to "supplementary" materials will initially appear to be the same (although pointing to the new locations), but future enhancements will allow for greater visibility or transparency of the associated materials. However, by separating the hosting of data archives (for historical materials, at ICPSR) from the referencing of those archives, we open the door to a more consistent model of linking to and citing data artifacts associated with published articles.

For future submissions, we will allow researchers to reference supplementary materials on a wide list of data archives or repositories.[6] In particular, while some journals are already curating a list of recommended *open* data archives (Nature Scientific Data, 2016) (ALSO PLOS, F1000), we will also allow authors to reference materials in reliable *restricted-access* data repositories. What constitutes a "reliable" data repository? For one, it needs *persistence*, and *accessibility*. Properly managed repositories have a *preservation policy* - they commit to maintaining deposits for a defined duration (often, but not always, in perpetuity), and only under very restrictive circumstances will remove deposits. Such repositories will also have a policy about access - who can obtain data deposited at the institution, and under what conditions. This characterization applies homogeneously to open and restricted-access repositories. A recommended repository will have been vetted (by the AEA Data Editor, or a reliable third party) to have acceptable and credible policies.

Authors submitting their work to the AEA journals will be affected in several ways. First, those authors who already deposit their (open access) code and data at known repositories will not have to do so again - a simple reference (and citation!) of the previously archived materials is sufficient. Authors who use data provided through institutional providers (Panel Study of Income Dynamics (PSID), Health and Retirement Study (HRS), the U.S. Census Bureau, and international equivalents), and who in some cases cannot deposit the data, will also reference the persistent location where they obtained their data from, and where others can do so as well. In the case of restricted-access, a better description of access procedures will be requested from authors, who in turn should ask their data providers to provide such procedural descriptions, in the form of web pages and (persistent and citable) documents.

If replicability is truly part of the research scientist's workflow, then by the time she submits an article to any journal, the intermediate and final data products as well as the code used for an article have already been deposited at appropriate repositories and archives. If all such repositories and archives are of sufficient quality, then the additional deposit at a journal is duplicative at best, and perturbative to the provenance chain at worst. The right solution is to reference those other repositories, not copy them. Of course, for those that have not used repositories and simply wish to provide a replication zip archive of the files on their laptop, an adequate deposit solution should also exist. By fundamentally relying on references to repositories instead of deposits, it also becomes possible to put public-use and restricted-access data on a comparable footing at the journal with regards to potential replicability. Well documented location of data (through DOI), access protocols (implicit or click-through license, contracts, etc.), and access mechanism (direct download, delivery of physical media or controlled download, sign-on to controlled secure access, etc.) are then available for any data.

[6] Do I need to distinguish the two? is there a difference?

IV. Goal 2: Improved Reliability of Replication Materials

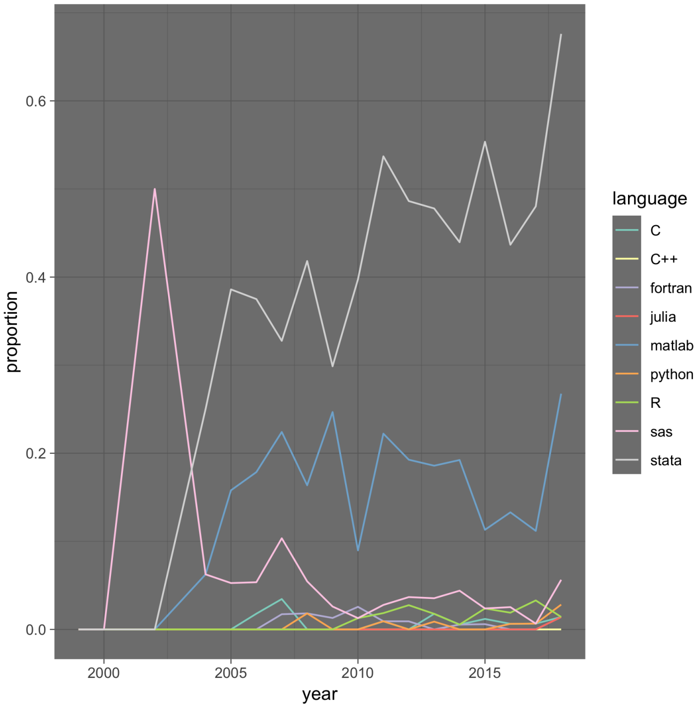

Figure 1. Popularity of statistical software in the AER

Figure provided by Patrick Baylis (UBC), based on filename extensions in ZIP files of replication materials on the AEA website.

V. Data Citations

Properly referencing data goes beyond just reproducibility - it is also proper scientific writing style. In the same way that we use bibliographic references to "printed" resources, we should also be using such references for data resources, to give and receive credit where credit is due. Not referencing an article or book is at best an oversight, and at worst plagiarism - and the same should apply to data objects. Numerous guides and tutorials exist (DataONE, 2011; ICPSR, 2018; Martone, 2014).

