The FASEB Journal • Life Sciences Forum

Mandated data archiving greatly improves access to research data

Timothy H. Vines,*,†,[1] Rose L. Andrew,* Dan G. Bock,* Michelle T. Franklin,*,‡ Kimberly J. Gilbert,* Nolan C. Kane,*,§ Jean-Sébastien Moore,* Brook T. Moyers,* Sébastien Renaut,* Diana J. Rennison,* Thor Veen,* and Sam Yeaman*

*Department of Biodiversity, University of British Columbia, Vancouver, British Columbia, Canada; †Editorial Office, *Molecular Ecology*, Vancouver, British Columbia, Canada; ‡Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada; and §Ecology and Evolutionary Biology Department, University of Colorado at Boulder, Boulder, Colorado, USA

ABSTRACT **The data underlying scientific papers should be accessible to researchers both now and in the future, but how best can we ensure that these data are available? Here we examine the effectiveness of four approaches to data archiving: no stated archiving policy, recommending (but not requiring) archiving, and two versions of mandating data deposition at acceptance. We control for differences between data types by trying to obtain data from papers that use a single, widespread population genetic analysis, STRUCTURE. At one extreme, we found that mandated data archiving policies that require the inclusion of a data availability statement in the manuscript improve the odds of finding the data online almost 1000-fold compared to having no policy. However, archiving rates at journals with less stringent policies were only very slightly higher than those with no policy at all. We also assessed the effectiveness of asking for data directly from authors and obtained over half of the requested datasets, albeit with ~8 d delay and some disagreement with authors. Given the long-term benefits of data accessibility to the academic community, we believe that journal-based mandatory data archiving policies and mandatory data availability statements should be more widely adopted.—Vines, T. H., Andrew, R. L., Bock, D. G., Franklin, M. T., Gilbert, K. J., Kane, N. C., Moore, J-S., Moyers, B. T., Renaut, S., Rennison, D. J., Veen, T., Yeaman, S. Mandated data archiving greatly improves access to research data. *FASEB J.* 27, 1304–1308 (2013). www.fasebj.org**

Key Words: science policy · reproducibility · Joint Data Archiving Policy · population genetics · STRUCTURE

ARCHIVING THE DATA UNDERLYING scientific papers is an essential component of scientific publication and its subsequent reproducibility (1–3), but very few papers actually make the underlying data available (4). In response to this gap between the needs of science and author behavior, a number of journals have introduced data archiving policies. Here, we evaluate the effectiveness of these policies by comparing journals that have no stated data archiving policy, journals that recommend data archiving, and journals that mandate archiving prior to publication. Journals that mandate data archiving fall into two further subgroups: those that require an explicit data availability statement and those that do not. We ask two questions: whether having any kind of data archiving policy improves the likelihood of the data being available online, and whether the type of data archiving policy has any effect on the likelihood of obtaining the data.

We recently assembled datasets from a range of journals for a study of the reproducibility of commonly used population genetic analyses (5). Here, we use this opportunity to examine whether data archiving policy (or lack thereof) was associated with the proportion of datasets we were able to obtain from a journal. As papers within even a single journal contain many different types of data, we restricted both this and our reproducibility study to articles using the population genetics program STRUCTURE (6). We chose STRUCTURE because it is widely used in ecology and evolution, and because the underlying data are compiled in a table of microsatellite, amplified fragment length polymorphism, or single-nucleotide polymorphism genotypes, and for the ease of archiving this type of dataset online. For example, the data could be uploaded as supplemental material, or archived on the Dryad repository (7). Dryad was established in 2010 for the preservation of a wide range of data types associated with ecology or evolution articles, and is often used to archive STRUCTURE datasets.

DATA COLLECTION

We used Web of Science to identify articles published in 2011 or 2012 that cited the original description of STRUCTURE (6). We selected journals for each of the four journal categories described above, and excluded those

Abbreviations: *BJLS*, *Biological Journal of the Linnean Society*; BMC, BioMed Central; CI, confidence interval; IF, impact factor; JDAP, Joint Data Archiving Policy; PLoS, Public Library of Science; *TAG*, *Theoretical and Applied Genetics*

[1] Correspondence: Biodiversity Department, University of British Columbia, 6270 University Blvd Vancouver BC, Canada, V6T 1Z4. E-mail: vines@zoology.ubc.ca

doi: 10.1096/fj.12-218164

0892-6638/13/0027-1304 © FASEB

that had <5 eligible papers. We complemented our list of papers by searching for additional articles that used STRUCTURE on the journal website. Papers that used DNA sequence data were excluded, as preparing raw sequence data from, for example, GenBank, for reanalysis with STRUCTURE was found to be very time consuming.

