ISSN 1936-5330

kc FED Research Working Papers

Safeguarding Research: A Review of Economics Journals' Preservation Policies for Published Code and Data Files

Courtney R. Butler, Brett D. Currier, and Kira M. Lillard

December 2021
RWP 21-14
http://doi.org/10.18651/RWP2021-14

Federal Reserve Bank of Kansas City

Safeguarding Research: A Review of Economics Journals' Preservation Policies for Published Code and Data Files

Courtney R. Butler
Brett D. Currier
Kira M. Lillard

December 2021

Abstract

For many years, economics researchers have discussed the importance of sharing code and data files to ensure replicability. The discussion, however, rarely includes questions about long-term access to those files. This paper looks in-depth at the code and data policies from top economics journals to understand the guidance provided to researchers regarding data sharing and asks if that guidance supports preservation of code and data files for access and use, long into the future. We used content analysis to review journal policies from 184 economics journals. We discovered that while most journals recommend code and data be released with papers and that a few journals recommend practices consistent with long-term preservation, almost no journals specifically or emphatically consider long-term preservation of those files.

The data replication file is anticipated to be available in December 2021.

1. Introduction

For at least a century, making data available has been understood to underpin the scientific methods which the field of economics depends on. Writing in the 1933 inaugural issue of the journal *Econometrica*, Ragnar Frisch stated, "In statistical and other numerical work presented in ECONOMETRICA the original raw data will, as a rule, be published, unless their volume is excessive. This is important in order to stimulate criticism, control, and further studies" (3). At the time Frisch was writing, "data" was generally publicly available (Vilhuber 2019, 4), and sharing one's data was simply a matter of printing tables for anyone to read. Not only has sharing data changed from printing tables to posting files to the internet, the field of economics and certainly the technology employed by the field have also evolved since the time Frisch wrote. Yet many economics researchers continue to support the idea of data sharing. The reasons for doing so have varied widely over the years, from a desire to confirm existing studies (Hunt 1979), to understanding whether a theory will be true for the internet or another country (Faber 2002), to concerns about errors in quantitative analysis (Albright and Lyle 2010), self-correction as part of the research process (Andreoli-Versbach and Mueller-Langer 2014), seeking scientific truth (Anderson, et al. 2008), and more. However, across these many reasons is a common agreement summarized by Ben Bernanke's (2004) remark that "[r]eplicability is essential if empirical findings are to be credible and usable as a starting point for other researchers" (404). Which is to say, the empirical findings – employing code and data – are essential to the community of researchers who would use them in their own work, at a given point in time.

In 2003, economics scholar Teresa Harrison replicated an article[1] published in 2000 in the *Journal of Money, Credit, and Banking (JMCB)* by Daniel Thornton titled "Lifting the Veil of Secrecy from Monetary Policy: Evidence from the Fed's Early Discount Rate Policy." Thornton's article has been cited as recently as 2017[2], and, fortunately, the data files remain available via the Ohio State University iteration of the *JMCB* journal archives[3]. Conversely, the data archive link for the 2014 *JMCB* article by Chang-Jin Kim, et al. titled "Trend Inflation and the Nature of Structural Breaks in the New Keynesian Phillips Curve," which has been cited as recently as 2020[4], returns a 404 Page Not Found error[5], as do all data archive links for 2014 on the OSU *JMCB* journal archive pages. The files from Kim's "Trend Inflation" do not appear in either the Wiley or JSTOR iterations of the journal archives, and the journal was not able to provide these files upon request. Kim's research remains, and the data may be available on a personal or academic page. However, the *JMCB* Data Archiving Policy (Journal of Money, Credit and Banking, n.d.) states that acceptance of empirical papers is conditioned on authors providing their data and programs to facilitate replication. Because those files have not been maintained, replication is no longer possible via the journal archive.

---

[1] Harrison, Teresa. 2003. "Successful Replication of Thornton's (2000) JMCB Article." *Indian Journal of Economics and Business*, 2, no. 2 (December): 285.

[2] Fitoussi, Jean Paul and Jérôme Creel. 2017. "The European Central Bank and the 'Economic Government of Europe'." in *The Political Economy of the European Constitution*, edited by Luigi Paganetto, 180-193. London: Routledge. https://doi.org/10.4324/9781351145763

[3] See https://web.archive.org/web/20210107071448/https://jmcb.osu.edu/sites/jmcb.osu.edu/files/jarchive-v32n2pp155-167.zip

[4] Kamber, Güneş and Benjamin Wong. 2020. "Global Factors and Trend Inflation." *Journal of International Economics* 122, https://doi.org/10.1016/j.jinteco.2019.103265.

[5] See http://web.archive.org/web/20210123002712/https://jmcb.osu.edu/files/12-304%20Data%20and%20Programs.zip

Researchers interested in replicability and data availability have chronicled economists' increased willingness to share data and the number of journals that encouraged data sharing. In this paper, we looked beyond the code and data availability policies which journals developed, and we considered how journals handle long-term preservation of that code and data, if at all. We reviewed the policies of 184 economics journals to evaluate their data availability guidelines, to search for anything suggestive of a preservation policy and to discover whether the journals provided guidance to support long term access and use. We found that 143 journals of the 184 sampled publications provided a statement related to expectations for data – enough to consider it a data policy – while almost none had a formal preservation policy. Despite the absence of a formal preservation policy, about half of the journals recommended measures that may help the data remain discoverable. Significantly, however, almost no journals' policies suggest how the files themselves, once discovered, will maintain their technical integrity and be usable in the future.

2. Background

2.1 Data Availability

Over a 30-year period, economics researchers studied the growing interest in replication and data sharing. In 1986, a study examining articles under review or published in the *Journal of Money, Credit and Banking* found that 58% of authors were willing and able to share their data files upon request (Dewald, Thursby, and Anderson 1986). By 2015, a study by Chang and Li showed that data sharing had become more formalized with authors for 64% of articles from across 13 different journals providing data as part of the publication process. These studies confirmed what appeared to be a trend: economists were releasing their data more frequently over time. However, Robert A. Moffitt, Editor of *The American Economic Review*, noted in 2010

that while there had been slow and steady progress, “the desired ‘Culture of Replication’ and data sharing that would be optimal is still quite a way off” (Sedransk et al. 2010, 43).

At the same time, the economists advocating replicability of research also argued that professional publications have an obligation to require submission of code and data. Given the significance that publishing in prominent journals has for academic researchers’ careers, many in the research community have recognized the large role that journal policies play in incentivizing data sharing (Crosas 2018; Sturges et al. 2015). Lars Vilhuber, Data Editor of the American Economic Association, said he was “summarizing studies old and new” when he noted in 2020 that, “the probability of obtaining sufficient data and code to actually attempt a reproduction is lower when no formal data or code deposit policy is in place” (11), and journals have gradually developed such policies. For example, the American Economic Association joined a small but influential group of journals and institutions when it implemented its mandatory data availability policy in 2004, and 14 years later, a study found that 74% of economics journals had some form of data policy in place (Crosas, et al. 2018, 9).

To this day, economists continue to discuss issues in sharing data, weighing the demands that replication imposes on economists as well as obstacles posed by changes in data technology and dissemination. The conversation has evolved from agreeing on data sharing in principle to grappling with its complex challenges in more practical terms, among them: the diversity of software programs, complexity of data formats, legal and ethical constraints, and the lack of uniformity in submission or storage requirements at the journals or the repositories (Duvendack, Palmer-Jones, and Reed 2015; Sturges et al. 2015; Vlaeminck and Herrmann 2015). Today, the researcher must contend with a wide range of practical issues, the through line of which is concern with long-term discovery, access, and usefulness of replication files. Though journals

have implemented code and data availability policies in the field of economics for almost 40 years, Chang and Li (2015) concluded that "economics research is generally not replicable" (11), largely due to the "missing data or code" for the majority of the 60 papers whose research they attempted to replicate.