The AEA uses the Chicago style for citations and bibliographies (American Economic Association, 2018). However, the Chicago Style Manual (Chegg, 2018; Chicago Manual of Style Online, 2018) does not provide examples for data citations, and neither does the Citation Style Language[7] used by applications like Zotero[8] and Mendeley Desktop[9].

As part of our activities, the AEA prepress department has started the process of updating AEA templates available through such software.[10] Some guidance for data citations is provided at

REFERENCES

American Economic Association. 2008. "Data Availability Policy." *https://www.aeaweb.org/journals/policies/data-availability-policy*, (accessed: 2017-04-06).

American Economic Association. 2018. "Sample References." *https://www.aeaweb.org/journals/policies/sample-references*, (accessed on 2018-08-10).

Bátiz-Lazo, Bernardo, and Thomas Krichel. 2012. "A Brief Business History of an On-line Distribution System for Academic Research Called NEP, 1998-2010." *Journal of Management History*, 18(4): 445–468.

Bollen, Kenneth, John T Cacioppo, Robert M Kaplan, Jon A Korsnick, and James L Olds. 2015. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science." Subcommittee on Replicability in Science, National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. 2016. "Evaluating replicability of laboratory experiments in economics." *Science*, 351(6280): 1433–1436.

Chegg. 2018. "Citation Machine: Chicago Manual Of Style 17th Edition (Author Date) format citation generator for journal article." *http://www.citationmachine.net/chicago-17-author-date*, (accessed on 2018-10-02).

Chicago Manual of Style Online. 2018. "Author-Date: Sample Citations." *https://www.chicagomanualofstyle.org/tools_citationguide/citation-guide-2.html*, (accessed on 2018-10-02).

[7] https://citationstyles.org/

[8] https://www.zotero.org/

[9] https://www.mendeley.com/download-desktop/

[10] For the technically inclined, this process involves updating an existing style or creating a new style on https://citationstyles.org/ and https://github.com/citation-style-language/styles, from where it propagates to a large number of software packages.

Clemens, M A. 2017. "The meaning of failed replications: A review and proposal." *Journal of Economic Surveys*, 31(1).

Coffman, Lucas, Muriel Niederle, and Alistair J Wilson. 2017. "Replications: A Proposal to Increase their Visibility and Promote them."

Cruz, Jose Manuel Barrueco, and Thomas Krichel. 2000. "Cataloging Economics Preprints." *Journal of Internet Cataloging*, 3(2-3): 227–241.

DataONE. 2011. "DataONE Tutorial on Data Citation." *http://www.dataone.org/sites/all/documents/L09_DataCitation.pptx*, (accessed on 2018-08-10).

Duvendack, Maren, Richard Palmer-Jones, and W Robert Reed. 2017. "What is Meant by 'Replication' and Why Does It Encounter Resistance in Economics?"

FORCE11. 2016. "THE FAIR DATA PRINCIPLES."

Fuentes, Montse. 2016. "Reproducible Research in JASA." *http://magazine.amstat.org/blog/2016/07/01/jasa-reproducible16/*, Accessed: 2017-4-4.

Gentzkow, M, and Jesse Shapiro. 2014. "Code and data for the social sciences: A practitioner's guide." University of Chicago.

Hamermesh, Daniel. 2017. "What is Replication? The Possibly Exemplary Example of Labor Economics."

Hoeffler, Jan H. 2017. "Replication and Economics Journal Policies."

ICPSR. 2018. "Data Citations." *https://www.icpsr.umich.edu/icpsrweb/ICPSR/curation/citations.jsp*, accessed on 2018-08-10.

Joskow, Paul L. 2015. "President's Letter, Alfred P. Sloan Foundation Annual Report 2014." Alfred P. Sloan Foundation.

Krichel, Thomas. 1997. "WoPEc: Electronic Working Papers in Economics Services." *Ariadne*, , (8).

Krichel, Thomas, and Christian Zimmermann. 2009. "The Economics of Open Bibliographic Data Provision." *Economic Analysis and Policy*, 39(1): 143–152.

Martone, M. (ed.). 2014. "Data Citation Synthesis Group: Joint Declaration of Data Citation Principles."

Moffitt, Robert. 2016. "Report: American Economic Association Committee on Statistics (AEAStat)." *American Economic Review*, 106(5): 788–793.

Nature Scientific Data. 2016. "Nature Scientific Data recommended repositories." *figshare*.

Open Science Framework. 2017. "Badges to Acknowledge Open Practices Wiki." *https://osf.io/tvyxz/wiki/home/*, Accessed: 2017-10-18.

RePEc: Research Papers in Economics. n.d..

Stodden, Victoria, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John P A Ioannidis, and Michela Taufer. 2016. "Enhancing reproducibility for computational methods." *Science*, 354(6317): 1240–1241.

Welch, Finis. 1973. "Education, Information, and Efficiency." National Bureau of Economic Research w0001.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K Teal. 2016. "Good Enough Practices in Scientific Computing."

Reviewers

The following replicators have assisted in the post-publication verification of publications:

Flavio Stanchi,

Thanks.