A journal's guide to choosing a repository for replication packages

Marie Connolly[1], Miklós Koren[2], Joan Llull[3], Peter Morrow[4], and Lars Vilhuber[5]

[1]UQAM
[2]CEU
[3]MOVE, Universitat Autònoma de Barcelona, and Barcelona GSE
[4]University of Toronto
[5]Labor Dynamics Institute, Cornell University. Corresponding author: lars.vilhuber@cornell.edu

April 3, 2023

Abstract

Your abstract here.

1 Goals

This document is meant to guide academic journal managers and editors (collectively referred to as "editors" for the rest of the document) on how to choose support mechanisms for replication packages. Here, support mechanisms are broader than a simple choice of repositories, as it may entail connections to publication platforms, data citation support, and usability for computational reproducibility. The scope is limited to journals in the social sciences, where our expertise lies, but we provide pointers to additional information in other disciplines.

'Replication packages' in the social sciences are 'composite' objects. They typically contain both data and computational scripts or code, generally in one of several commonly used statistical software packages such as R, Stata, Matlab, or SPSS. This has implications for the ideal structure of a repository, which, as far as we know, no single repository satisfies. Gentleman and Temple Lang (2007) defined "research compendia" in a similar way, though they also envisagend that the manuscripts themselves might be dynamic, which is not typically the case for journals in the social sciences. The challenge to journal editors is how to balance several competing requirements and desired features.

The intended audience are academic journal managers, editors, and directors of learned societies who own or manage their own journals. Journals that are owned and managed by large publishers may already have an infrastructure provided or imposed by their owning publisher. Society journals, even when contracting with a publisher or publishing platform, may have the ability to craft or choose their own solution. Others may find the issues identified in this document of interest to the journals under their guidance or control, and may request additional features.

2 Goals of Data and Code Availability Policies

Data and code availability policies (most often simply referred to as "data availability policies") describe a journal's approach to ensuring transparency and availability of auxiliary materials commonly associated with academic articles published in the journal. While the earliest approaches targeted the availability of data — the Journal of Applied Econometrics has been requiring authors to deposit data since 1985 — researchers investigating their effectiveness very quickly identified computational issues as well (McCullough and Vinod, 1999). De facto, most replication packages contain both data and code.

Data and code availability policies have a range of features. Authors may be required to promise to deliver replication packages upon later request, either generically, or by incorporating an explicit data and code availability statement into the manuscript. For instance, the *Journal of Human Resources* states that

> "Authors are responsible for making available complete replication materials, including the data and code, models, and other materials necessary for replication [...]" (Journal of Human Resources, 2018).

In other instances, authors may be required to provide replication packages upon submission, conditional acceptance, or publication. Replication packages may be simply vetted for content or metadata, or their completeness and accuracy may be explicitly tested prior to publication (Christian et al., 2018; Vilhuber, 2019). The stricter such policies, the more authors may need to interact with a variety of journal-related systems that go beyond the classic manuscript submission systems.

3 Trusted Data Repositories

In order to provide the highest level of preservation, journals might want to leverage existing "trusted repositories." The concept has both some generally accepted criteria, as well as a (voluntary) certification process in the form of the "CoreTrustSeal[1]." Sansone et al. (2020) summarize key criteria, and include

- the ability to guarantee long-term persistence and preservation of datasets (though what "long-term" may mean can be surprisingly short)
- create and maintain stable and persistent identifiers for deposits, typically in the for of Digital Object Identifiers (DOIs)
- provide clear terms of accessing the data.

Additional technical criteria might include anonymous reviewer access. Lists of trusted or simply accepted repositories are maintained by Nature,[2] F1000Research,[3] and PLOS.[4] A curated database listing many repositories and their attributes is maintained by re3data.org[5].

Many repositories that do not get officially designated as trusted repositories may nevertheless comply with the basic criteria: longevity and preservation, persistent identifiers, and clear terms. For instance,

[1] `https://www.coretrustseal.org/`, accessed April 3, 2023.
[2] `https://www.nature.com/sdata/policies/repositories`, accessed April 3, 2023.
[3] `https://f1000research.com/for-authors/data-guidelines`, accessed April 3, 2023.
[4] `https://journals.plos.org/plosone/s/recommended-repositories`, accessed April 3, 2023.
[5] `https://www.re3data.org/search`, accessed April 3, 2023.

the World Bank Microdata Catalog[6], the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB)[7], and the Centre d'accès aux données (CASD)[8] all do long-term preservations, and the latter two assign DOIs to their datasets. Journals may decide to consider these and similar locations as acceptable, in particular when, as in the case of the IAB and CASD, data are confidential and cannot be included within public deposits. Note that neither of these are listed (as of the writing of this manuscript) in the re3data.org database, nor are any of them certified by CoreTrustSeal.

Note on the other hand that several platforms are not "trusted" repositories because they make no guarantees as to the permanence of a depositor's files (though they may make promises as to the persistence of the platform). In other words, they do not prevent the depositor from later deleting the data. Such platforms include commercial "cloud storage" companies (Dropbox, Box, etc.), but also (somewhat surprisingly) the OSF platform.[9]

4 Code Repositories

Code repositories are very popular in the computer sciences as locations for supplementary code, and sometimes data. Github, Gitlab, Bitbucket are commercial services, often free for academics . In some cases, institutions (universities) may have on-premise versions of these platforms.