While the issue of missing data and code that Chang and Li observed continues to be a pressing issue for the non-replicability of economics research, we asked the question: what about the papers which are replicable? For the researchers who agree with the principle of replicability in general, comply with data availability policies, and submit their code and data for publication, what happens to their code and data files over decades?

2.2 Data Preservation

Denis Huschka (2013) listed the lack of long-term preservation among the challenges researchers and publications face related to replication, noting that "[p]utting data on CDs and flash sticks is NOT long-term preservation!" (8). In fact, according to Crosas (2018), "the basis of all data policies is the need for data access and preservation" (3). This need is particularly relevant for economics where the work of influential economists is cited and referenced for decades and sometimes centuries after publication (Anderson, Levy, and Tollison 1989). According to Web of Science's 2019 *Journal of Citation Reports* (JCR), the aggregated cited half-life of economics journals is greater than 10 years[6]. This means that, on average, the majority of economics articles – even less prominent ones – are in publication for more than 10

[6] The Journals of Citation Reports caps the cited half-lives at 10 years. Items scoring more than 10 years are reported as >10. See Clarivate (2017).

years by the time they receive half of their total citations. By extension, it is not unusual for an economist to utilize another economist's 10-year-old research in their own work; that economist may expect, reasonably, that the accompanying code and data is available, accessible, and usable. Thus, researchers who advocate for the importance of replication should also consider the mechanisms that need to be in place to ensure associated data remains available over time. Yet, we found few examples of replication and data availability discussed in the same breath as preserving data. Even where the word "archiving" appears, it may not connote much more than data storage and usually does not point to discernible preservation measures.

For example, McCullough, McGeary, and Harrison (2008) pointed to several cases of failed replication that had significant potential to impact public policy including the Phillips Curve, the relationship between police staffing and crime levels, the relationship between a woman's lifetime earnings and her number of children, and the size of the underground economy. They suggested that the issues associated with each of these would have been significantly reduced or eliminated if the publishing journals maintained a mandatory data+code archive (1409–1411). They examined four journal archives to determine if the research in them was replicable and discovered issues related to non-compliance with data sharing requirements, data-only submissions that did not provide accompanying code, and incomplete or insufficiently documented packages. Their recommendations for an effective archive included points related to documentation and file formats, but there was no discussion related to ensuring the integrity and continued accessibility of the digital files or reference to preservation tools, methods, or standards. A number of studies since have continued to focus on the enforcement of journal archives, often with mentions of persistent and discoverable links but, again, with no discussion

of preservation standards or requirements (Anderson, Greene, McCullough, & Vinod, 2008; Fear, 2015; Vlaeminck & Herrmann, 2015).

This was not for lack of discourse around the issue; the National Library of Congress began the National Digital Information Infrastructure and Preservation Program in 2000 (Library of Congress, n.d.), and in 2005 a consensus of academic librarians, university administrators, and others who participated in a meeting to discuss electronic journal preservation at the Andrew W. Mellon Foundation offices released a statement titled "Urgent Action Needed to Preserve Scholarly Electronic Journals" that stated:

> "In field after field, research and teaching are generating data, reports, publications, teaching materials, and other forms of scholarly communication in digital formats. Research and teaching are also increasingly dependent on data mining tools and other computer-based techniques that require the long-term persistence of these various forms of digital information to advance knowledge. Yet as the creation and use of digital information accelerate, responsibility for preservation is diffuse, and the responsible parties—scholars, university and college administrators, research and academic libraries, and publishers—have been slow to identify and invest in the necessary infrastructure to ensure that the published scholarly record represented in electronic formats remains intact over the long-term. Inaction puts the digital portion of the scholarly record—and the ability to use it in conjunction with other information that is necessary to advance knowledge—increasingly at risk[.]" (Association of Research Libraries 2005)

This risk has been realized in very real ways. Though McCullough et al., (2008) did not discuss missing or corrupted data in their study, they did reference Dewald, Thursby, and Anderson's 1986 replication of two articles published in the *JMCB*. One of the two articles

replicated was Robert Engle's 1983 paper, "Estimates of the Variance of U.S. Inflation Based Upon the ARCH Model," which continues to be cited in 2021[7]. While it is unclear if that data remained available in 2008 when McCullough et al., were writing, it is presently unavailable via the *JMCB* archive. Much like Kim's 2014 data files, Engle's files do not appear in either the Wiley, JSTOR, or Ohio State University iterations of the journal archives, and the journal was not able to provide these files upon request. We should note, however, that while the *JMCB* archive is frequently studied due to its proactive efforts in this area over the last 40 years and is an illustrative example of the fragility of data files over time, it is certainly not the only economics publication to face such challenges, and observations should not be construed as pointed criticism.

Still, even more recent literature on replication for publications continues to discuss the issue of preservation in opaque terms. For instance, Willis and Stodden (2020) found that most of the reproducibility efforts they studied required long-term accessibility of artifacts, but these requirements were not defined. Instead, the discussion focused on the observation that reproducibility initiatives commonly "rely on established repositories for artifact preservation, stewardship, and long-term access" (14). Crosas et al., (2018), while still not explicit, came closer when examining whether economic journal policies directed authors to share data in a repository to ensure the files remain FAIR (Findable, Accessible, Interoperable, and Reusable) (12) and in recommending more collaboration among journals, publishers, and associations to

[7] Liu, Fred, and Lars Stentoft. 2021. "Regulatory Capital and Incentives for Risk Model Choice under Basel 3." *Journal of Financial Econometrics* 19, (1): 53-96. https://doi.org/10.1093/jjfinec/nbaa029.

promote the inclusion of data sharing best practices from information science and data professionals (18).

Overall, the literature we reviewed underlined the importance of journals' data policies and enumerated many of the difficulties in enforcing the data policies. What we did not find in the literature, by and large, was awareness of the connection between replication and what should happen after journal policies were complied with and the code and data was shared and available.

3. Methodology

3.1 Research Question

Using a snapshot taken in August 2019 of the journal policies in our sample, this paper asked: what are journals doing today to support access and use of code and data files at least 10 years from now?

3.2 Approach

We used content analysis as the research method for this paper to study journal and publisher policies in an unobtrusive and context sensitive vacuum (Yoon and Schultz 2017, 923). Rather than counting the number of times the words, "data availability," appeared in a paragraph, we read in and around the context of the entire policy and attempted to tease out the implied meaning of the words. We read the policy content from beginning to end, in order to understand what the language implied as well as what was explicitly outlined as guidance for the researcher to follow.

3.3 Sample

We identified a sample of top economic and business journals to have a pool of publications covering diverse research areas that were also representative of the most highly

respected, influential, and widely read journals. We consulted three separate sources that each used distinct ranking criteria (see Figure 1). While one source would likely have provided an acceptable list of leading journals, using three helped ensure we were reviewing the standard-bearers in this field.

The first source we consulted was Google Scholar, which ranked journals in topical sub-categories using the h-5 index. The second source was Web of Science, which ranked journals in a single unified list based on selected keywords using impact factors. We filtered Google Scholar and Web of Science using a broad range of categories, trying to capture a wide universe of economics- and business-related titles. Examples include accounting and taxation, business, economics, finance, marketing, and probability and statistics with applications. Finally, the third source we consulted was IDEAS/RePEc, which is dedicated to economics and related disciplines and did not require any filtering. IDEAS/RePEc ranked journals in a single, unified list using a combination of simple impact factors, recursive impact factors, discounted impact factors, recursive discounted impact factors, and the h-index.

From these three sources, we initially identified 772 journals: 272 from Google Scholar, 250 from Web of Science and 250 from IDEAS/RePEc. More journals were included from Google Scholar because of its sub-category structure versus the single unified list structure of Web of Science and IDEAS/RePEc. We took the top 20 journals from each selected sub-category in Google Scholar, combined them into a single list, and de-duplicated to arrive at 272. This unified list was then sorted by h-5 index.

