Data policies of highly-ranked social science journals[1]

Mercè Crosas[2], Julian Gautier[2], Sebastian Karcher[3], Dessi Kirilova[3], Gerard Otalora[2], Abigail Schwartz[2]

Abstract

By encouraging and requiring that authors share their data in order to publish articles, scholarly journals have become an important actor in the movement to improve the openness of data and the reproducibility of research. But how many social science journals encourage or mandate that authors share the data supporting their research findings? How does the share of journal data policies vary by discipline? What influences these journals' decisions to adopt such policies and instructions? And what do those policies and instructions look like?

We discuss the results of our analysis of the instructions and policies of 291 highly-ranked journals publishing social science research, where we studied the contents of journal data policies and instructions across 14 variables, such as when and how authors are asked to share their data, and what role journal ranking and age play in the existence and quality of data policies and instructions. We also compare our results to the results of other studies that have analyzed the policies of social science journals, although differences in the journals chosen and how each study defines what constitutes a data policy limit this comparison.

We conclude that a little more than half of the journals in our study have data policies. A greater share of the economics journals have data policies and mandate sharing, followed by political science/international relations and psychology journals.

Finally, we use our findings to make several recommendations: Policies should include the terms "data," "dataset" or more specific terms that make it clear what to make available; policies should include the benefits of data sharing; journals, publishers, and associations need to collaborate more to clarify data policies; and policies should explicitly ask for qualitative data.

---

[1] Authors contributed equally to the analysis and writing of this paper and are listed alphabetically. Replication data and code (Crosas et al., 2018) are available on the Harvard Dataverse at https://doi.org/10.7910/DVN/CZYY1N

[2] Institute for Quantitative Social Science, Harvard University

[3] Qualitative Data Repository, Syracuse University

Introduction

As scientific research has become more data-intensive due to the application of information technologies to scientific problems (Kowalczyk & Shankar, 2011) and the amount of data collected, stored, and analyzed has increased (Tenopir et al., 2011), academics from a wide array of disciplines (Fienberg, Martin, & Straf, 1985; King, 1995; Dalrymple, 2003; Esanu & Uhlir, 2003; Freese, 2007; Piwowar & Chapman, 2008; Hanson, Sugden, & Alberts, 2011; Borgman, 2012) have argued that data are a critical part of the scholarly output, and therefore scholars have a responsibility to preserve and share the data used in their research. Data sharing enables transparency, benefits other researchers allowing them to re-use data to answer new research questions or to replicate previous studies, facilitates methods training and collaboration across disciplines, and overall advances knowledge discovery and innovation. These intertwined interests in data sharing and research transparency have become prominent within the academic community, funding agencies, professional associations, peer-reviewed journals, and publishers, as well as among governments and international organizations.

As a result, a pioneering "data availability movement" (Gherghina & Katsanidou, 2013) has taken shape across these different loci across the research enterprise. While different actors have a role to play in promoting the data sharing imperative, the research community has recognized that meeting journals' expectations for data sharing by researchers provides an essential incentive due to the significance that getting one's work published holds as the standard for scientific progress, academic recognition, and career development. Therefore, many journals have created data sharing policies that aim to archive the data underlying scientific papers (Vines et al., 2013) and implement access and preservation for the rest of the scientific community. According to Sturges et al. (2015), the pivotal aspect of journal-based data sharing policies is their pragmatism, as it recognizes that presenting these policies in the "instructions for authors" at a time where authors are submitting their manuscript creates both an "incentive for compliance and the opportunity to do so." (p. 2446)

Therefore, the study of journal data policies is central to understanding the latest changes (or lack thereof) in a data sharing culture. The literature suggests that in general, journals have agreed to enforce some kind of data policy, and, overall, these policies have become more rigorous over time (Borgman, 2012). Nowadays, apart from access to data, some journals require authors to

share the computer code involved in the data analysis (Borgman, 2012), a reference list with data citations, a data statement regarding the availability of data (Vines et al., 2013) and the location where they have deposited the data or use of a community-endorsed venue (Freese, 2007). Although the basis of all data policies is the need for data access and preservation, specifics of policies vary dramatically (Parsons, 2013). According to Sturges et al. (2014), the best way for journals to address data sharing issues is through strong and clear data policies that mandate authors to submit their data during manuscript submission. Recently, there have been a number of efforts from professional associations to push journals to implement strong data policies. A good example of these efforts is the Data Access and Research Transparency (DA-RT) statement that was incorporated into the 2013 Code of Ethics of the American Political Science Association (APSA). In 2014 a group of editors of leading political science journals committed to enacting data policies compliant with DA-RT by January 2016 (see https://www.dartstatement.org/ for more on both DA-RT and the Journal Editors' Transparency Statement).

Publishers such as Taylor & Francis and Springer; professional associations such as APSA and the American Psychological Association (APA); and community-driven efforts such as the Transparency and Openness Promotion Guidelines (TOP Guidelines) have had a key role in the effort to standardize journal data policies. Nevertheless, journal data sharing policies present many differences among disciplines, and thus, researchers who have studied journal data sharing policies in the social sciences have mostly focused on their own discipline instead of conducting multi-disciplinary studies (e.g. Wicherts, Borsboom, Kats, & Molenaar, 2006; Gherghina & Katsanidou, 2013; and Zenk-Möltgen & Lepthien, 2014).

The main goal of our study is to provide an overview from across the major social science disciplines of the presence and quality of data sharing policies by journals. In order to do so, we review the data policies of the 50 most influential international peer-reviewed journals according to the Claryvate (formerly Thomson Reuters) Journal Impact Factor in the disciplines of political science and international relations, economics, sociology, history, psychology, and anthropology. While history is typically classified as a humanistic discipline, its clear reliance on empirical evidence (i.e. data) makes it an appropriate inclusion in the list of domains we review.

We classify the data sharing policies in each discipline on a number of dimensions meant to represent policy standards / best practices from the point of view of the data management

community, and engage in a cross-disciplinary comparison. Where possible, we compare the current data sharing policies with those presented in past studies to explore how the policies have evolved in the last years. Finally, based on a follow-up interaction with the editors of journals where no data policy was publicly available, we try to understand some of the reasons that might prevent or delay the adoption of a data policy. This is the first cross-disciplinary study focused exclusively on the social sciences. The ability to compare across disciplines allows us to highlight differences in disciplinary priorities and trajectories both in the presence of data policies and in their content. Moreover, we hope to inform and prompt journals that lack a data sharing policy to invest in developing one and all journals to refine theirs to provide the right amount of detailed requirements for authors with the expectation that such clear instructions on data provision will in turn increase policy compliance and actual data sharing.

Methods

Research Questions

We aimed to find out, at a most basic level, whether significant inter-disciplinary differences exist with respect to the presence or absence of journal-based data policies in the fifty highest-ranked peer-reviewed journals for each of the above listed fields. Secondarily, we wanted to discover whether some disciplines have more rigorous data policies, both with respect to more specific expectations of what authors need to do to provide access to their research data (e.g., does the policy describe where the data should be archived or how the author should cite both his or her own and others' data) and with respect to the repercussions if they did not comply (e.g., no article will be published without either the underlying data or an author's statement explaining why sharing cannot happen being provided). Lastly, we explored correlations among the existence and strictness of data policies and characteristics of those journals, such as ranking and age.