We found four eligible journals with no stated data archiving policy: *Conservation Genetics*, *Crop Science*, *Genetica*, and *Theoretical and Applied Genetics* (*TAG*).

There were four eligible journals that had some sort of data archiving policy but stopped short of mandating archiving for all data [*BMC Evolutionary Biology* (BMC, BioMed Central), *Biological Journal of the Linnean Society* (*BJLS*), *Journal of Heredity*, and *PLoS One* (PLoS, Public Library of Science)]. These policies were retrieved from the author guidelines in mid-2011 and are available on Dryad (doi: 10.5061/dryad.6bs31). The latter three journals ask that the data be placed onto an online archive whenever one exists. For STRUCTURE data, Dryad is the most commonly used repository, and indeed the policies of the last two journals (*Journal of Heredity* and *PLoS One*) explicitly mention Dryad. There is thus an expectation for three of these four journals the data should be available somewhere online, most likely on Dryad. For *BMC Evolutionary Biology*, the data will only be online if the authors have decided to share it at publication. The individual policies are as follows:

First, *BMC Evolutionary Biology* states that "submission . . . implies that readily reproducible materials described in the manuscript, including all relevant raw data, will be freely available to any scientist wishing to use them for noncommercial purposes," and at that time did not require that data appear in an online archive. This policy has been in place since 2009.

Second, the *Biological Journal of the Linnean Society* has the policy "we recommend that data for which public repositories are widely used, and are accessible to all, should be deposited in such a repository prior to publication." This policy was introduced in January 2011, and we hence only considered papers submitted after this date.

Third, *Journal of Heredity* "endorses the principles of the Joint Data Archiving Policy [see below] in encouraging all authors to archive primary datasets in an appropriate public archive, such as Dryad, TreeBASE, or the Knowledge Network for Biocomplexity." As with *BJLS*, this policy was introduced in January 2011, and we hence only considered papers submitted after this date.

Fourth, *PLoS One* has had a policy on data sharing in place since 2008, and one statement is as follows: "If an appropriate repository does not exist, data should be provided in an open access institutional repository, a general data repository such as Dryad, or as Supporting Information files with the published paper."

Finally, there were four journals that adopted a mandatory data archiving policy [known as the Joint Data Archiving Policy (JDAP); ref. 1], which states "[Journal X] requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive." For these journals, we excluded papers submitted before the policy came into force: January 2011 for *Molecular Ecology*, *Journal of Evolutionary Biology*, and *Evolution*, and March 2011 for *Heredity*. Of these four, two (*Molecular Ecology* and *Heredity*) additionally require that authors include a data availability statement within each accepted manuscript; these sections describe the location (typically the database and accession numbers) of all publicly available data.

For all 229 eligible papers, we then checked whether the STRUCTURE genotype data were available either as supplemental material or elsewhere online, such as on the Dryad archive (7). Our results are shown in **Table 1** and **Fig. 1**, and the data and *R* code used in the analysis are archived on Dryad (doi: 10.5061/dryad.6bs31).

STATISTICAL ANALYSIS

To evaluate the statistical support for an association between the presence/absence of an archiving policy and whether the STRUCTURE data could be found online, we fitted a mixed effects logistic regression. The response variable was whether the data from a paper were available online, coded as 0 for not available and 1 for available. The predictor variable was either "no policy" or "archiving policy," and journals were included as a random effect within each category.

Having any sort of archiving policy did lead to a significant improvement in the probability of the data being online (likelihood ratio test statistic: 4.27, $P=0.038$), such that the odds of getting the data were ~25 times higher [95% confidence interval (CI): 1.5–416.7].

We then tested how well each type of archiving policy compared to having no policy at all. As above, we used a mixed effects logistic regression. Again, the response variable was whether the data from a paper were available online, coded as 0 for not available and 1 for available. The predictor variable was policy type, and the categories were "no policy," "recommend archiving," "mandate archiving, no data statement," and "mandate archiving, with data statement." Journals were a random effect within each policy type. The overall model found that policy type did have a very significant effect on data availability (likelihood ratio test statistic: 28.06, $P<0.001$).