All journals ranked within the top 50 by each source were automatically included in the preliminary sample. In addition, all three sources were cross-referenced, and any journal that

appeared more than once, regardless of ranking, was also included. This sample was then de-duplicated, and the remaining 212 journals were selected for eligibility review.

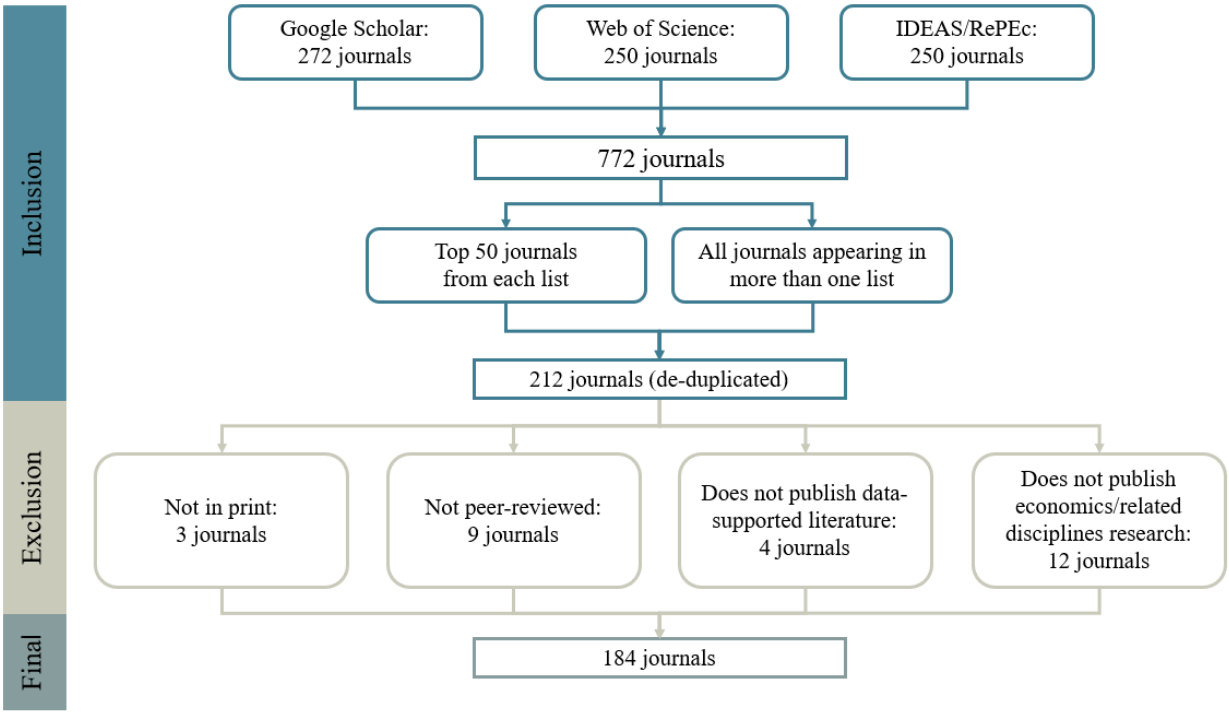

Figure1

Within the month of August in 2019, we downloaded all the relevant content for analysis from the public websites of all 212 journals; content types included, but were not limited to: editorial statements, instructions for authors, manuscript guidelines, and formal policies. We downloaded publisher documentation only when it was referenced directly within a journal's documentation. We did not reach out to any journals or publishers directly.

We then reviewed the preliminary sample for eligibility. Because this study focused on point-in-time data policies for academic economic journals, a journal was only considered eligible if it met all the following criteria:

- In print at the time of data collection;
- Peer-reviewed;

- Publishes data-supported literature (as opposed to strictly narrative literature reviews, for example); and
- Publishes research specific to economics, business, or other related disciplines.

A total of 28 journals lacked one or more of these features and were removed from further evaluation. The remaining 184 journals comprised the final sample.

3.4 Variables: What We Measured

Our variables fell within three categories: data availability policies, data preservation policies, and features consistent with preservation and long-term use. Content analysis allowed us to capture some of the nuances and complexity within these components. We anticipated that the exact language within our sample would vary widely, but the categories gave us enough structure to see differences across a wide spectrum of journals. To that end, we interpreted 'data' broadly to mean quantitative results, inputs and intermediary files, code, data availability statements, or any other materials that were intended to enable reproduction of results.

For the first category, data availability policy, we considered a journal to have a data policy if it provided any statement in its public-facing documentation about expectations related to data availability, archiving, or replication.

For the second category, data preservation policy, we considered a journal to have a data preservation policy if it provided any statement in its public-facing documentation about expectations related specifically to preservation or long-term access of data. The term 'archive' was not sufficient – it had to be explicit that the policy created some expectation of active, long-term management of the files. For instance, direction to maintain the files for a set number of years was not considered as having a preservation policy since 'not deleting' is not the same as proactively preserving.

Finally, our third category comprised elements of a journal's policy which connoted – or denoted – data preservation. This category required that we examine journal preservation practices in more detail, closely investigating the presence of features consistent with preservation. Even if no official preservation policy or guidance was present, we hypothesized that journals may still be encouraging practices consistent with preservation within their data guidance generally, even if it was done inadvertently or partially.

The third category was necessarily complex because we wanted to conceive of the category of 'preservation features' as broadly as possible. We chose preservation features based on the 2013 National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation as well as two features which fall under the rubric of 'reference rot': referrals to a data repository and the assignment of persistent identifiers.

About NDSA

The 2013 National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation are a tiered set of recommendations for organizations seeking to build or enhance digital preservation activities, typically within a repository. There are many standards we could have selected, and we chose the NDSA standard because we felt it offered broad categories that we could use as a benchmark.[8] We used all of their categories as our variables. Those variables are:

- storage and geographic location
- file fixity and data integrity

[8] As a note, we used the Version 1.0 of the NDSA's Level of Digital Preservation, as Version 2.0 came out after instrument testing and data collection had begun.

- information security
- metadata
- file formats

The NDSA categories are clearly defined with a number of ways to satisfy their requirements using a spectrum of maturity characteristics (NDSA, 2013). We did not analyze the maturity of practices within journals' policies, but the maturity characteristics provided a broad range of language we could use to identify these variables. For instance, the 'information security' category is described within the NDSA standard as the management and monitoring of files' read and write access for authorized individuals (Table 1). Thus, any journal that discussed active access management for data and code files was interpreted as including information security features consistent with preservation. Another example was the category of file formats. If a journal policy mentioned ASCII as a recommended or required format for submission of data files, we considered it to meet the first, most basic level of NDSA file format preservation because it defined the 'use of a limited set of known open formats and codecs.

About Reference Rot

The NDSA standard allowed us to target specific terms and practices to get at features of a policy consistent with preservation. However, we knew those standards were not sufficient to capture all the nuances of preservation features – features that do not resemble what they might have, only a few generations ago. While Frisch may have imagined code and data files living in a

physical or print archive, today's code and data files exist almost entirely online. As such, these files are subject to all issues that digital objects face, including link rot, reference rot, and bit rot[9].

Link rot, reference rot, and bit rot present challenges for long-term use and replication, but each can be mitigated through applying active preservation measures such as unique identifiers and file fixity checks and repairs (Baker, at al. 2006, 9; Zittrain, Albert, & Lessig 2014, 189). We characterized these preservation measures as the variables 'persistent identifiers' and 'referral to a data repository' because integrity monitoring and fixity checks are part of the Open Archival Information System (OAIS) standard, which is a key component of the CoreTrustSeal for Data Repositories[10]. It is worth noting here that these preservation measures – which serve as variables in our study – also correspond with guidance found in data management requirements from major funding institutions, such as the National Science Foundation[11] and the Economic and the Social Research Council[12].