Case selection

We reviewed the public websites of the highest-ranked journals according to the Journal Citation Index. In this we mirror the approach used for other single-discipline studies of journal data sharing policies in the social sciences and beyond (Gherghina & Katsanidou, 2013; Gleditsch & Metelits, 2003; Piwowar & Chapman, 2008; Stodden, Guo, & Ma, 2013; Vines et al., 2013;

Zenk-Möltgen & Lepthien, 2014). While impact factors and rankings of this sort have well-known problems, they are very suitable for the purposes of establishing a set of highly influential journals in a given field. We looked at the full text of the submission instructions, searching for a number of phrases, which we treated as interchangeable: data availability policies, data archiving policies, replication policies, and simply data policies. In cases where we could not identify a publicly available policy, we followed up with an email to the journal's editor asking whether such a policy existed either explicitly but "behind the scenes" and was simply not posted online, or informally.

Variables and Coding Procedures

Which parts of data policies we chose to assess (for example, when data submission is expected, what location for the data publication is recommended, etc.) was driven by our understanding as data management professionals of how researchers, journals, and professional data stewards can work together to share data effectively and best enhance the research transparency of their related scholarly articles. Similar studies also use many of these variables.

We defined our key variable of interest in the broadest possible way: a journal was considered to have a data / empirical materials policy if data were mentioned explicitly in any portion of the instructions; we did not restrict that value to only journals that had a dedicated section with such a label. We compared the absence versus presence of policies across the disciplines and found some clear differences, as discussed further in the Analysis section.

Where publications included instructions for the possible submission of supplements without explicitly referring to their contents as data, we did not code these as data policies, although one possible interpretation of such instructions is that they might envision *data* supplements. However, the very fact that such terminology leaves so much open to interpretation suggested to us that such broad instructions do not meaningfully function to encourage data sharing as instructions that talk about data *per se* do.

For the journals where this dichotomous variable received a "yes," we further coded ten variables describing specific details of the policy (for example, whether there were instructions where data are supposed to be archived; what the expected timing of data submission was;

whether qualitative data were addressed explicitly[4]), as well as the source of the policy and a general sense of the strictness of the data sharing expectation. Lastly, we were interested in learning what nomenclature journals used for data and made a list of the phrasing used in the policies.

While strictness is one of the aspects we were most interested in and we considered creating an index that captured how much detail the policy went into on the ten characteristics, in the end we decided to start with a simpler approach: code anything expressing an obligation to share data as "required." To avoid subjective judgment, we based coding on the exact verbiage used by the journal: "have to," "are expected to," "must" were all treated as expressing obligation, even if there was no obvious information about checks / verification steps or repercussions. Other verbs (the most common were "suggested" and "encouraged," in one case "encouraged to consider") or the mere mention of the possibility of sharing data were treated as "encourage" only.

We also distinguished the origins of data policies. In addition to journals' own editorial boards writing them, in many cases a disciplinary association or a publisher provided either a template that journals adapted or a wholesale text that editors used without alteration.

Recoding and Inter-coder Reliability Checks

After reviewing our first round of coding, we improved our coding instructions and variable definitions and added a variable to mark whether the policy recommended or mandated data sharing. Then every collected journal policy was re-visited in a second round of coding. We reconciled our first and second rounds of coding by finding the coding differences between them and re-visiting the journal policies we collected to determine which coding was correct. This reconciliation step was conducted by two of the authors (Gautier and Karcher).

In a final step, we conducted various plausibility checks and re-visited every inconsistent coding: examples of such coding included the same journal in different disciplines receiving different codings or journals being coded as having no policy with non-zero codes for the *attributes* of a data policy (such as location of data or source of the presumably non-existent policy).

[4] The full list of variables and the values assigned are available in the Appendix.

Follow-up with journal editors

When a selected journal did not have a publicly available data policy, as defined above, we emailed the editors with a brief list of questions trying to understand whether such a policy might be used informally or whether there was an intentional choice to not have expectations that empirical data used in articles would be shared. Representatives of 125 journals were emailed in September and October 2017 and we received some information about 47 of those journals (37.5% response rate), although not all of them addressed all the questions posed. The most common response, from 30 journals, was that the journal indeed had no policy and did not intend to introduce one (more on some reasons given below). Eight wrote that they are currently developing a policy, often under the recommendation of the journal's publisher. Four wrote that they do provide some uniform guidance on the possibility of data sharing in their communications with authors, even though they do not have publicly posted policies. Two representatives corrected our initial coding of their journals as not having data policies (these responses have been incorporated in the data reported here).

A few interesting trends emerged in the answers from these inquiries. Several respondents reported that their journal is currently in conversations with its publisher on this topic, and an additional one informed us that the relevant committee of their disciplinary association has it on its agenda for an upcoming meeting to discuss ways to encourage data sharing, including via journals.

Editors of journals that focus on research based on archival, interview-based, and other fieldwork frequently responded that since they virtually never publish quantitative articles, their journals do not need data sharing policies. Some explained that this was driven by a lack of easy answers to the concerns of confidentiality, foreign language competency, or form of the collected empirical materials. We believe this is an incomplete understanding of what empirical data are and how they can be made available to other interested scholars in a legal and ethical manner. (In the questions we emailed the words "quantitative" or "qualitative" were not mentioned. We asked about data or empirical materials.) Other editors had almost the reverse explanation: that since most of the work they publish is not based on any "confidential" sources, they do not find data

sharing to be relevant for the publication either. However, direct access to the empirical materials cited in a given article is not guaranteed even where they are apparently public. For example, materials in historical archives may be prohibitively difficult to obtain, hyperlinks may become obsolete, and published documents (especially online) may not be preserved for long-time access. While it is understandable that many journal editors do not feel they have the resources to maintain their own repositories of supplemental materials of this sort, working with existing social science repositories might provide a win-win solution for them.

An interesting common undertone across several titles whose editors wrote that the topic of data sharing has been part of internal discussions suggests that because their respective disciplines do not have a culture of data sharing, creating such a policy is of potential interest, but is not a high priority. Editors recognize that the standard practices of the scholarly communities they serve are changing, but where any given journal falls on the spectrum of data sharing seems in many cases to depend on the specific editorial team in place and the sub-discipline the journal mainly serves. Several wrote to say either that they just started in the position and are reviewing policies and potential gaps in them, or that they are the outgoing editors and so are not sure what their successors might choose to do in this regard. (To our knowledge no journal that at any point adopted a data sharing policy has reversed it at a later point.)