Code repositories are transitory by nature. They generally have no overall long-term preservation strategy,[10] and in general, individual code owners can delete a deposit, or a release, at any time. They are, to that extent, no better than authors' personal websites. Some journals appear to accept provision of code via code repositories; others (e.g., AEA) explicitly exclude code repositories as an alternate location.

5 Data Curation

In addition to depositing and preserving data, data repositories may also perform professional data curation ("preventing bitrot"). This generally entails verifying that data are well described (codebooks are available, all variables or database columns are well described, including every value occurring), and transforming them into preservable, non-proprietary formats. Data curation focuses on preserving the content, if not necessarily the original form, of the data deposit. Doing so may sometimes involve technical means (Dataverse), or professional curation staff at the repository (Dryad, ICPSR) or at a third party (Odum Institute). Not all trusted data repositories offer such services: The openICPSR service at ICPSR and Zenodo offer no data curation services themselves.

[6] `https://microdata.worldbank.org/`, accessed April 3, 2023.

[7] `https://fdz.iab.de/en.aspx`, accessed April 3, 2023.

[8] `https://casd.eu`, accessed April 3, 2023.

[9] When making a project public, OSF warns that "Once [the project is] made public, you should assume they will always be public. **You can return them to private later,**.." Even after assignment of a persistent identifier (DOI), the depositor can permanently delete the project. The DOI then resolves to a page stating that "Resource deleted".

[10] `Github.com`, owned by Microsoft, has recently created the "Arctic archive," taking snapshot of some of their hosted repositories, but it is not known how often such an archive is made. It cannot be triggered by a code owner or a third-party.

6 Data citations

A data citation is the minimal data provenance indicator, in the spirit of the Data Citation Principles (Martone, 2014). They have the same basic components as literature citations, and should be used and listed the same way. An in-text reference to "Bureau of Labor Statistics (2000-2010)" would be resolved in the references as

> Bureau of Labor Statistics. 2000–2010. "Current Employment Statistics: Colorado, Total Nonfarm, Seasonally adjusted - SMS08000000000000001." *United States Department of Labor*. http://data.bls.gov/cgi- bin/surveymost?sm+08 (accessed February 9, 2011).

Equivalently, an article's replication package can be cited in the text as "Romer and Romer (2010b)" (where "2010a" would be the actual article), and resolved as

> Romer, Christina D., and David H. Romer. 2010. "Replication data for: The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks." *American Economic Association* [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E112357V1.

The exact style will depend on the citation style chosen by the journal. Many default styles (APA, Chicago) do not align well with the Data Citation Principles, so auxiliary guidance may be necessary. Many typical but complex examples are described at `https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html`.

While all repositories provide support for citability of deposits, and most display at least one version of a "suggested citation", they are not all equal in their ease of integration into the journal workflow. When replication packages are made public prior to the journal workflow, citation is trivial. However, in many cases, authors or journals may want to make the deposit public at the same time as, and not earlier than, the manuscript's publication date, in particular when double-anonymous review is required. Copy-editors need to know how to reference the replication package by its future permanent identifier. When such a permanent identifier only becomes known at publication time, this may cause problems. Some repositories allow to reserve a DOI (Zenodo, Dataverse), others are predictable (openICPSR), alleviating the potential issue.

7 Computational Platforms

Since the purpose of replication packages or research compendia is to facilitate the computational reproduction of a manuscript, computational integration, or at least access to a computational resource, would seem a very desirable attribute for an academically oriented repository. Researchers can use cloud computing services, such as Google Colab and various continuous integration (CI) services, in their research, but such integration is rare in repositories used in the social sciences. CodeOcean[11] is a commercial platform that has both a preservation policy as well as a computational platform underlying its offerings. It explicitly targets academics, verifies reproducibility, and creates permanent archives with DOI. However, while it supports some commercial statistical software (Stata, Matlab) and most standard open source statistical software (R, Python, Julia), it does not support arbitrary combinations of software, and has some limitations as to the size of the data supported as part of a single deposit. It cannot handle replication packages when no data is publicly accessible. Nevertheless, it is useful in many contexts, and used by engineering and some political science journals as the primary platform for research compendia. A non-commercial platform with similar attributes and integration (for preservation) to Dataverse and Zenodo is Wholetale (Chard et al., 2020).

[11] `codeocean.com`, accessed April 3, 2023.

8 Reproducibility services

When journals or authors create deposits at repositories, such *proposed* replication packages may be accepted as-is, or may be verified to assess their actual reproducibility. Some journals conduct reproducibility verifications in-house, others may outsource to organizations. The scale of the effort may also depend on the scope of tasks handled by such a team. Some journals may handle the deposit process for authors, others may require that authors self-deposit. For instance, assignment of a DOI to a deposit by `Codeocean.com` signals that their team has verified computational reproducibility. The Odum Institute conducts computational reproducibility checks on behalf of several journals in political science (Christian et al., 2018). Similarly, `cascad.tech` conducts reproducibility checks for openly accessible data, or data where access rules require physical presence in Europe (e.g. Swedish data), and subsequently issues a reproducibility report. Furthermore, through an arrangement with the French secure data access service `casd.eu`, they are able to access restricted-access French administrative and survey data, and conduct reproducibility checks for papers that use such data (Pérignon et al., 2019).