[9] Link rot and reference rot directly influence a user's ability to find digital content. Link rot refers to a URL that no longer returns any content (404 errors are a common example), and reference rot, an even larger phenomenon, happens when a link still works but the cited information is no longer present or has changed (Zittrain, Albert and Lessig 2014, 177). Bit rot, on the other hand, is degradation of the storage medium (Baker, et al. 2006). File corruption is one of the most common manifestations of bit rot and results in a user's inability to open or use a file.

[10] Examples of Core Certified Repositories include the Inter-university Consortium for Political and Social Research (ICPSR), the Cornell Institute for Social and Economic Research, the UK Data Archive, the Roper Center for Public Opinion Research, the Qualitative Data Repository, Mendeley Data, and more. See https://www.coretrustseal.org.

[11] See https://www.nsf.gov/news/special_reports/public_access/

[12] See https://www.ukri.org/wp-content/uploads/2021/07/ESRC-250821-ESRC-Research-Funding-Guide-V2.pdf

Levels of Obligation

In trying to capture a journal's level of commitment to availability and preservation, we tried to measure the degree of seriousness or pressure the journal was willing to place on the authors who submitted research articles. We therefore further classified each policy – data availability and preservation – by identifying the level of obligation: required, recommended, or optional. A policy was considered required if the statement included non-ambiguous terms such as "must", "expects", or "required"; recommended if the statement included positive but less strict terms such as "should", "encourages", or "recommended"; and optional if the statement included neutral terms such as "may" or "optional." Journals that did not mention data at all or included statements explicitly discouraging sharing were considered as having no policy. The distinctions among required, recommended, and optional were constructions we imposed for the purpose of understanding the landscape of data policies. The distinctions were useful because they provided insight into journals' commitment to these practices as well as correlations between the level of obligation and any accompanying infrastructure that enabled long-term availability.

Each of the consistent-with-preservation variables was classified by obligation as well. Those variables did not need to be part of an official data or preservation policy to be included; they could appear anywhere in the journals' public-facing guidance.

Coding Procedures

We tested our coding procedures using the top five political science journals as reported by Google Scholar in September 2019: *American Journal of Political Science, American Political Science Review, Journal of European Public Policy, Journal of Politics,* and *Comparative Political Studies*. These journals were selected based on their relatively close

relation to business and economics as a social science, but none were included in the final sample.

The two reviewing authors scored all five journals independently, and the third author served as arbitrator. This process allowed final adjustments to be made and for all authors to develop a shared understanding of the intended measurements prior to beginning data analysis. Once scoring began, the first reviewing author independently reviewed all journals in alphabetical order while the second reviewing author independently reviewed in reverse alphabetical order to help mitigate chronological bias.

After both authors completed coding for all journals, we compared data to assess agreement. In cases where the coding was not aligned, the reviewing authors discussed the reasons for their scores, and the third author made final decisions. We utilized interrater reliability (IRR), a quantified measure of consensus among raters, to evaluate the trustworthiness of our analysis. For this study, IRR agreement was at or above 80%, which is considered a commonly accepted threshold (O'Connor and Joffe 2020, 9). The only exception was the persistent identifier variable, which had 53.8% agreement. After further review, this discrepancy was found to be based on template language present within a broad range of journals from the same publisher and did not reflect dissent across the overall sample.

4. Findings

4.1 Overall

Of the 184 journals we reviewed, 143 (77%) provided a statement related to expectations for data availability or replication in their public-facing documentation and thus were considered to have a data policy (see Figure 2). This is consistent with Crosas et al.'s (2018) findings of

74%. However, only 14 journals (8% of the total sample) explicitly mentioned preservation and were considered as having a preservation policy.

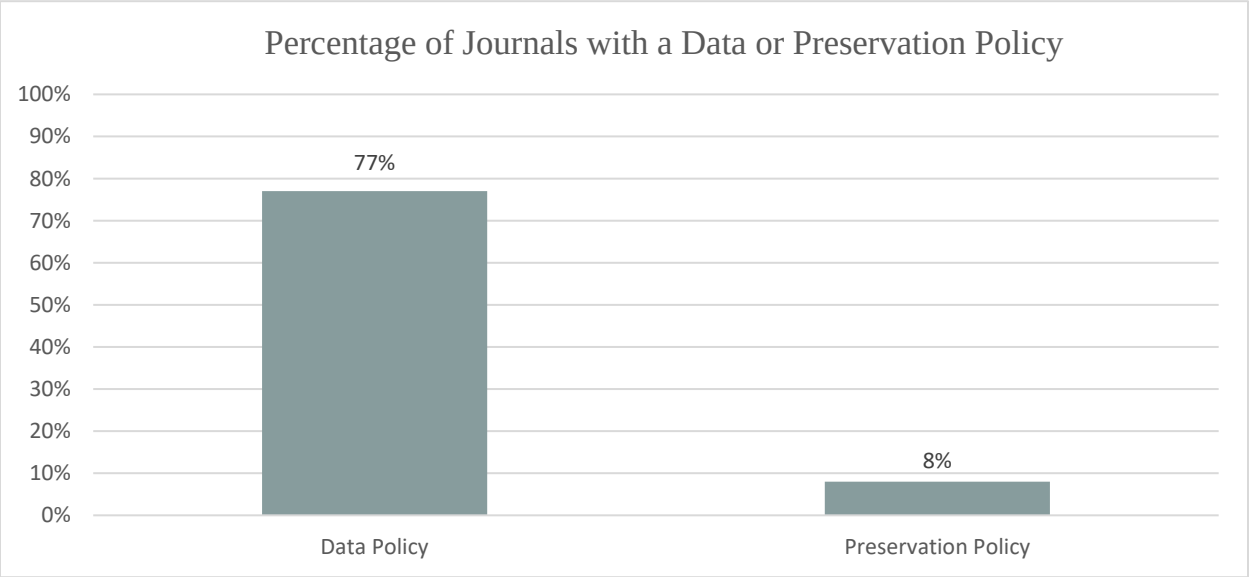

Figure 2

Though few journals codified preservation into policies, many had mitigations for reference rot built into their guidance, with 59% of journals pointing authors to a repository for data and code deposits and 47% discussing the use of persistent identifiers (see Figure 3). Conversely, the five NDSA variables had a more limited presence. Two of the variables – 'storage and geographic location' and 'file fixity and data integrity' – went unmentioned by any journal, and only one journal included a reference to information security. 'Metadata' and 'file formats' both had a slightly larger – though still small – presence within the sample with 15% (28 journals) and 7% (13 journals) respectively.