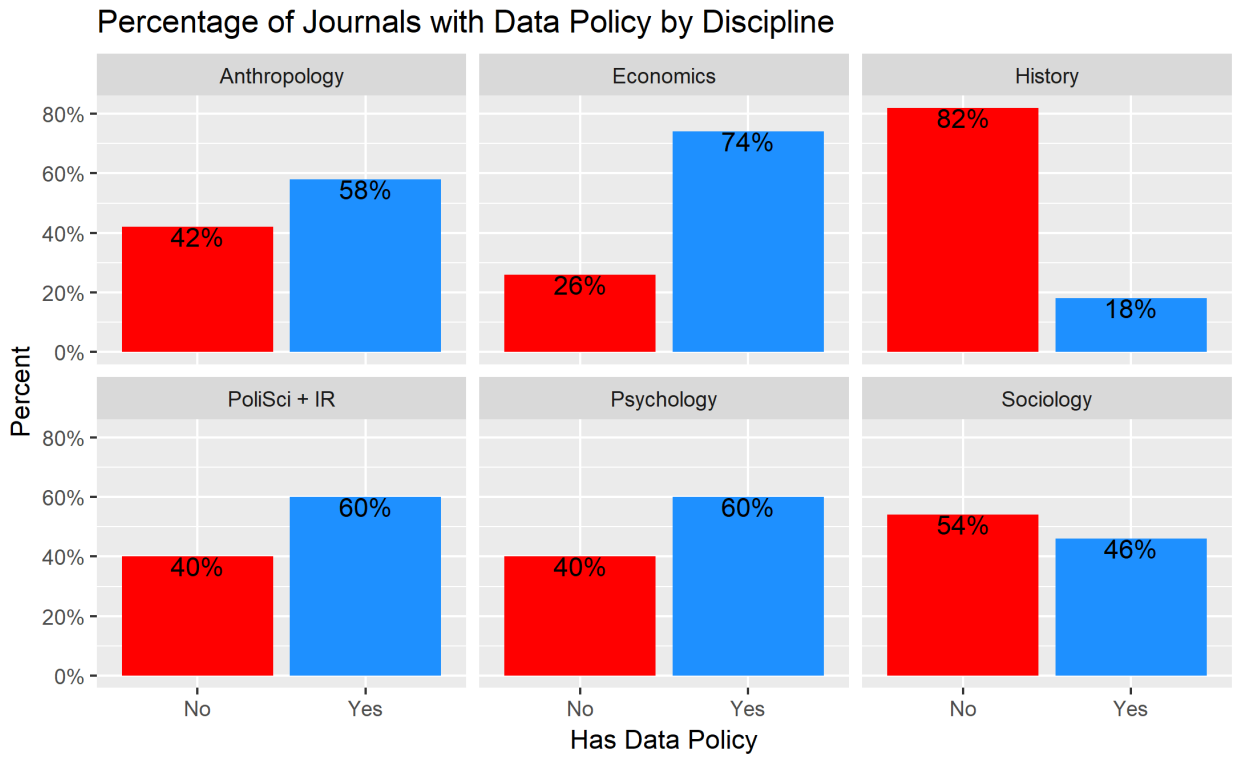
Analysis

Among the highest-ranked 50 journals in the six disciplines we studied, seven journals were in two disciplines; one journal (*Socio-Economic Review*) was in three. This leaves us 291 unique journals. In the analysis below, we count these journals as duplicates for all analyses by discipline, while using only the 291 unique journals for all other analyses. We found data policies for a little over half (155) of these journals.

Data policies by discipline

The share of journals with data policies varied strongly by discipline as shown in Figure 1: While three out of four economics journals have such a policy, only one in five history journals does. The other disciplines are between these two, with data policies in about 60% of anthropology, political science, and psychology journals and in 46% of sociology journals.

Figure 1: Data Policies by Discipline



Strictness of data policies

There is significant variation within these data policies. One crucial feature of data policies is whether they *require* the sharing of data or whether the sharing of data is merely expected, encouraged, or supported. In political science, Key (2016) finds among highly-ranked political science journals, that more than 80% of articles in journals that require data sharing include replication data, whereas less than 30% of articles in journals that have no stringent requirement include replication files.

Among the 291 journals, about one in six (53) have stringent requirements to publish data alongside articles. Again, data policies vary significantly by discipline as shown in figure 2. economics has the most stringent requirements, with more than one third of journals requiring data sharing. It is followed by political science and psychology, where 30 and 22% of journals have such requirements. They are largely absent from policies in anthropology, sociology (both 10%), and history (0).