In-house reproducibility checks will depend on the scale of the journal process. Some journals may be able to conduct such checks with the help of a single part-time graduate assistant. Others, on the other hand, may have a team of associate editors (e.g., JASA) or a team of reproducibility analysts working under a designated data editor (as is the case with three of the authors of this manuscript).[12]

9 Choosing a repository

Given the preceding facets, journal editors may be consdering the following practical questions:

- What are key technical aspects that I should consider?
- What licenses should I allow?
- How will this be tracked at the journal?
- What is the cost going to be?
- Should I work with a dedicated repository, or allow, or require, deposit at a general purpose repository?

9.1 Key technical aspects

Sansone et al. (2020) propose several criteria which are designed to assist journal editors and managers to select repositories more generally. While this kind of assessment is likely to be a one-time activity when choosing a dedicated repository, it is a non-trivial exercise in time and skill to conduct when allowing for third-party repositories. Curated lists (PLOS One, 2021; Nature - Scientific Data, n.d.) can help, but are often incomplete when, as pointed out earlier, repositories are discipline-specific and not embedded in the global "research data management" community.

More generally, however, the criteria selected Sansone et al. (2020) are heavily data-centric, and do not encompass the full spectrum of needs for a policy that emphasizes computational reproducibility checks. We discuss these in more detail. Based on our collective experience, we rate these as "required," "optional," or "good to have."

[12] See Vilhuber et al. (2022*a*,*b*) for a description of the training required for a "replication lab."

Data curation As mentioned earlier, data curation is a value-added service. Generally, it can take significant time, and can be expensive. It may also impose a substantial burden on authors, which may not be appropriate when there is much shared knowledge within the discipline on these data. This should not imply that no data documentation should be requested from authors, only that the level of required documentation may be flexible. **Recommendation: Optional.**

User support Related to data curation is the support to authors depositing their materials at the repository. Depending on the complexity of the workflow, this may be troublesome for a smaller or larger number of authors. It may be sufficient, depending on number of articles and availability of staff, to have a knowledgeable staff member at the editorial office. This will be an option that depends on the cost of user support. **Recommendation: Optional.**

Pre-publication (private) access General purpose repositories do not uniformly provide the ability to have private or anonymous pre-publication access to a deposit. Based on our personal experience, among generalist repositories, openICPSR can provide non-anonymous access to repositories prior to their publication, Dataverse can provide single-blind private access. Figshare and Dryad also appear to offer such an option, possibly as part of a paid options. Zenodo does not offer pre-publication access. **Recommendation: Good to have, but depends on journal policy**

Citation of related publication A key aspect is the connection between manuscript and the associated replication package. Both directions (link from manuscript to replication package, and link from replication package to manuscript) should be present, should rely on DOIs, and should be recorded in the DOI metadata (technically, the `RelIdentifier` tag). A particular question is how difficult it is for the link to be updated after publication of the replication package: is there an API, can this be done manually, or is it not possible? Most current repositories that we have experience with require administrator or depositor control over the deposit, and none currently have the ability to auto-discover such links (based for instance on metadata in the Crossref DOI registry) **Recommendation: Required**

More generally, it is also theoretically possible to identify other publications that also cite the replication package. After all, the whole point of re-use is that data and code actually be re-used. Identifying such usage is best done via citations, and showcasing such reverse links is a key re-inforcement mechanism. Unfortunately, it relies heavily on either data citations being used at other journals, or on curation of data-specific bibliographies. **Recommendation: Good to have**

Programmatic download abilities Since the primary purpose of replication packages at journal repositories is to be re-used, the ability to access data would seem to be important. Naturally, all repositories support download of some sort. Repositories may require that users login (for free) to agree to terms of use, but not all do so. Some repositories offer up an application programming interface (API) to download data through programmatic commands, for instance in Python or R. Both API-based downloads as well as login-less downloads therefore allow for easy integration into program sequences. Depending on the typical size of deposits, a journal may prefer to have the latter capability. Currently, OSF, Zenodo, and Dataverse have APIs that are well documented. **Recommendation: Good to have**

9.2 What licenses to allow or require

Data Reuse Conditions and Copyright Transfer A non-negligible issue is the license that is affixed to replication packages, and relatedly, the copyright. Typically, a journal will request that copyright of the

manuscript be transferred from authors to the journal.[13] Traditionally, the replication package is subject to the same copyright transfer; however, this does not have to be the case. The AEA has removed that requirement (Vilhuber, 2019), and leaves copyright (and any liability) with the article authors (manuscript copyright is still transferred). The journal must then, however, require that the author make the replication package available under a license sufficient to achieve the goals of the data and code availability policy. Options range from limited licenses to open-source and open-access licenses (Stodden, 2009).

However, not all repositories allow for all licenses. Thus, the journal may want to select a repository that allows for a reasonable default policy (e.g., Creative Commons Attribution [CC-BY] licenses and their variants), but allow for the choice of other licenses. Ideally, such licenses are also captured by the DOI-related metadata, which is often not the case. Finally, there is little support in the metadata and across repositories for multiple licenses: the CC-BY is not appropriate for software, and Stodden (2009) suggests to use open-source licenses. Vilhuber (2019) describes the dual license suggested by the AEA, though it is not implemented by default at the AEA repository due to technical barriers. `Codeocean.com` defaults to separate licenses for code and data.