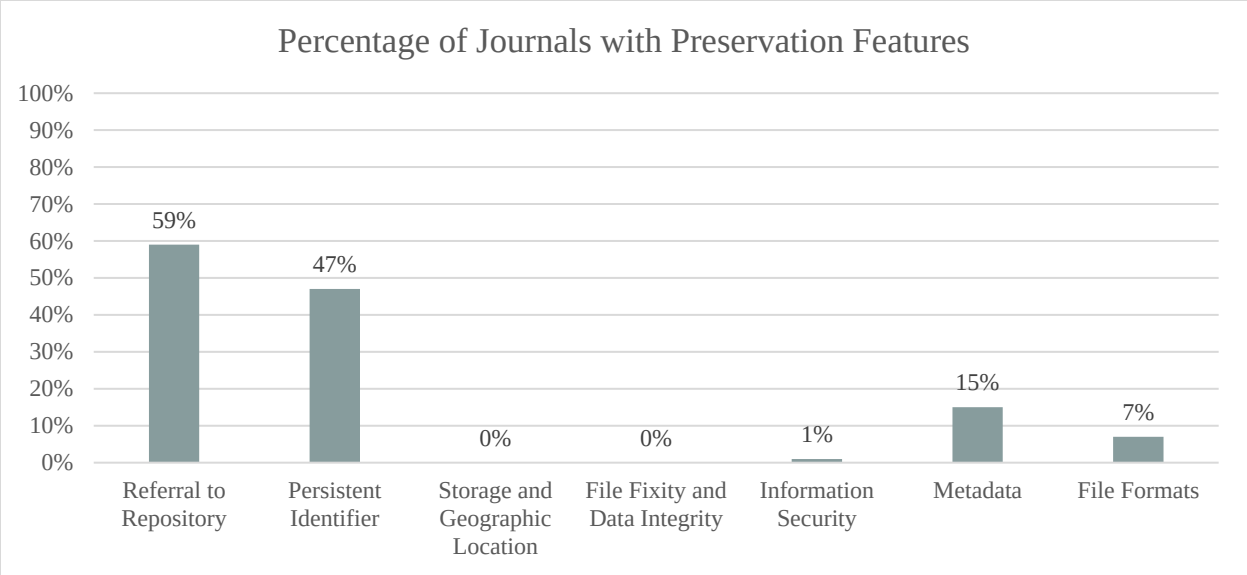

Figure 3

4.2 Obligation

As can be seen in Figure 4, data and preservation policies tended to be primarily recommended rather than optional or required. Of the 143 journals that had a data policy, only 34% of the policies were required while almost double that amount (60%) were recommended and only 6% had optional policies. Virtually all preservation policies were recommended with only one required policy and none that were optional.

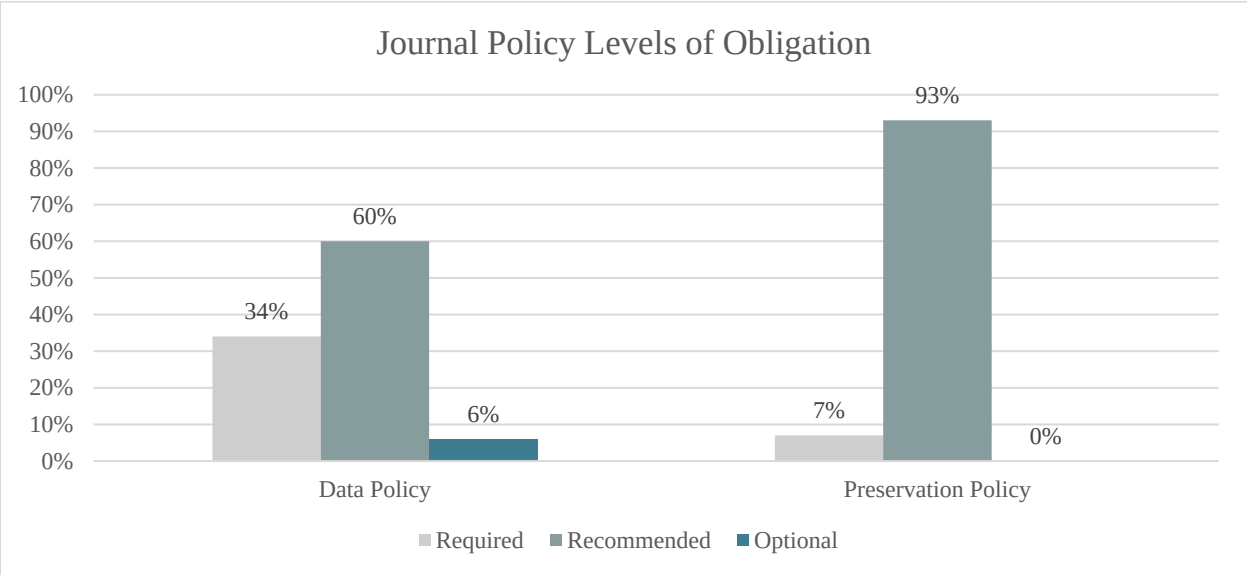

Figure 4

Guidance related to reference rot variables, on the other hand, was primarily optional. Of the 109 journals that referred authors to a repository, only 14% required deposit and 28% recommended it (see Figure 5). The remaining majority (59%) of journals provided optional guidance. Similarly, 83% of journals that discussed the use of persistent identifiers made the practice optional while only 7% required it and 11% recommended it.

The level of obligation for the NDSA variables did not reveal a distinct trend. Within the 28 journals that discussed metadata, 57% required authors to provide metadata and 43% recommended it. None of the guidance provided was optional. Only one of the 13 journals that mentioned file formats required authors to follow guidance while 77% (10 journals) recommended it and the rest made it optional. The only journal that discussed information security required compliance with their guidance. None of the journals discussed 'storage and geographic location' or 'file fixity and data integrity.'

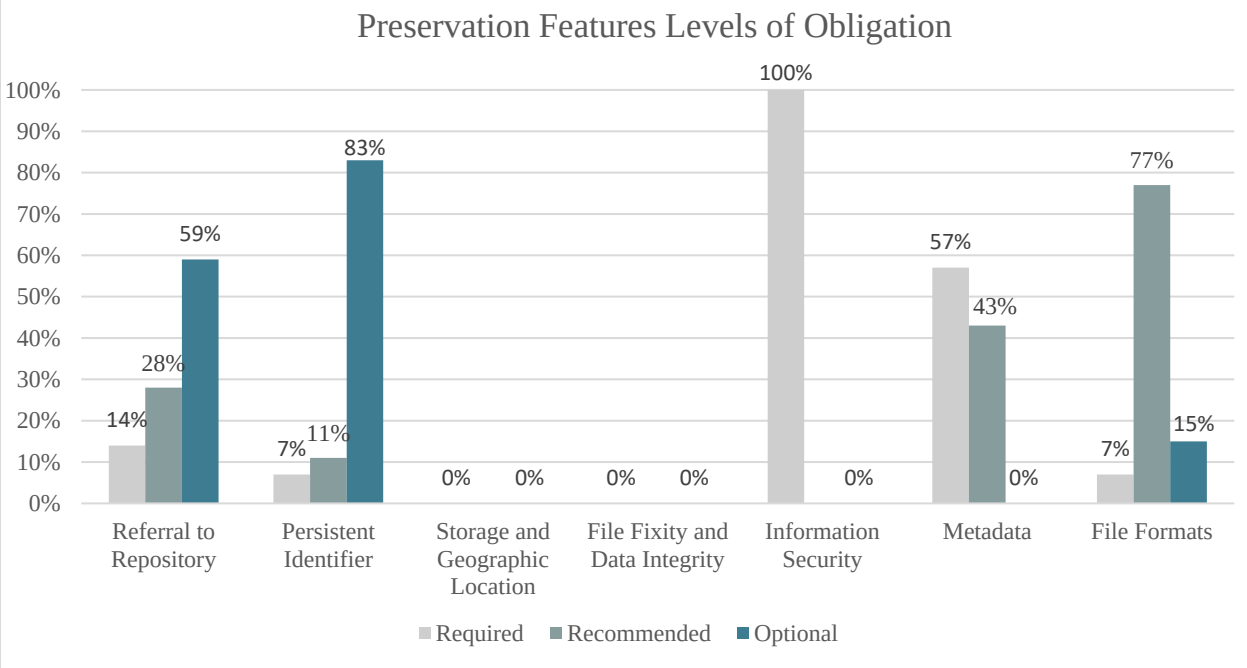

Figure 5

4.3 Relationship Between Policies and Preservation Features

As can be seen in Figure 6, approximately 10% of all journals that had a data policy also had a preservation policy. We observed a slight relationship between data policy obligation and the existence of a preservation policy; approximately 12% of journals that required data sharing also had a preservation policy compared to only 9% of journals that recommended data sharing. None of the journals that had optional or non-existent data sharing policies had a preservation policy.