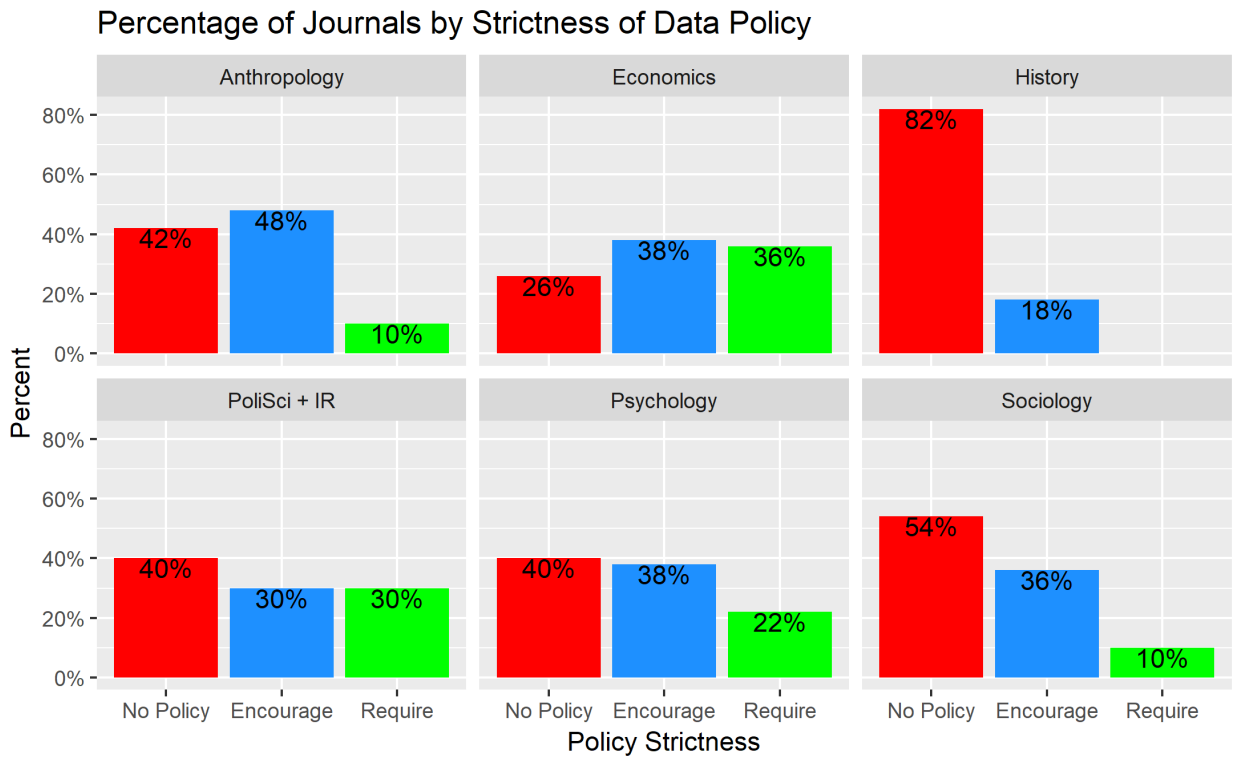


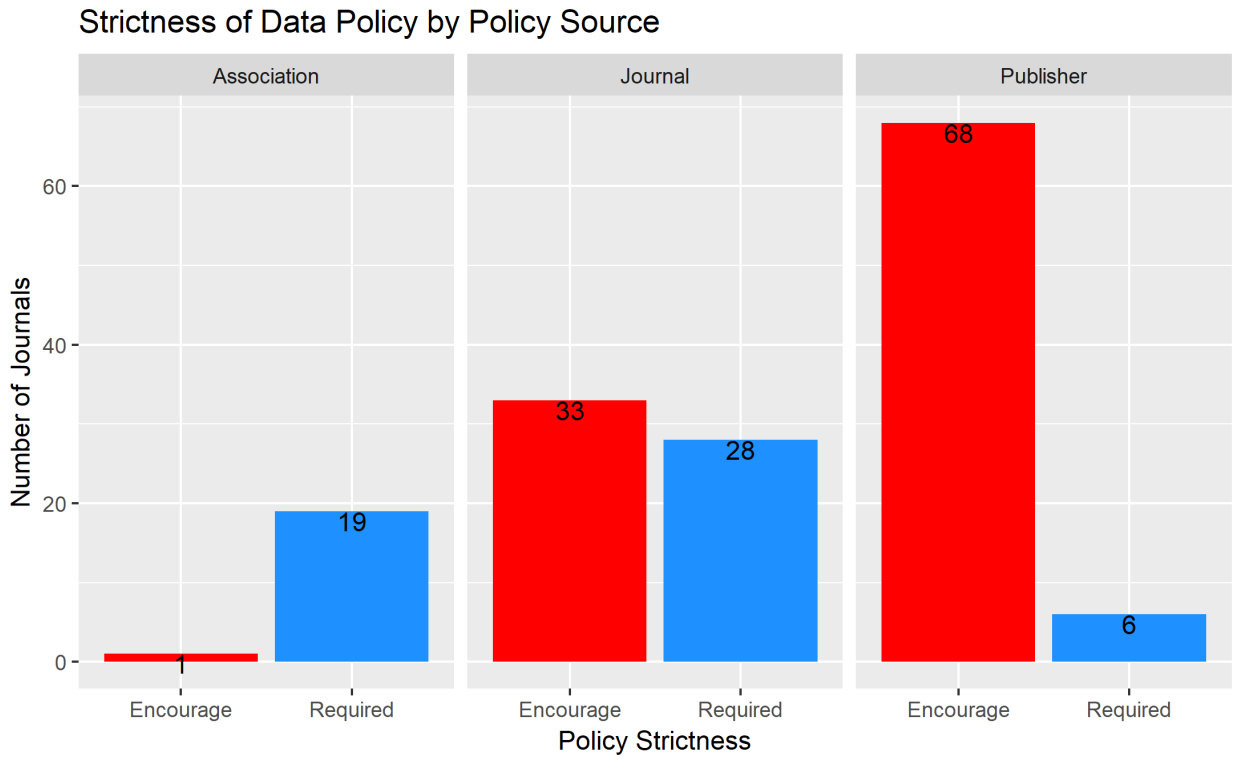
Figure 2: Data Policy Strictness by Discipline

Data policies by source

Data policies for journal articles have originated from three main sources: the journals (and their editorial boards) themselves, their publishers, and professional associations that may either publish journals or exert influence over journals in a field. Figure 3 suggests that the source of a data policy significantly influences its contents. Publishers are the most common source of data policies, and where publishers (such as Springer, Elsevier, and Taylor & Francis) have taken an active role in promoting data policies, we found that almost all of the journals they publish have data policies. But the figure also suggests a possible downside. Only a very small share (8%) of journals with publisher-generated policies require data sharing. Among journals' own policies, about half require data sharing. Strikingly, almost all policies by professional associations (such as the American Economic Association [AEA], the American Psychological Association [APA], the American Political Science Association [APSA], and the American Sociological Association [ASA]) require data sharing.[5]

[5] The figure likely understates the importance of associations for strict data policies. In political science, most journals implement their own version of the "Data Access and Research Transparency" (DA-RT) policy instituted by the APSA (see below).

Figure 3: Policy Strictness by Source



Data policy by journal ranking and age

Does the prominence of a journal affect whether it has a data policy or the strictness of its policy? We measure prominence by a journal's impact factor and by its rank within the respective discipline by impact factor. The two measures are quite similar, but rank discounts outliers at the top and increases the prominence of small differences among lower-ranked journals. We assess the question using a series of logit regressions, all controlling for journals' disciplines. We find a moderate and statistically significant association between a journal's prominence (measured either as impact factor or ranking) and the existence of a data policy. At the lowest impact factor in our study (0.25), we estimate a probability of 40% for a journal to have a data policy (18%-62% 95CI). For a journal with the highest impact factor in our sample (19.95), our estimate of this probability is 99% (96%-100% 95CI). Similarly, a journal ranked 50th in its discipline has an estimated probably of 48% to have a data policy (30%-65% 95CI) while a journal ranked first has a data policy with an estimated probability of 71% (57%-86%

95CI).[6] We find no association between impact factor and the strictness of policies among journals with data policies. We do, however, find a strong association between a journal's rank and the strictness of its data policies. Among journals with data policies, we estimate that the first-ranked journal has an 81% probability (66%-97% 95CI) to require data sharing, whereas the same probability is only 16% (2%-31% 95 CI) for the 50th ranked journal. This finding is in line with expectations: as strict journal policies impose costs on authors and may discourage submissions, higher-ranked journals will find it easier to impose such costs, as they are unlikely to see a loss of submissions as a consequence.

A similar question concerns the relationship between a journal's age (i.e., time since the publication of the first issue). We do not find any relationship between journals' age and the existence of a data policy, but find a fairly strong association between age and strictness of data policy. Among those journals with a data policy, the probability of the oldest journal, at 131 years old, requiring data sharing is estimated at 94% (85%-100% 95CI), with 25% estimated for the youngest journal, at 4 years old (8%-41% 95CI). We suspect that a journal's age serves as a proxy for reputation and standing in its disciplines, which allows for the imposition on higher hurdles for submitting authors.