Journals may require varying levels of restrictions - *JHR* requires a public domain license, *Nature - Scientific Data* mentions "no unnecessary restrictions," but does not allow restrictions on commercial reuse, whereas the *American Economic Association* generically requires only that access for reproducibility purposes be possible as the minimal acceptable level. We note that this requirement, if implemented consistently and without exceptions, may limit the types of manuscripts that are submitted to a journal. For instance, most "public-use" survey data are not actually subject to a public domain, open data, or CC-BY license. US government data has by law been in the public domain for decades, but this has not been the case for many other countries' national agencies,[14] and is rarely the case for data created by and for academics. **Recommendation: Good to have**

Support for computation Sansone et al. (2020) do not address any further criteria, and yet for compound objects that include computational elements, these are as important. First on that list is support for computation that happens either on the repository platform, or is enabled via an API. For instance, among non-repositories, it is possible to instantiate a computational instance on `mybinder.org` from a repository at `github.com`, and run the code, without manual intervention (though with lots of setup on the part of the author). Such support is typically absent from repositories that focus on preservation. Exceptions are, to the best of our knowledge, the aforementioned `Codeocean.com`, as well as Wholetale in conjunction with `zenodo.org` or Dataverse deposits. **Recommendation: Good to have, but unlikely**

Support for online browsing and selective downloading of files Much of the transparency benefits accrue by being able to peruse the replication packages before downloading. Many of the ZIP files that are deposited on journal websites, under archaic deposit policies, have to be downloaded in their entirety before their contents can be inspected. Since sizes of such ZIP files range from several kilobytes to multiple gigabytes, this is not efficient. Most repositories allow to view individual files, but the support to view all types of files, in particular data files in common proprietary formats (Stata, R) or programs in statistical software languages (Stata, R, Python) are often limited. Even when it is possible to view the contents of ZIP files, it may not be possible to download or preview individual files from within ZIP files. Many of the code repositories support this much more robustly, allowing to preview modern formats such as Markdown, Python, or Jupyter notebooks natively on the platform, before downloading. **Recommendation: Good to have, but unlikely**

[13] At a minimum, a license to publish the manuscript is granted by the author to the journal, an option used by OA journals.

[14] See Statistics Canada (2012) and UK Government (2014) for recent open data licenses outside of the U.S.

At a minimum, the README describing the repository contents needs to be either viewable or downloadable separately in a variety of formats. **Recommendation: Required**

9.3 How will this be tracked at the journal

Ability to communicate with authors on deposit-related matters When identifying issues with code and data, can the journal editor (data editor) communicate with the authors through the normal channels used for manuscript (manuscript submission system), through a system dedicated to the deposit platform, or out-of-band (e.g., per personal email)? Does the system allow to record such interactions? **Recommendation: Required**

Control over content and publication When using a dedicated repository, additional decisions can be made as to how content is accepted, and how and when publication decisions are made. For instance, a person linked to the journal may need to accept requests to be listed in the journal community (Zenodo), or take action to publish a replication package submitted by authors within the journal's repository (openICPSR with journal workflow). This means that the repository publication process becomes part of the article's workflow through the journal's editorial system. Integrating into the editorial system may be manual, if the overall quantity of deposits is not too large, or automated (see below). **Recommendation: None**

A related issue are withdrawals and deaccessions. Deposits may be rendered inaccessible because it turns out they contain data that should not have been published (e.g., personally identifying information, or data subject to terms of use that prevent redistribution). Authors may, in some cases, retain the technical ability to change visibility or accessibility of deposits, for instance by changing the license.[15] When deposits are made at third-party repositories, it may be difficult to learn of or detect such changes. Control over such deaccessions may be higher at dedicated repositories, but journals may also want to put in place agreements with authors requiring them to notify a journal of such events, or preventing them from making spurious changes in the first place, for instance through archive agreements. **Recommendation: Good to have**

Ability to integrate with the manuscript or journal system Can the repository (dedicated or third-party) communicate with the manuscript or journal system? For instance, can deposits easily be linked to manuscripts through APIs or integrations, during the review process or the publication process? This can imply that the journal systems are notified that a deposit has been submitted for a particular manuscript, or in the other direction, triggering a coordinated publication of the replication package when the manuscript is published.

Such integration is offered for a fee for dedicated repositories hosted at Dryad and Figshare, but is generally absent for other repositories. The ability to do this kind of integration when using third-party repositories is currently not possible. We are not aware of a systematic review of such integration abilities. **Recommendation: Optional**

Branding The journal also needs to consider whether it wants a branded repository or not. ZIP files on a journal website are clearly associated with the journal. Most repositories allow for some superficial branding (providing a color scheme and a journal logo), some may allow for deep integration (repository as a service), in particular when the repository owner is the same publisher as the journal publisher (the big commercial publishers). Some repositories do not support anything but the most superficial branding (communities on Zenodo, collections at the Harvard Dataverse). **Recommendation: None**

[15] In general, authors do not have the ability to delete a deposit made in a trusted repository; see cautionary note about OSF, above.

9.4 Dedicated or not?

The question arises whether to require that authors deposit in a journal-specific repository, whether a looser association is envisioned, or whether any third-party repository is acceptable. To some extent, there are at least two high-level questions.

For one, from a data curation point, any object should be curated only once, and moved or copied only when curation is not possible in the original location. Thus, if a replication package is already deposited, say, at the Harvard Dataverse, it would not be acceptable to copy the package to openICPSR unmodified and assigning it another DOI. However, when the deposit is modified (versioned), while it is desirable that closely linked versions be maintained at the original repository, an argument can be made that the second version could exist at a different repository, as long as the two versions are linked. While supported by the existing metadata structure (DOIs are version-agnostic, but DOI registries contain the fields to identify prior versions), it is uncommon. This argues against moving replication packages between trusted repositories solely for the purpose of providing a journal-specific replication package.