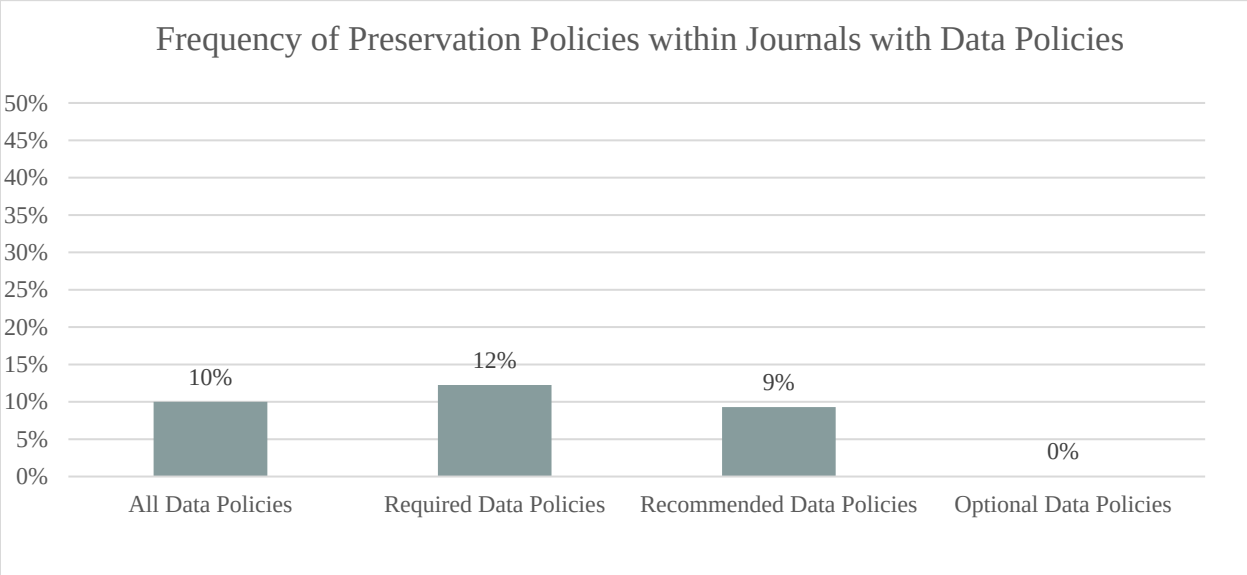

Figure 6

There also appeared to be a relationship between data policies and reference rot considerations (Figure 7). Of journals with a data policy, 76% referred authors to a repository and 60% discussed persistent identifiers. Over half of journals that required data sharing referred authors to a repository and one-third addressed persistent identifiers. The vast majority of journals that recommended data sharing addressed these issues also (93% for repositories and 85% for persistent identifiers). Even journals with optional data sharing policies addressed this issue relatively often, with 38% referring authors to a repository and 25% referencing persistent identifiers.

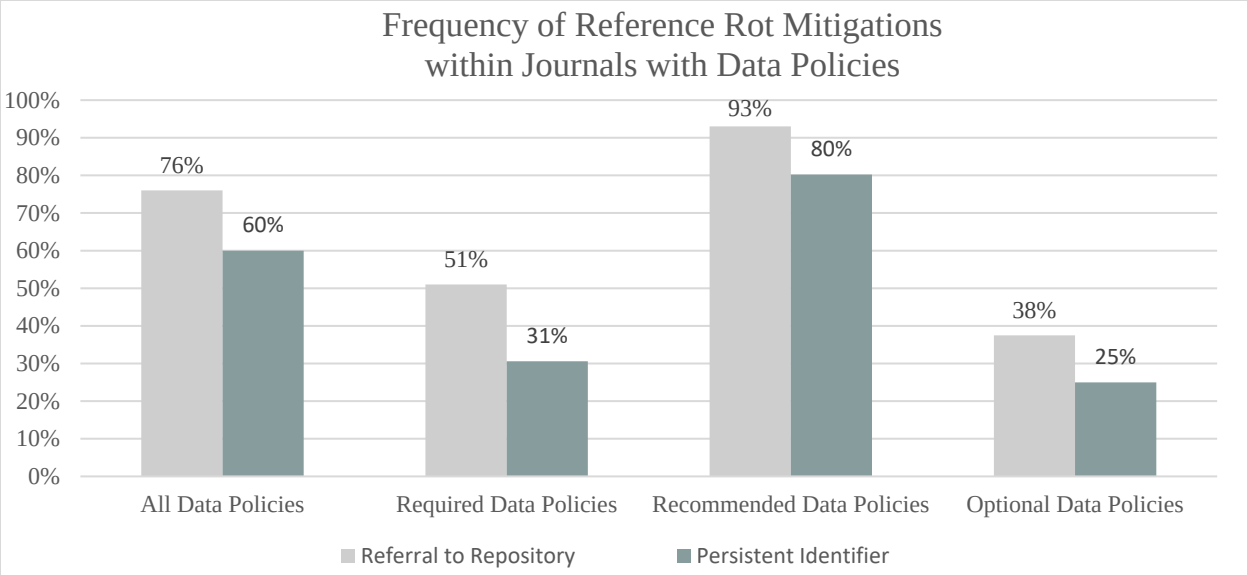

Figure 7

Significantly, we found that 100% of journals with a required or recommended preservation policy referred authors to a repository and 86% mentioned using a persistent identifier (Figure 8). This indicates that repositories and persistent identifiers are foundational to journals' preservation policies. There were no journals with optional preservation policies.

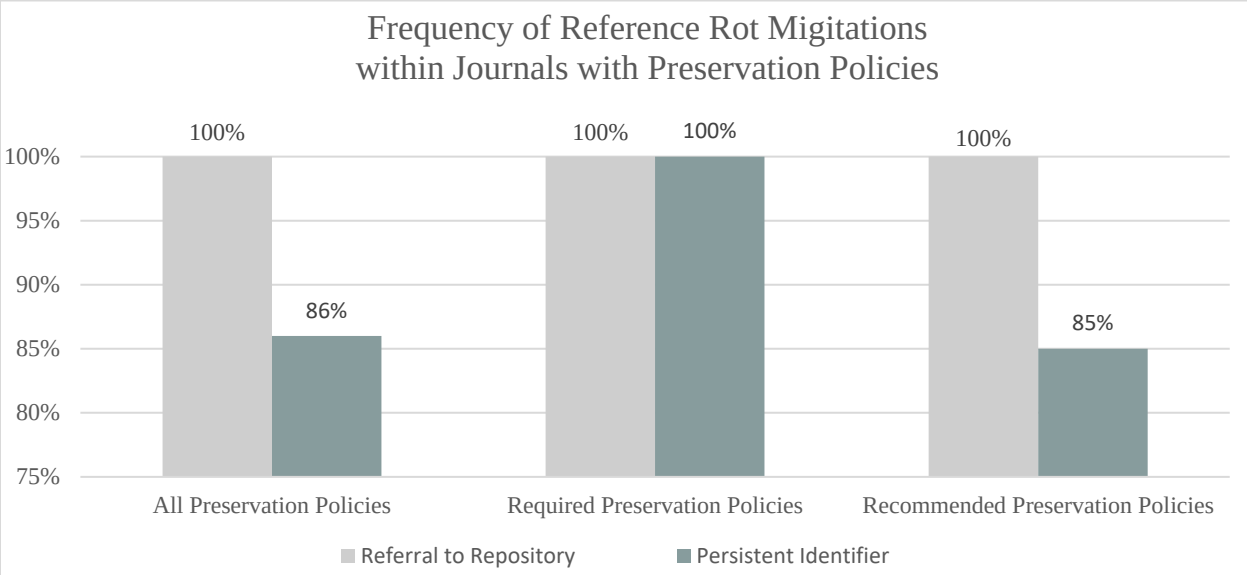

Figure 8

There was also a relationship between journals with required data sharing policies and the NDSA variables we applied, as can be seen in Figure 9. For example, 96% of journals that discussed metadata and 85% that discussed file formats also required data sharing. The only journal that discussed information security had both a required data policy and a recommended preservation policy, but no other journal with a preservation policy of any sort discussed any of the other NDSA variables (which may reflect journals' dependence on repositories for these practices). The variables 'storage and geographic location' and 'file fixity and data integrity' were not mentioned by any journals.