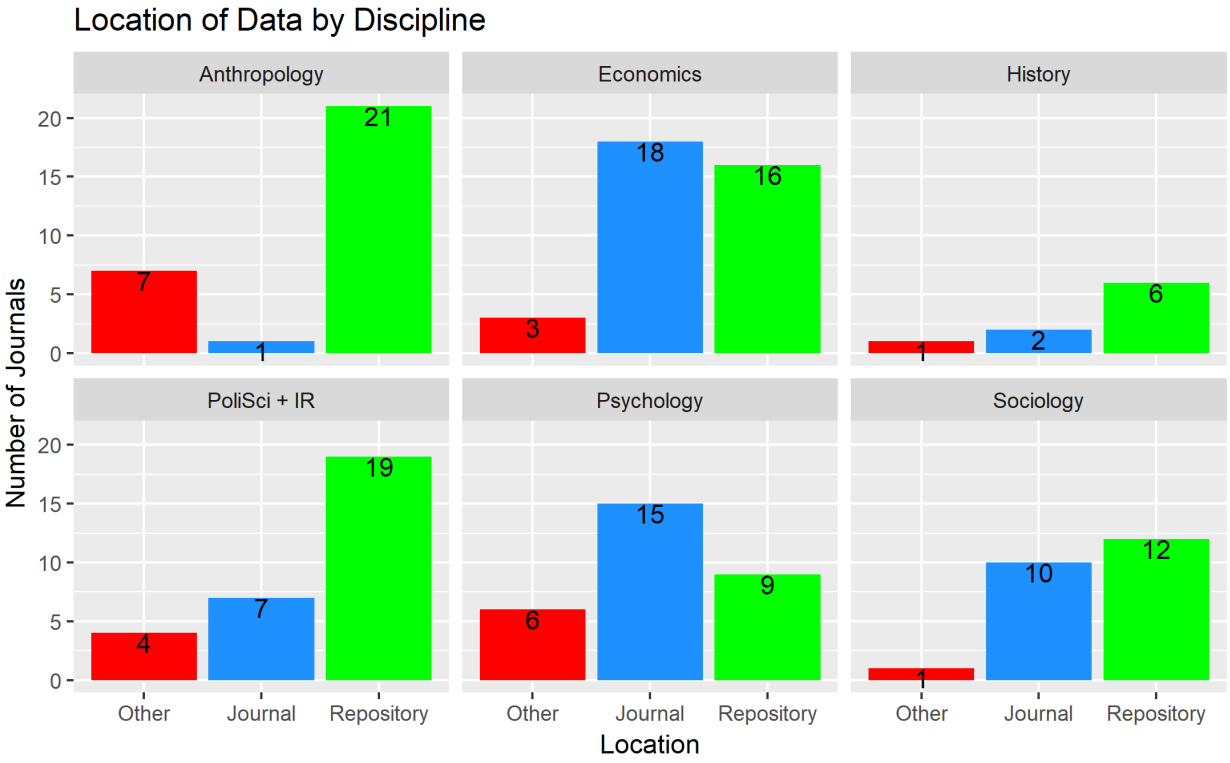
Where do journals tell authors to share data?

Among data professionals there is a broad consensus that data should be shared in a data repository to ensure that they remain FAIR (i.e., Findable, Accessible, Interoperable, and Re-usable; see Wilkinson et al. 2016). To what extent do journal data policies follow these recommendations and what contributes to differences among journals? Among the 155 journals with data policies, 80 recommend or require a data repository for shared data, another 53 promote data sharing through the journal's own site as supplementary materials, and 22 journals either do not specify a location or specify another mode of sharing such as "by request." Given the results in Stodden, Seiler, & Ma (2018), a recently published study that underlines the deficiency of data sharing by request, these numbers are encouraging. There is significant variance on the recommended location for shared data based on both the discipline of the journal and the source of the data policy. As Figure 4 shows, economics and psychology journals see a strong role for journal supplementary materials, whereas the other disciplines, in particular

[6] Details on these regressions and predicted effect graphs are available in the replication materials.

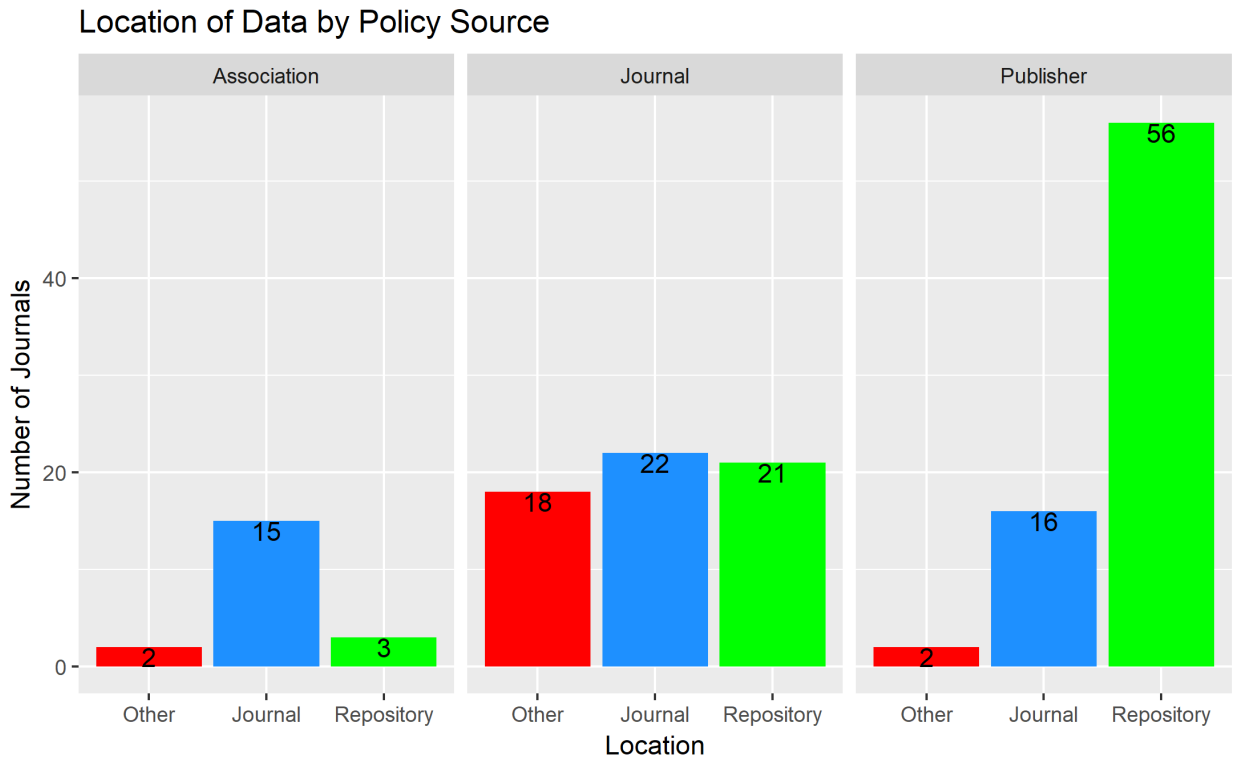
anthropology and political science, show a preference for data repositories. In political science, we believe this is due to the long-standing importance of a disciplinary data repository, ICPSR, as well as the more recent development of Dataverse as a widely known and used repository among political scientists due to its association with Harvard's Institute for Quantitative Social Science (IQSS). In anthropology, the role of repositories is likely due to the strength of major publishers (see below) and journals in biological anthropology, which share stronger norms of data sharing with the biosciences.