The second question is whether to request data deposit at manuscript submission, or prior to publication (conditional acceptance). If the former is the intent, then it must be possible to access the replication package in a pre-published state, because if the manuscript is rejected and not published, the deposit must also be deleted. This can only happen when the repository allows for private (pre-publication) access. This may not be possible with all repositories. If the latter is the intent, then deposit may be more flexible, since the association with the specific journal is now known before the deposit is initiated.

Finally, the issue of branding, and association more generally with the specific journal comes to mind. While third-party repositories may be able to link back to the article, a journal can always link from articles to replication packages. A searchable archive of all the journal's associated replication packages is not a default feature, but can be easily accomplished if using the full capabilities of the DOI registries. The branding can be stronger or weaker. Commercial publishers sometimes have dedicated repositories that are tightly associated with the publisher, albeit not the journal. For instance, Mendeley Data[16] is an Elsevier property, and hosts replication packages for many Elsevier publications, without strong journal-specific branding. figshare[17] and ICPSR are known to provide heavily branded repositories, see for instance DataLumos[18] hosted at openICPSR, the National Archive of Criminal Justice Data (NACJD)[19] hosted at ICPSR, and the Carnegie Mellon KiltHub repository[20], hosted by Figshare. Lighter branding, which retains parts of the hosting repository's branding, is often available, see f.i. repositories hosted at openICPSR[21], or simply "collections" (Dspace), "communities" (Zenodo[22]), "dataverses" within the Dataverse software hosted at an institution, or similar. **Recommendation: None**

9.5 Cost

A key parameter is, of course, the cost of hosting replication packages, including support, superficial or in-depth verification of packages, and integration. If allowing for any third-party repository, the cost for hosting the packages tends to be zero for the journal (authors may have to carry some costs). A dedicated repository need not cost more money. Both Harvard Dataverse and Zenodo allow for some minimal branding for free,

[16] `https://data.mendeley.com/`, accessed April 3, 2023.
[17] `https://figshare.com`, accessed April 3, 2023.
[18] `https://www.datalumos.org/datalumos/`, accessed April 3, 2023.
[19] `https://www.icpsr.umich.edu/web/pages/NACJD/index.html`, accessed April 3, 2023.
[20] `https://kilthub.cmu.edu/`, accessed April 3, 2023.
[21] `https://www.openicpsr.org/openicpsr/repository/`, accessed April 3, 2023.
[22] `https://zenodo.org/communities`, accessed April 3, 2023.

and relatively generous amounts of storage for deposits. Deeper integration does come with an additional cost.

10 Implementing the linked repository

Once a repository model has been decided upon, journals will need to implement and transition to the desired repository scheme. If using author-initiated deposit, guidance to authors is strongly recommended. Some guidance can be found at `https://social-science-data-editors.github.io/guidance/`. Guidance to staff members tasked with monitoring and following the deposit process is also required. Legal counsel may be suggested for data and code archive agreement.[23]

Finally, unless the journal's repository is a fresh start, journals may want to consider migrating historical replication packages from the journal website to the new repository. Such a migration can be challenging due to the potential size and quantity of deposits. The migration should retain information about the original date of publication, authorship, and other metadata, and will need to be relinked to the article landing page. We refer to the AEA's migration for a particular large-scale example.

11 Case Studies

11.1 American Economic Association and openICPSR

On July 16, 2019, the AEA announced an updated and modernized "data and code availability policy" (American Economic Association, 2019), also published as American Economic Association (2020), in particular to bring the policy in line with the Findable, Accessible, Interoperable, Re-usable (FAIR) principles (Hagstrom, 2014). At the same time, the AEA began using a branded data and code repository hosted by ICPSR. The AEA selected ICPSR for a number of reasons.

As all trusted repositories, deposits receive their own DOI, and can be cited on their own. Deposits are tagged with keywords (e.g., "Current Population Survey" or "behavioral study"). Deposits are findable through [24]Google Dataset Search or through DOI registries such as [25]DataCite.

Importantly for the AEA were some additional criteria. ICPSR has data-centric approach, and authors can provide additional structured methodological information, such as the time period or geographic region covered by the data collected, or the survey method used. This information is encoded into the archive's metadata, and is used by various search engines, such as the native search engine on ICPSR. Furthermore, ICPSR was able to modify the metadata structure and user interface to allow for JEL codes to be added, which are the standard classification in economics. It is thus possible to search for data deposits with a discipline-specific taxonomy. This was not possible at the time with other trusted repositories.

A second key criterion was the ability to have pre-submission access to deposits. ICPSR has a robust system of sharing unpublished (draft) deposits. The AEA Data Editor can thus inspect and verify the deposits before they are published. ICPSR also added an unpublished metadata field that allowed for manual linkage to the manuscript identifier in the journal management system. While a more structured integration with the journal management system would have been preferable, this second-best solution was deemed sufficient.

[23] An example can be found in the form of the AEA Data and Code Archive Agreement - `https://www.aeaweb.org/journals/forms/data-code-archive-agreement`.

[24] `https://toolbox.google.com/datasetsearch`, accessed April 3, 2023.

[25] `https://search.datacite.org/`, accessed April 3, 2023.

We note that even if the AEA and ICPSR had been able to implement a structured approach on the repository side, the journal management system itself (ScholarOne) did not support a more structured approach.