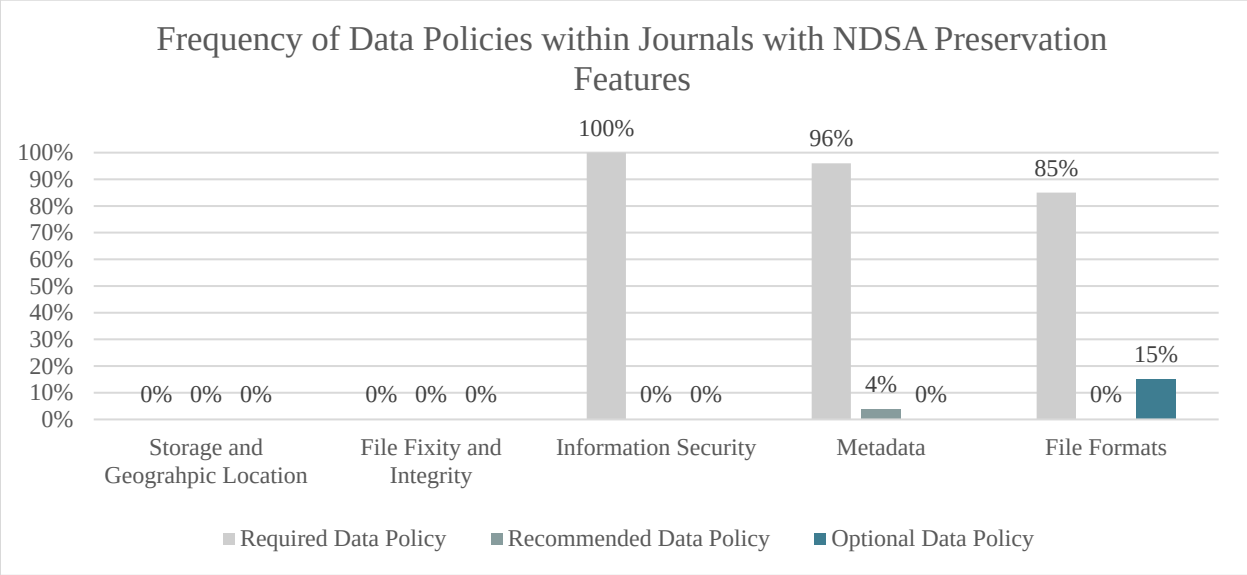

Figure 9

4.4 Summary

Our findings confirmed Crosas et al.'s (2018) conclusions that most economics journals have data availability policies. A data policy, however, does not necessarily extend to a preservation policy and our findings confirmed that also: only 14 journals (8% of the total sample) explicitly mentioned preservation and were considered as having a preservation policy. Our initial hypothesis that journals may recommend practices consistent with preservation was

borne out: nearly 60% of journals mentioned a repository and nearly 50% include mention of a persistent identifier. Yet, repository and persistent identifier are only two of the seven variables we examined to suggest data preservation practice. In short, we interpreted the policies as broadly as we could but found the practices consistent with preservation were few and far between.

5. Discussion

5.1 The Problem

Since 2017, when we first began research on this topic, we have seen an increase in journals with data policies and an increase, albeit small, in the number of journals with preservation policies (Butler and Currier 2017). Yet, where we saw guidance that could be interpreted as a preservation policy, journals only focused on short-term verification or reproducibility but not long-term replication and preservation. In other words, journals are encouraging data availability, economists are providing data, and as a result, more data is available for replication for future researchers. But if data is made available, is that sufficient? For how long will the data be available – and usable – for replication?

The question of how long is an important one: physical degradation of a file renders the question of its findability moot if a researcher finds a data file to use ten years from now, but it is too corrupted or obsolete a format to open and use. And is ten years a long time? In economics research, it is not. Looking at the length of time in which economic research is considered current, 10 years is the shortest amount of time the data should be usable. Yet, for data that is submitted to journals who do not have preservation policies, there are no guarantees the data will be viable in a decade or more. Most journals may encourage data availability, but their policies do not outline the technology and infrastructure that would ensure the data's long-term viability.

If replicability is the gold standard of economic research, then the code and data which informs that research must be, first, available but also viable and usable over the long-term, decades from now when the economic research is still read and consulted. The unavailability of Engle's and Kim's data via the *JMCB* journal archive was not inevitable and, we argue, could have been avoided with the extension of measures we see in place at some journals already. Not only should journals more explicitly require code and data sharing, but they should also stress the importance of long-term availability of data and provide more specific guidance to economists.

5.2 Suggestions for Further Research

One area of further research would be to understand what economists need to fulfill the requirements for providing their data with their research. We have tried to show here that even where a journal has a data policy, the actual obligation for authors is often ambiguous. Economists who are willing and actively seeking to share their data are not supported by clear guidelines. Many journals provide detailed and explicit instructions for topics like the fonts and headings that authors should use, but it was not uncommon in our survey to find no more than a sentence or two directing authors to clearly document their code and data without any instruction as to how to accomplish that. For authors using sensitive or restricted data, the barriers become that much higher. Without clear expectations as to what it means to successfully make data available in the long-term, it should be no surprise when an author requests an exception than try to wade through the complexities unaided.

Another area of future research would be to examine what journals are doing with the data once it is submitted. What are the journals doing to verify the data's viability? Are they checking that the code and data files are still accessible years after the economists submitted them? To what extent have files in journal archives been lost due to lack of preservation

measures? Referral to a repository with a long-term preservation policy – a refrain commonly found in journal policies – at least points authors to a structured space that is dedicated to making data available and provides guidance for doing so. Yet, as we discussed, reference to a repository or requiring a DOI is not sufficient to constitute a preservation policy. The AEA, for example, has made considerable headway in this area since the time of our sample collection in 2019 and a few other journals we reviewed have policies with features consistent with preservation, but still nothing is declarative, clear, and instructive.

Finally, on the subject of planning for the future, it is worth asking if there are proto-preservation measures to put in place now, knowing we cannot safeguard every code or data file that is released with economics research. Is there documentation about the data which should accompany the data itself? Would a detailed description of the data be sufficient to help the researcher in 2040 to replicate the original findings? A text file with a detailed description has a greater chance of remaining stable and intact. What would that description look like in order to be genuinely valuable to future researchers? Guidance and standards for these have been well established by information science and data professionals. How can journals leverage this work to promote and elevate these practices among authors in the more immediate term?

6. Conclusion

In this paper, we conducted content analysis of 184 top economics journals' code and data availability policies. We looked particularly closely at the guidance which journals provide regarding long-term preservation of code and data files. In our August 2019 snapshot, we found that a majority of the 184 journals provided a statement related to sharing data, while almost none had a formal preservation policy. We found that about half of the journals have policies which include recommendations about a repository or a persistent identifier, two things which

may help ensure long-term findability. However, a negligible number of journals' policies recommend anything concrete for the researcher to do to maintain their physical integrity and technical accessibility. A researcher may release code and data, actively encourage replication, and welcome another economist's findings, but it is not enough that the code and data are shared. Without active preservation of code and data files, there is no guarantee that future generations of researchers can access or use them. As the drivers of publication standards and practices, there is significant room for journals to take a more active role in promoting the safeguarding of these files.