Figure 4: Location for sharing data by discipline



The source of data policies is also associated with stark differences in the recommended location for sharing data (see Figure 5). Publishers strongly favor data repositories, while associations appear to favor journals' supplementary materials as a location for shared data (with the results being driven almost entirely by journals in Psychology and Economics). Journals' own policies vary widely, with the highest share of "other" (typically poorly specified and/or ill considered). This is a curious mirror image of the effect of a policy's source on policy strictness above. Taken together, these results suggests that collaboration between different stakeholders in the publishing process may help to produce optimal data policies.

Figure 5: Location for sharing data by policy source



Effect of TOP and DA-RT on data policies

There are several initiatives to promote more open research practices in general and data sharing in particular. Probably most widely known are the "Transparency and Openness Promotion" (TOP) guidelines, published by the Center for Open Science (https://cos.io/our-services/top-guidelines/) and based on Nosek et al. (2015), signed by hundreds of journals across the scientific spectrum. A smaller but similar intiative is the "Journal Editors' Transparency Statement" (JETS, https://www.dartstatement.org/2014-journal-editors-statement-jets), in which a number of prominent journals in political science pledged to implement the APSA's recent commitment to data sharing by requiring authors to share data as a condition for publication. We do find that signatories of both policies are much more likely to have data policies and, in the case of JETS, these data policies are stricter. Of the 25 journals that have signed TOP, we found data policies for 22 (88%), compared to 50% of non-TOP journals. Similarly, we found data policies for 14 out of 15 JETS signatories (93%), compared to 46% of political science journals that had not signed. Of those 14, 11 had a strict data policy, requiring data sharing, whereas only four out of the 16 other political science journals with data policy did.

Data policies and qualitative data

Debates about data sharing have more recently turned to qualitative data such as images, interview transcripts, and archival documents (see e.g. Elman & Kapiszewski, 2014). Only a small minority of journals in our study (11) explicitly mentioned qualitative data, more than half of those (6) in political science, where debates about qualitative data sharing have been most visible. Many journals use language that implicitly acknowledges the possibility of qualitative data (e.g. by mentioning videos or transcripts). In economics, in line with the quantitative orientation of the discipline, we find such an implicit mention absent from many journals – a justified omission where journal editors do not expect any qualitative data used in articles.

Discussion

Comparing our findings with previous studies

We look to previous studies to gauge the progress that the academic publishing community has made in implementing and strengthening data availability policies. We have reviewed the data of four studies in which researchers also assessed the policies of journals categorized as social science journals. We summarize these studies and ours in Table 1.

Table 1: Summary of social science journal data policy studies

Study	Journal selection criteria	# of journals	# of data policies found
Gleditsch, N., & Metelits, C. (2003)	15 most frequently cited journals in political science, 13 most frequently cited journals in international studies in the 1999 Journal Citation Reports. (One journal is in both lists)	27	8
Gherghina, S., & Katsanidou, A. (2013)	170 political science journals in the 2010 Thomson Reuters' Social Science Citation Index, minus journals with no functional website and journals that did not publish articles using data	120	18
Zenk-Möltgen, W., & Lepthien, G. (2014)	The 140 sociology journals listed in the Thomson Reuters' Social Science Citation Index	140	101
Sturges et al.	Deduplicated list of 100 most-cited, 100 least-	193	63

(2015)	cited social science journals in Thomson Reuters' 2011 Journal Citation Report		
Crosas et al. (2018) (this paper)	50 highest-ranked journals in six social science disciplines according to the Claryvate Journal Impact Factor	291	155

Each study uses different variables to compare policies, such as whether the policy encourages or requires data sharing, the journal's founding date, and its impact factor. But all of the studies determine if each journal has a publicly available data policy (e.g., on its website). We used this common variable to attempt to measure how the number of social science journals with data policies has changed, keeping in mind that:

- Because of differences in scope and sampling methodology, there is limited overlap between the journals in our study and those in the other four studies. There are 642 unique journal titles in all five studies, including ours. Only 102 of the 291 journals in our study are also in one or more of the other four studies.
- We suspect that our study's definition of a data policy is narrower or broader than the other four studies' definitions. For example, we ignored policies that only mention supplemental material or information, whereas Paul Sturges et al. "found it impossible to totally ignore what are described as supplemental materials which might be deposited along with data." (2013, p. 2446) We also did not code a journal as having a policy if a professional organization it belonged to did have a policy for sharing data, but that policy is not mentioned in the journal's submission guidelines. For example, among the journals that Zenk-Möltgen & Lepthien (2014) coded as having data policies are 21 journals that we report do not have data policies. We think it is highly unlikely that so many journals have abandoned their data policies. Instead, we think that the authors recorded journals as having data policies because they were using a broader definition of a data policy or because they always took into account journals' relationships with professional organizations that had data policies, while we ignored these relationships unless the journals' policies explicitly stated that authors are bound by such policies.

To understand changes across time, we are therefore ignoring cases in which studies report that a journal had a data policy when we found that they do not. Of the 102 journals in our study that

are also in one or more of the four previous studies, one or more previous studies recorded 34 journals as not having a data policy. We found that 16 of those 34 journals still have no data policies, while 18 journals have since adopted data policies. 16 of those 18 journals were included in studies published in 2013 and 2014, which suggests that most of these changes happened within the last four to five years.

Limitations

This study has several limitations.

First, unlike many similar studies (e.g. Zenk-Möltgen & Lepthien, 2014, discussed above), we did not attempt to measure how well authors abide by the journal policies we reviewed. We focused on the clarity and thoroughness of data policies, some of which we and the other studies we include in our discussion have found difficult to interpret. However, we think it would be beneficial to investigate how easy or difficult it is for authors to figure out whether journals have data policies and what those policies' requirements are.

Second, we used a definition of a data policy that is narrower than other studies' definitions (see Appendix for other studies' definitions). As noted in our discussion, while our study's coders were asked to code journals as having no data policy if their policies mentioned only "supplemental information" (or something along those lines), we suspect that other studies considered such policies to be data policies. We also required any policy to be on or directly linked to a journal's author guidelines. Being published by an organization or publisher that supports data sharing does not, in our view, mean that functionally a journal has a data policy if authors are unlikely to find or follow the broader entity's suggestions.

Third, our study and the other studies we have reviewed rely on lists of journals in Journal Citation Report, which have been criticised for favoring "journals published in English and circulated in the main academic centres." (Altbach, 2005, p. 151) We do agree that ours is *not* a representative sample. Since, as Zenk-Möltgen & Lepthien note, "the higher ranked journals tend to be trend setters for the rest of the journals," (2014, p. 719) we believe our sample is nonetheless highly instructive for understanding data policies across social science's most influential journals.

Recommendations

Include the terms "data and "dataset" in policies:

Data availability policies should phrase their instructions to more clearly use "data" terminology. In some disciplines, the terminology may include even more specific terms, such as configuration and script files for simulation data in economics research. But we think that the common term being used to refer to research outputs that cannot be printed with the article, which appears to be "supplemental material," is so broad that any data sharing recommendation or mandate using such terms will most often be interpreted as very weak and easily ignored.