ICPSR also developed a journal-centric workflow that gave the journal manager, and not solely the author, an active role in the publication process on the repository platform. Dataverse has been successfully integrated into certain journal workflows,[26] but no such integration into ScholarOne or other journal workflows was available at the time the AEA chose its platform, and there was no equivalent two-party workflow on the platform itself. Zenodo has no support for such mediated publication. The AEA has been a key driver behind the journal-centric workflow on openICPSR, and it is hoped that many features implemented and lessons learned can be leveraged by future adopters of the platform.

A final facet which the AEA assessed to be important was the ability to create and preserve hierarchical data and code structures. In prior work, the AEA Data Editor had observed that the vast majority of authors have an often deep directory structure to their replication package. At a minimum, code and data are stored in separate directories, as per various best practices. Most repositories only support a flat file structure, leading to difficulties in making the code reproducible.[27] ICPSR was, at the time, the only trusted repository that supported such hierarchical structures.[28] We note that a branded repository on openICPSR is not free. The AEA pays a yearly fee for the usage of the repository, which is meant to cover both administrative effort and storage costs.

The AEA migrated the entire "back catalog" of replication packages to ICPSR - a first wave added nearly 100,000 files in 2,500 replication packages, totalling nearly 0.5 terabytes of data. In a typical year, the AEA processes about 500 deposits. About 15% of deposits are larger than 2GB, and 1% are larger than 20GB (Vilhuber, 2021).

The openICPSR repository, as many other repositories, continues to have its challenges. The user interface is of an older vintage, and some authors face issues with uploading large and complex packages. Many authors have their code and data in Dropbox, or their code in Github, but there currently is no way to conduct a "cloud-to-cloud" transfer between such systems. The absence of an application programming interface (API) for users to download files more robustly through code remains a (minor) issue.

Currently, the AEA requests deposits of code and data prior to final acceptance at the AEA Data and Code Repository, though deposits at other trusted repositories are accepted as long as the transparency criteria are satisfied. Detailed guidance to authors is available at `https://aeadataeditor.github.io/aea-de-guidance/`, providing them with step-by-step instructions.

11.2 Review of Economic Studies and Zenodo

The Review of Economic Studies (REStud) has had a Data Availability Policy since 2006, which mandated sharing research data and code as supplementary materials on the journal publisher's website. In practice, this amounted to a ZIP archive of code and data, with often minimal documentation. This was briefly reviewed by the Managing Editor handling the paper after acceptance, before uploading to the publisher website. This model had a number of shortcomings: the review process was not standardized and often remained superficial, and the packages were hard to find and explore, especially as the journal has switched publishers during this period.

REStud appointed its first Data Editor in 2019 to help with the enforcement of the existing Data Availability Policy and to conduct more thorough reproducibility checks. One of the first changes this meant to

[26] The integration into an earlier version of the Open Journal System (OJS) is described at `https://projects.iq.harvard.edu/ojs-dvn/home`, see also Castro and Garnett (2014).

[27] For an assessment of how much flattening the directory structure might matter, see Trisovic et al. (2021).

[28] Due to technical limitations, a small number of supplements that provide more than 1,000 files will still be partially zipped.

authors was the adoption of a trusted repository for replication packages. The primary motivation was to make packages more findable and accessible.

REStud selected Zenodo as a freely available trusted repository. Zenodo records contain rich metadata, which make them easily searchable. Records can be organized into a curated list, called a "community." This way, all the packages relating to journal articles can be browsed together. Importantly, records receive their own DOI and can be cited in Data Availability Statements and elsewhere.

Zenodo also interacts with datacite and OpenAIRE. The first helps indexation by Google Dataset Search and other similar services. The latter helps authors acknowledge grant funding in a machine readable way. REStud also makes use of the download and view statistics collected by Zenodo as a key metric of replication packages.

Authors are asked to upload their replication package to Zenodo directly. Even though Zenodo can handle multiple files for each record, it does not allow for a hierarchical folder structure. Because of that, authors are required to upload a single ZIP archive. This limits readers in browsing individual files within the archive before downloading. This is a relevant constraint because many of the packages are several GB in size. Zenodo does show, however, a list of the files in the ZIP archive.

A key limitation of Zenodo is that records can only be shared after they are published. For the journal workflow this means that authors first publish their replication package then send its link to the Data Editor for review. If the Data Editor requests revisions (which is almost always the case), authors have to publish another version on Zenodo. This is a potential source of confusion for both authors and readers. The accepted version, however, is cited in the journal manuscript by its specific DOI, which fixes the version of record at the time of publication.

Another potential risk following from the inability to share incomplete packages is that authors may inadvertently publish data they are not allowed to. In that case, the Data Editor cannot "unpublish" the package, and authors have to turn directly to Zenodo to remove the package. More broadly, the repository model ensures that authors, not publishers or editors are responsible for the content of their packages. This is also made very salient to authors at the time of their package submission.

11.3 Economic Journal and Zenodo

After several years with a Data and Code Availability Policy in place, in 2019 the Economic Journal (EJ) started implementing pre-acceptance reproducibility checks. To coordinate and monitor the process, a Data Editor was appointed at that time. Originally, the system involved a submission of a ZIP file via the journal's platform and, after the necessary level of interaction between the Data Editor and the authors, the package would be published at the journal's website along with the article as "Supplementary Material". In 2020, EJ decided to move the publication of replication packages to an external repository. The motivation for the change was three-fold. First, it allows authors to retain the copyright of the contents of the replication package without sharing it explicitly with the journal, also allowing the list of authors to differ from that in the article, and moving the sole legal responsibility for the contents published in the replication package to the authors. Second, it was understood that publishing the packages in an external repository with its own DOI increases visibility to the articles and replication packages, and it makes it easier to cite them both. Third, it makes navigation through replication packages easier, providing extra branding and visibility to the journal as a whole.