References

Albright, Jeremy J., and Jared A. Lyle. 2010. "Data Preservation through Data Archives." *PS: Political Science and Politics* 43 (1): 17-21. http://www.jstor.org/stable/25699287.

Alliance, Digital S, Carol Kussmann, Matt Schultz, Lauren Work, Nathan T Tallman, Paige Walker, and Kathryn Michaelis. 2021. "National Digital Stewardship Alliance (NDSA)." OSF. November 4. osf.io/4d567.

Anderson, Gary M., David M. Levy, and Robert D. Tollison. 1989. "The Half-Life of Dead Economists." *The Canadian Journal of Economics / Revue Canadienne D'Economique* 22 (1): 174-83. https://doi.org/10.2307/135467.

Anderson, Richard G., William H. Greene, B. D. McCullough, and H. D. Vinod. 2008. "The Role of Data/Code Archives in the Future of Economic Research." *Journal of Economic Methodology* 15 (1): 99-119. https://doi.org/10.1080/13501780801915574.

Andreoli-Versbach, Patrick, and Frank Mueller-Langer. 2014. "Open Access to Data: An Ideal Professed But Not Practised." *Research Policy* 43 (9): 1621-1633. https://doi.org/10.1016/j.respol.2014.04.008.

Association of Research Libraries. 2005. "Urgent Action Needed to Preserve Scholarly Electronic Journals." https://www.arl.org/resources/urgent-action-needed-to-preserve-scholarly-electronic-journals.

Baker, Mary, Mehul Shah, David SH Rosenthal, Mema Roussopoulos, Petros Maniatis, Thomas J. Giuli, and Prashanth Bungale. "A Fresh Look at the Reliability of Long-Term Digital Storage." In *Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems* 2006, 221-234. https://doi.org/10.1145/1217935.1217957.

Bernanke, Ben S. 2004. “Editorial Statement.” *The American Economic Review 94* (1): 404. https://www.jstor.org/stable/3592790.

Butler, Courtney R., and Brett D. Currier. “You Can’t Replicate What You Can’t Find: Data Preservation Policies in Economic Journals.” Presentation, 2017 Annual International Association for Social Science Information Services & Technology (IASSIST) Conference, Lawrence, KS, May 2017. http://doi.org/10.17605/OSF.IO/HF3DS.

Chang, Andrew C., and Phillip Li. 2015. “Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'.” Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System. http://dx.doi.org/10.17016/FEDS.2015.083.

Council of the Consultative Committee for Space Data Systems. 2012. “Reference Model For an Open Archival Information System (OAIS).” https://public.ccsds.org/pubs/650x0m2.pdf.

Crosas, Mercè, Julian Gautier, Sebastian Karcher, Dessi Kirilova, Gerard Otalora, and Abigail Schwartz. 2018. “Data Policies of Highly-Ranked Social Science Journals." *SocArXiv*. March 30. https://doi.org/10.17605/osf.io/9h7ay.

Dewald, William G., Jerry Thursby, and Richard G. Anderson. 1986. "Replication and Scientific Standards in Empirical Economics: Evidence from the JMCB Project." *American Economic Review* 76, no. 4 (September): 587-603.

Duvendack, Maren, Richard W. Palmer-Jones, and W. Robert Reed. 2015. “Replications in Economics: A Progress Report.” *Econ Journal Watch 12* (2): 164–191.

Faber, Ronald. 2002. "From the Editor: A Glance Backward and the View Ahead." *Journal of Advertising* 31, no. 4 (Winter): V-VII. http://www.jstor.org/stable/4189232.

Frisch, Ragnar. 1933. "Editor's Note." *Econometrica* 1 (1): 1-4. http://www.jstor.org/stable/1912224.

Hunt, H. Keith. 1979. "From the Editor." *Journal of Advertising* 8 (2): 3-4. https://doi.org/10.1080/00913367.1979.10717968.

Huschka, Denis. 2013. "Why Should We Share Our Data, How Can It Be Organized, and What Are the Challenges Ahead?" RatSWD Working Paper Series, 216. Berlin: Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://nbnresolving.org/urn:nbn:de:0168-ssoar-427918.

Journal of Money, Credit and Banking. n.d. "Data Archiving Policy." Journal Index and Archive. Accessed August 24, 2021. https://jmcb.osu.edu/archive.

Kim, Katherine, Digital S Alliance, Bethany Nowviskie, Wayne Graham, Becca Quon, Carol Kussmann, Winston Atkins, and Aliya Reich. 2020. "2013 Levels of Digital Preservation." OSF. November 16. osf.io/9ya8c.

Library of Congress. n.d. "National Digital Information Infrastructure and Preservation Program." Accessed August 24, 2021. https://www.loc.gov/loc/lcib/0601/ndiipp2.html.

McCullough, B. D., Kerry Anne McGeary, and Teresa D. Harrison. 2006. "Lessons from the JMCB Archive." *Journal of Money, Credit and Banking* 38 (4): 1093-107. http://www.jstor.org/stable/3838995.

McCullough, B. D., Kerry Anne McGeary, and Teresa D. Harrison. 2008. "Do Economics Journal Archives Promote Replicable Research?" *The Canadian Journal of Economics* 41(4): 1406-1420. http://www.jstor.org/stable/25478330.

Moffitt, Robert. A. 2011. "Report of the Editor: American Economic Review (with Appendix by Philip J. Glandon). *American Economic Review* 101 (3): 684–693. https://doi.org/10.1257/aer.101.3.684.

O'Connor, Cliodhna, and Helene Joffe. 2020. "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines." *International Journal of Qualitative Methods* 19: 1-13. https://doi.org/10.1177/1609406919899220.

Sedransk, Nell, Linda J. Young, Katrina L. Kelner, Robert A. Moffitt, Ani Thakar, Jordan Raddick, Edward J. Ungvarsky, et al. 2010. "Make Research Data Public? - Not Always so Simple: A Dialogue for Statisticians and Science Editors." *Statistical Science* 25 (1): 41-50. https://doi.org/10.1214/10-STS320.

Shah, Ubaid and Sumeer Gul. 2019. LOCKSS, CLOCKSS & PORTICO: A Look Into Digital Preservation Policies. *Library Philosophy and Practice* 2841. https://digitalcommons.unl.edu/libphilprac/2481.

Sturges, Paul, Marianne Bamkin, Jane H.S. Anders, Bill Hubbard, Azhar Hussain, and Melanie Heeley. 2015. "Research Data Sharing." *Journal of the Association for Information Science and Technology* 66: 2445-2455. https://doi.org/10.1002/asi.23336.

Vilhuber, Lars. 2019. "Report by the AEA Data Editor." *AEA Papers and Proceedings* 109: 718–729. https://doi.org/10.1257/pandp.109.718.

Vlaeminck, Sven, and Lisa-Kristin Herrmann. 2015. "Data Policies and Data Archives: A New Paradigm for Academic Publishing in Economic Sciences?" In *Proceedings of the 19th International Conference on Electronic Publishing*, edited by Birgit Schmidt, Milena Dobreva, 145–155. Amsterdam: IOS Press. https://doi.org/10.3233/978-1-61499-562-3-145.

Willis, Craig, and Victoria Stodden. 2020. "Trust but Verify: How to Leverage Policies, Workflows, and Infrastructure to Ensure Computational Reproducibility in Publication." *Harvard Data Science Review* 2 (4). https://doi.org/10.1162/99608f92.25982dcf.

Yoon, Ayoung and Teresa Schultz. 2017. “Research Data Management Services in Academic Libraries in the US: A Content Analysis of Libraries’ Websites.” *College & Research Libraries* 78 (7): 920-933. https://doi.org/10.5860/crl.78.7.920.

Zittrain, Jonathan, Kendra Albert, and Lawrence Lessig. 2014. “Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations.” *Legal Information Management* 14 (02): 88-99. https://doi.org/10.1017/S1472669614000255.