Include the benefits of data sharing:

Only about a third of the journal data policies we found (46 of 155 journals) describe the benefits of sharing data, most of which (40) are from the journal publisher. A statement of *why* data should be shared does not just explain to prospective authors the additional burden imposed by sharing data, it also helps to clarify for authors what they need to share and how best to share it.

More collaboration among journals, publishers, and associations will improve data policies:

Our analysis has shown that publishers, associations, and journals have different comparative advantages in formulating and propagating data policies and should work to align their individual messaging on this topic for maximum effect. Publishers have the widest reach and can easily set "baseline" policies encouraging data sharing among all of their journals that publish empirical work. They are also often most attuned to discussions in information science and among data professionals and thus will be most likely to follow best practices such as recommending data to be shared through repositories. Professional association can set standards for good academic practice within their disciplines and can do so with significant legitimacy, as they are membership-run. They are thus best suited to strongly influence norms for data sharing, but may be ill-prepared to formulate specifics or to reach far beyond their society journals. Journals, finally, are in direct contact to authors and thus in the best position to formulate policy specifics. They can judge the types of mandates appropriate for their audience, provide adequate instructions, and consider if and how any requirement can be enforced. Taken together, we believe each stakeholder should play an active role in promoting data policies, and should do so in communication with other stakeholders to design optimal data policies for each journal.

Don't interpret "data" as exclusively quantitative:

Only 11 of our study's 155 journals with data policies explicitly address what authors should or must do with their qualitative data, six of which publish political science research. The remaining 144 data policies either use broad language that does not exclude qualitative data or use language that define data in exclusively quantitative terms. Researchers need to be made aware of the importance of sharing qualitative data, something journals can help with by singling out in their data policies the types of qualitative data that facilitate verification and reproducibility of research findings and steps researchers can take to make such data safe to share.

Concluding remarks

Scholars generate new research outputs and scientific results at a rate higher than ever before. There are now more than 30,000 peer-reviewed journals used by scholars to publish their research outcomes across all disciplines ("Subscriptions," n.d.). Even though in principle this growth of the number of scholarly articles is good for science and the advance of knowledge, the *article* alone is not sufficient to accurately describe the underlying research in a rigorous and complete way. Empirical research findings need to be supported by their underlying (well-documented) data for others to fully understand the results. Moreover, the underlying data are needed to reproduce the outputs and verify the validity of these results, an essential part of advancing science. Thus, in order for scholarly communication to continue being reliable and complete, articles need to be accompanied with the underlying data (and often also the code).

Journals then play a critical role in making this needed transformation of scholarly communication successful. We show in this study that this change is already being adopted, or even led, by an increasing number of journals by establishing data policies that encourage or require data to be shared at the time of article submission. Only five to 10 years ago, the number of journals requiring data sharing was negligible and mostly limited to biomedical disciplines. We find here that the social sciences have started to embrace this change in the last years. Based on the top 50 journals from each discipline studied here (economics, political science, psychology, anthropology, sociology, and history), more than half of the journals now have a data policy and around a third of the top-ranked journals in economics and political science

require data sharing. We see this as a welcome change in social science. However, many of these data policies have room to improve in clarity and explicit data sharing requirements, especially in history, anthropology, and sociology where data sharing is *required* only for zero, 10, and 10% of the journals respectively.

Our study is also among the first to investigate data policies based on their source, i.e., whether they were created by a publisher, a professional organization, or the journal itself.

References

Altbach, P. (2005). Academic challenges: The American professoriate in comparative perspective. In A. Welch (Ed.), *The Professoriate: Profile of a Profession* (Vol. 7, pp. 147–165). Berlin/Heidelberg: Springer-Verlag. https://doi.org/10.1007/1-4020-3383-4_9

Bierer, B. E., Crosas, M., & Pierce, H. H. (2017). Data authorship as an incentive to data sharing. *New England Journal of Medicine*, *376*(17), 1684–1687. https://doi.org/10.1056/NEJMsb1616595

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Crosas, M., Gautier, J., Karcher, S., Kirilova, D., Otalora, G., & Schwartz, A. (2018. *Replication Data for: Data policies of highly-ranked social science journals* [Data set]. https://doi.org/10.7910/DVN/CZYY1N

Dalrymple, D. (2003). Scientific knowledge as a global public good: Contributions to innovation and the economy. In J. M. Esanu & P. F. Uhlir (Eds.), *The role of scientific and technical data and information in the public domain: Proceedings of a symposium* (pp. 35–51). Washington D.C.: National Academies Press.

Dataverse Project. (n.d.). About the project. Retrieved March 26, 2018, from https://dataverse.org/about

Elman, C., & Kapiszewski, D. (2014). Data Access and Research Transparency in the Qualitative Tradition. *PS: Political Science & Politics*, *47*(01), 43–47. https://doi.org/10.1017/S1049096513001777

Esanu, J. M., & Uhlir, P. F. (Eds.). (2003). *The role of scientific and technical data and information in the public domain: Proceedings of a symposium*. Washington, D.C.: National Academies Press.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (1985). *Sharing research data*. Washington, D.C.: National Academy Press. https://doi.org/10.17226/2033

Freese, J. (2007). Replication standards in quantitative social science: Why not sociology? *Sociological Methods & Research*, *36*(2), 153–172. https://doi.org/10.1177/0049124107306659

Gherghina, S., & Katsanidou, A. (2013). Data availability in political science journals. *European Political Science*, *12*(3), 333–349. https://doi.org/10.1057/eps.2013.8

Gleditsch, N., & Metelits, C. (2003). Replication in International Relations Journals: Policies and Practices. *International Studies Perspective*, *4*(1). https://doi.org/10.1111/1528-3577.04105

Hanson, B., Sugden, A., & Alberts, B. (2011). Making data maximally available. *Science*, *331*(6018), 649–649. https://doi.org/10.1126/science.1203354

Key, E. M. (2016). How Are We Doing? Data Access and Replication in Political Science. *PS: Political Science & Politics*, *49*(02), 268–272. https://doi.org/10.1017/S1049096516000184

Key Perspectives Ltd. (2010). *Data dimensions: Disciplinary differences in research data sharing, reuse and long term viability* (DCC SCARP Synthesis Report). Edinburgh: Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, *28*(3), 444–452. https://doi.org/10.2307/420301

Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual Review of Information Science and Technology*, *45*(1), 247–294. https://doi.org/10.1002/aris.2011.1440450113

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Piwowar, H. A., & Chapman, W. W. (2008). A review of journal policies for sharing research data. *Nature Precedings*, (713). https://doi.org/10.1038/npre.2008.1700.1

Springer Nature. (2018). Research data support. Retrieved March 26, 2018, from https://www.springernature.com/us/authors/research-data-policy

Stodden, V., Guo, P., & Ma, Z. (2013). Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLOS ONE, 8(6), e67111. https://doi.org/10.1371/journal.pone.0067111