With these premises in mind, and after some search, Zenodo was found to be the most appropriate repository given the journal's needs. Several considerations motivated this decision. Being a society journal, the cost was one of the main considerations: Zenodo is free. EJ editors also valued that Zenodo allowed the journal to create a community curated by the Data Editor in which all packages appear together, creating

some branding and providing easy navigability across packages of papers published at EJ. Obviously, Zenodo provides the standard requirements for being considered a "trusted" repository in the terms described above (perpetual depositing, assigning DOI, etc.). Finally, EJ editors also appreciated that Zenodo is quite user-friendly and very easy to explain to authors, which makes the deposit an effortless final step for the process.

The current workflow still keeps authors submitting their packages as a re-submission via the journal's platform, and the interaction between the Data Editor and the authors is kept as in the original system as well. The key difference is that, once the package has gone through the reproducibility checks (and some other anti-plagiarism checks), and the paper is ready for acceptance, the authors are requested to publish the finally accepted package on Zenodo, under the EJ's community dedicated to that effect.

11.4 Canadian Journal of Economics and Borealis, the Canadian Dataverse Repository

The Canadian Journal of Economics (CJE) adopted a data availability policy in 2008.[29] In 2015, the Executive Council of the Canadian Economics Association (CEA) approved the formation a committee to review this policy. This committee recommended the creation of the post of Data Editor. The role of the editor at that point was to enforce the existing data availability policy as well as shape it moving forward. The deposit of replication packages at the time worked as follows: Authors were asked to send a ZIP file with their data package and, after approval by the data editor, the ZIP file was made available on the paper's page (under "Supporting Information") on Wiley's website. This process had several drawbacks: data curation was not guaranteed, the contents of the ZIP file are not (easily) searchable, the journal archive itself is not available in one location and searchable (one has to go through individual papers' pages), and the deposit on Wiley's website is not incorporated in the journal system.

The journal thus decided to look for a repository that would help solve some of these issues. Being based in Canada, it was important for the CJE to look for a Canadian repository, which helps, among other benefits, justify future funding through national funding agencies. A natural choice was to turn to Borealis, the Canadian Dataverse Repository[30], a service of the Ontario Council of University Libraries that provides a shared technology infrastructure and shared collections to university libraries in Ontario and across Canada. Borealis hosts a Dataverse, the open-source repository platform first developed at Harvard University's Institute for Quantitative Social Science.

In 2021, the Canadian Journal of Economics set up its own Dataverse on Borealis, with the support of the McMaster University Library. Because CJE is the first journal to have its own dedicated journal repository on Borealis, the Canadian Dataverse Repository, CJE is contributing back experience and information that will hopefully for other journals in the near future. Borealis provides the CJE Dataverse for free. The process around the Dataverse creation was relatively smooth and quick. As soon as it was created, the CJE instructed authors to submit their replication package to the Data Editor directly on Dataverse, greatly improving the workflow around package verification. A more tedious task was to migrate all previous replication packages. In order to become familiar with the deposit process and because the data files were historically not organized in a consistent fashion, the migration was performed manually, one by one. Overall, while still relatively recent, the Canadian Journal of Economics' experience with Dataverse has been very positive. The platform is easy to use and authors have not given negative feedback. Managing the files on Dataverse is much easier than sending over email, which was the previous method. Each package comes with its own DOI, which is

[29] The CJE was published at the time by Blackwell Publishers. In 2012, the journal moved to Wiley.

[30] Until June 2022, Borealis was known as Scholars Portal Dataverse. Though the name of the repository changed, DOIs remain the same.

then given to the journal's publisher to put on the article's page. As of the writing of this manuscript, the CJE Dataverse hosts 240 replication packages, comprising 10727 files.

12 Final words

The needs of journals vary according to their specific policies, available resources, and internal organization. The authors' experiences, handling the replication packages (research compendia) for four different associations on three different platforms, can cover only a part of the possible scenarios. None of the journals that use the repository the authors manage have double-blind review, and none of the journals do more intense scrutiny of the data itself, letting the articles speak for the data. The repository experiences described here thus do not address those issues.

Each of the platforms described here, as well as all the others available, are in constant evolution. Decisions are made to implement a journal archive at a particular point in time, and this article is meant to provide assistance in making that decision, based on the four case studies the authors can speak for. Platforms add features, new integrations, and updated technology. Journals adapt their policies. Even the authors and journals represented here might conceivably make a different decision in the current, rather than the historical, environment. The authors nevertheless hope that both the criteria outlined here, as well as the decisions made, can help others make similarly well-informed decisions, as they create the infrastructure support of their journals in support of policies that improve transparency and reproducibility.

References

American Economic Association. 2019. "Updated AEA Data and Code Availability Policy." *AEA Member Announcements: Updated AEA Data and Code Availability Policy (July 16, 2019)*. https://web.archive.org/web/20191208160745/https://www.aeaweb.org/news/member-announcements-july-16-2019 (accessed 2019-09-21). tex.nonote: (accessed: 2019-12-08 via Archive.org) tex.notusedurl: https://www.aeaweb.org/news/member-announcements-july-16-2019 tex.timestamp: 2019-09-21T23:07:30Z.