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, *115*(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Sturges, P., Bamkin, M., Anders, J. H. S., Hubbard, B., Hussain, A., & Heeley, M. (2015). Research data sharing: Developing a stakeholder-driven model for journal policies: Research Data Sharing. *Journal of the Association for Information Science and Technology*, *66*(12), 2445–2455. https://doi.org/10.1002/asi.23336

Sturges, P., Bamkin, M., Anders, J., & Hussain, A. (2014). *Access to Research Data: Addressing the Problem through Journal Data Sharing Policies*. UK Jisc. Retrieved from https://jordproject.wordpress.com/access-to-research-data-addressing-the-problem-through-journal-data-sharing-policies/

Subscriptions. (n.d.). Retrieved March 30, 2018, from https://www.library.pitt.edu/subscriptions

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLOS ONE*, *6*(6), e21101. https://doi.org/10.1371/journal.pone.0021101

Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., … Yeaman, S. (2013). Mandated data archiving greatly improves access to research data. *The FASEB Journal*, *27*(4), 1304–1308. https://doi.org/10.1096/fj.12-218164

Wicherts, J., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–727. https://doi.org/10.1037/0003-066X.61.7.726

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Zenk-Möltgen, W., & Lepthien, G. (2014a). *Data from: Data sharing in sociology journals* [Data set]. GESIS Data Archive. https://doi.org/10.7802/65

Zenk-Möltgen, W., & Lepthien, G. (2014b). Data sharing in sociology journals. *Online Information Review*, *38*(6), 709–722. https://doi.org/10.1108/OIR-05-2014-0119

Acknowledgements

The authors thank Arianna Galluzzo and Amery Sanders (Syracuse University) for research assistances. The authors also thank Benjamin Campbell, Simo Goshev, Thu-Mai Christian and UNC's Odum Institute for their advice and guidance. Finally, the authors thank the authors of previous studies, all of whom made their data available.

Appendix

Template of first email to editors of journals whose data policies we could not find

Subject: Share your journal's data policies and help improve research publishing

To <name>,

At the Harvard Institute for Quantitative Social Science, we are studying the policies and instructions that scholarly journals provide regarding what authors should do with the data supporting their manuscripts. You may be aware that a range of journals have implemented or are considering creating such policies and guidance in order to improve data accessibility and preservation. As we have been unable to find public information about an explicit data policy or related guidance on the website of <journal_title>, we would appreciate if you could reply to this email with answers to the following requests for information by September 22, 2017.

By contributing to this study, your journal is helping the scholarly publishing community learn more about how to incentivize and incorporate data sharing in publishing workflows. This request for information about your journal should take about five minutes to answer. If you are unable to answer or you are not sure, please consider sharing this email with the member of your journal staff best equipped to.

1. Does <journal_title> have a data sharing policy or provide instructions to authors for sharing the data or supplemental materials associated with their manuscripts?

1a. If yes, please paste the address(es) of the webpages(s) where these instructions are publicly available online so that we can access them for analysis. Also, if your journal sends instructions in documents (e.g. PDFs or MS Word files) that are not publicly available, please attach those documents to your reply email.

1b. If no, does <journal_title> plan to introduce instructions or policies for sharing data or supplemental materials?
2. Please feel free to share additional information about your journal's instructions or data policies here.

3. Please let us know if you would like to be notified when the results of this study are made available.

We thank you for your time spent answering these questions. If you have any questions or concerns, feel free to reply to this email.

Template of second email to editors of journals whose data policies we could not find
Subject: Help improve research publishing by sharing your journal's data policies
To <name>,

Last month we sent an email asking if <journal_title> has a policy or provides any instruction to its authors for sharing the data or supplemental materials associated with their manuscripts. The Harvard Institute for Quantitative Social Science would like to include your journal in its study of such policies and instructions. To share information about <journal_title> that we were unable to find online, please consider replying to this email to answer the questions below by October 18, 2017.

You may be aware that a range of journals have implemented or are considering creating such policies and guidance in order to improve data accessibility and preservation. By contributing to this study, your journal is helping the scholarly publishing community learn more about how to incentivize and incorporate data sharing in publishing workflows.

This request for information about your journal should take about five minutes to answer. If you are unable to answer or you are not sure, please consider sharing this email with the member of your journal staff best equipped to.

1. Does <journal_title> have a data sharing policy or provide instructions to authors for sharing the data or supplemental materials associated with their manuscripts?

1a. If yes, please paste the address(es) of the webpages(s) where these instructions are publicly available online so that we can access them for analysis. Also, if your journal sends instructions in documents (e.g. PDFs or MS Word files) that are not publicly available, please attach those documents to your reply email.

1b. If no, does <journal_title> plan to introduce instructions or policies for sharing data or supplemental materials?

2. Please feel free to share additional information about your journal's instructions or data policies here.

3. Please let us know if you would like to be notified when the results of this study are made available.

We thank you for your time spent answering these questions. If you have any questions or concerns, feel free to reply to this email.

Definitions of data policies used in similar studies

Do authors of other studies define what they do and do not consider to be a data policy? When they do, what are their definitions?

1. "Replication in International Relations Journals: Policies and Practices" Gleditsch & Metelits 2003
   Authors do not define what they consider to be a data policy

2. "Data Availability in Political Science Journals" Gherghina & Katsanidou 2013
   Study's authors do not explicitly state what they and do not consider to be a data policy, but they write in their "Elements of a Data Availability Policy" section that the "first element the detailed list of what has to be provided by the authors" (p. 337). They write in their "Research Design" section that the journals they determined do have publicly available data policies "provide extensive guidelines for authors" (p. 339).

   However, there's evidence that they have or would consider a policy a data policy if it includes the term supplemental material but not the term data or dataset: "American Politics Research, Communist and Post-Communist Studies, Electoral Studies, and IPSR ... encourage their contributors to submit supplementary materials" (p. 343)

3. "Data Sharing in Sociology Journals" Wolfgang Zenk-Möltgen and Greta Lepthien 2014
   The study's authors don't define what they consider to be a data policy. However, when comparing their findings to ours, it appears that our definition of a data policy is narrower: This study and our study reviewed the policies of the same 46 journals. 21 of those journals that Wolfgang Zenk-Möltgen and Greta Lepthien coded as having data policies in 2014 were coded in our 2018 study as not having data policies. We think it's highly unlikely that so many journals abandoned their data policies. Instead, we think that the authors determined that those journals had data policies either because they were using a broader definition of a data policy or because they took into account each journal's association with a professional organization, such as the Association of Learned

and Professional Society Publishers, which we didn't see mentioned in those journals' policies.

4. "Research data sharing: developing a stakeholder-driven model for journal policies" Paul Sturges et al.

   The study's authors "found it impossible to totally ignore what are described as supplemental materials which might be deposited along with data." But "[w]ithout ignoring such materials we sought to concentrate on the essential data that supports results." It's unclear how many of the journals they code as having a data policy use terms such as supplemental material or supplemental information.