American Economic Association. 2020. "Data and code availability policy." *AEA Papers and Proceedings*, 110: xxx. https://doi.org/10.1257/pandp.110.xxx.

Castro, Eleni, and Alex Garnett. 2014. "Building a Bridge Between Journal Articles and Research Data: The PKP-Dataverse Integration Project." *International Journal of Digital Curation*, 9(1): 176–184. https://doi.org/10.2218/ijdc.v9i1.311. Number: 1.

Chard, Kyle, Niall Gaffney, Mihael Hategan, Kacper Kowalik, Bertram Ludaescher, Timothy McPhillips, Jarek Nabrzyski, Victoria Stodden, Ian Taylor, Thomas Thelen, Matthew J. Turk, and Craig Willis. 2020. "Toward Enabling Reproducibility for Data-Intensive Research using the Whole Tale Platform." *arXiv:2005.06087 [cs]*. https://doi.org/10.3233/APC200107. arXiv: 2005.06087.

Christian, Thu-Mai, Sophia Lafferty-Hess, William Jacoby, and Thomas Carsey. 2018. "Operationalizing the Replication Standard: A Case Study of the Data Curation and Verification Workflow for Scholarly Journals." *International Journal of Digital Curation*, 13(1). https://doi.org/10.2218/ijdc.v13i1.555.

Gentleman, Robert, and Duncan Temple Lang. 2007. "Statistical Analyses and Reproducible Research." *Journal of Computational and Graphical Statistics*, 16(1): 1–23. https://doi.org/10.1198/106186007X178663.

Hagstrom, Stephanie. 2014. "The FAIR Data Principles." https://www.force11.org/group/fairgroup/fairprinciples (accessed 2018-05-20).

Journal of Human Resources. 2018. "Detailed JHR Policy on Replication and Data Availability." https://uwpress.wisc.edu/journals/journals/jhr_replication.html (accessed 2022-02-19).

Martone, M. (ed.). 2014. "Data citation synthesis group: joint declaration of data citation principles." https://doi.org/10.25490/a97f-egyk.

McCullough, B. D., and H. D. Vinod. 1999. "The Numerical Reliability of Econometric Software." *Journal of Economic Literature*, 37(2): 633–665. https://doi.org/10.1257/jel.37.2.633.

Nature - Scientific Data. n.d.. "Recommended Data Repositories | Scientific Data." https://www.nature.com/sdata/policies/repositories (accessed 2018-11-03).

PLOS One. 2021. "Recommended repositories." https://journals.plos.org/plosone/s/recommended-repositories (accessed 2021-02-10).

Pérignon, Christophe, Kamel Gadouche, Christophe Hurlin, Roxane Silberman, and Eric Debonnel. 2019. "Certify reproducibility with confidential data." *Science*, 365(6449): 127–128. https://doi.org/10.1126/science.aaw2825.

Sansone, Susanna-Assunta, Peter McQuilton, Helena Cousijn, Matthew Cannon, Wei Mun Chan, Sarah Callaghan, Ilaria Carnevale, Imogen Cranston, Scott Edmunds, Nicholas Everitt, Emma Ganley, Chris Graf, Iain Hrynaszkiewicz, Varsha Khodiyar, Adam Leary, Thomas Lemberger, Catriona MacCallum, Kiera McNeice, Hollydawn Murray, Philippe Rocca-Serra, Kathryn Sharples, Marina Soares E Silva, and Jonathan Threlfall. 2020. "Data Repository Selection: Criteria That Matter." Zenodo, https://doi.org/10.5281/zenodo.4084763.

Statistics Canada. 2012. "Statistics Canada Open Licence." https://www.statcan.gc.ca/en/reference/licence (accessed 2022-02-20). Last Modified: 2021-10-29.

Stodden, Victoria. 2009. "Enabling Reproducible Research: Open Licensing For Scientific Innovation." *International Journal of Communications Law and Policy*, , (13). http://web.stanford.edu/~vcs/papers/ijclp-STODDEN-2009.pdf (accessed 2018-10-05).

Trisovic, Ana, Matthew K. Lau, Thomas Pasquier, and Mercè Crosas. 2021. "A large-scale study on research code quality and execution." *arXiv:2103.12793 [cs]*. http://arxiv.org/abs/2103.12793 (accessed 2021-07-27). arXiv: 2103.12793.

UK Government. 2014. "Open Government Licence for public sector information V3." https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/ (accessed 2022-02-20).

Vilhuber, Lars. 2019. "Report by the AEA Data Editor." *AEA Papers and Proceedings*, 109: 718–29. https://doi.org/10.1257/pandp.109.718.

Vilhuber, Lars. 2021. "Report by the AEA Data Editor." *AEA Papers and Proceedings*, 111: 808–817. https://doi.org/10.1257/pandp.111.808.

Vilhuber, Lars, Hyuk Harry Son, Meredith Welch, David N. Wasser, and Michael Darisse. 2022*a*. "Teaching for large-scale Reproducibility Verification." arXiv 2204.01540v1. https://arxiv.org/abs/2204.01540v1 (accessed 2022-04-05).

Vilhuber, Lars, Hyuk Harry Son, Meredith Welch, David N. Wasser, and Michael Darisse. 2022*b*. "Teaching for large-scale Reproducibility Verification." *Journal of Statistics and Data Science Education*, forthcoming. https://arxiv.org/abs/2204.01540v1.