TABLE 1. *Number of eligible articles per journal and number for which data were obtained from online databases*

Policy	Journal	Eligible articles	Data online
No policy	*Conservation Genetics*	47	1
	Crop Science	12	1
	Genetica	8	1
	TAG	21	0
Recommend data archiving	*BMC Evolutionary Biology*	13	1
	BJLS	13	3
	Journal of Heredity	12	0
	PLoS One	51	6
Mandatory data archiving	*Journal of Evolutionary Biology*	10	3
	Evolution	6	3
	Heredity	7	7
	Molecular Ecology	28	27

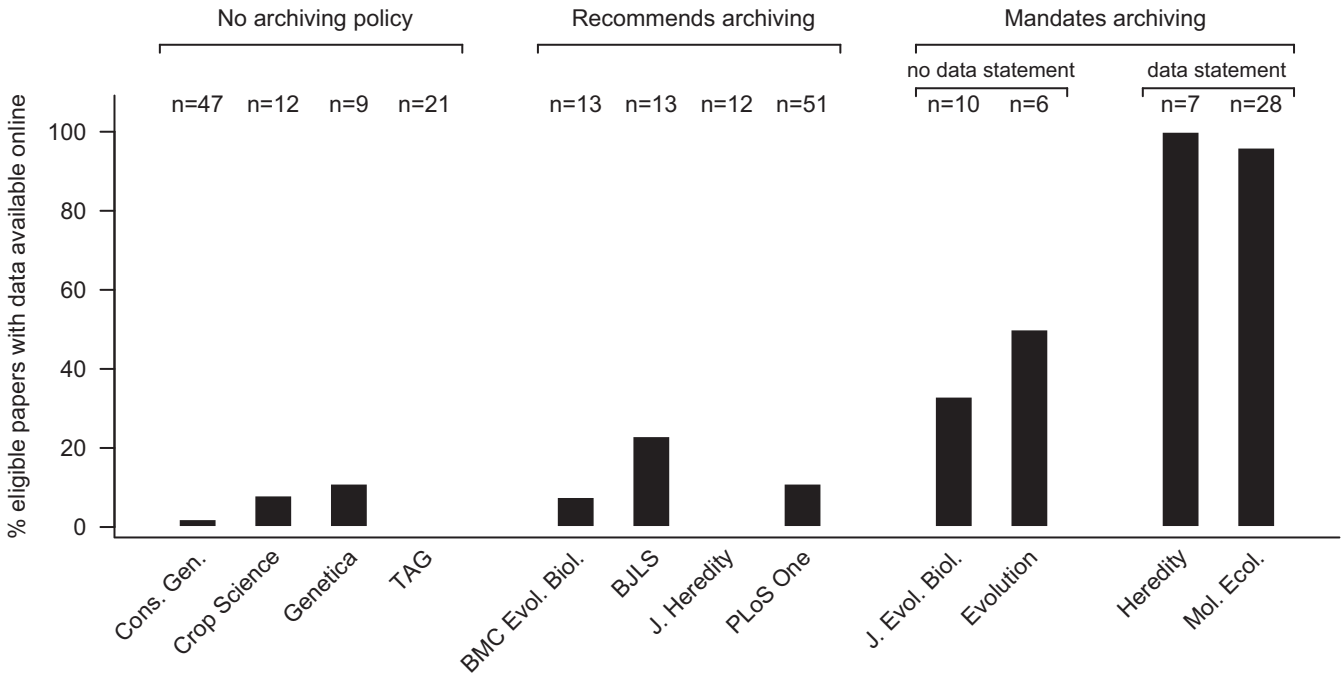

Figure 1. Percentage of eligible papers published in 2011 that made their data available online, by journal. Number of eligible papers is shown above each column. Within the "mandate archiving" group, "data statement" denotes the journals that require a data accessibility statement in the manuscript, and "no data statement" denotes those that do not.

Since this is a logistic model, we can readily calculate the effect that the different policy types have on the likelihood that the data will be available. We explore these odds for each type of policy below, using "no policy" as the baseline.

Having a "recommend archiving" policy made it 3.6 times more likely that the data were online compared to having no policy. However, the 95% CI overlapped with 1 (0.96–13.6); hence, this increase in the odds is not significant. Overall, recommending data archiving is only marginally more effective than having no policy at all.

The data were 17 times more likely to be available online for journals that had adopted a mandatory data archiving policy but did not require a data accessibility statement in the manuscript. This odds ratio was significantly >1 (95% CI: 3.7–79.6).

For "mandate archiving" journals where a data accessibility statement is required in the manuscript, the odds of finding the data online were 974 times higher compared to having no policy. The 95% CI on these odds is very wide (97.9–9698.8), but nonetheless shows that the combination of a mandatory policy and an accessibility statement is much more effective than any other policy type.

REQUESTING DATA DIRECTLY FROM AUTHORS

A number of the "recommend archiving" policies state that the data should also be freely available from the authors by request (see the Journal Policies file at doi: 10.5061/dryad.6bs31); hence, we wanted to evaluate whether obtaining data directly from authors is an effective approach. Part of the dataset collection for our reproducibility study (5) involved e-mailing authors of papers from two of the "recommend archiving" journals (*BMC Evolutionary Biology* and *PLoS One*) and requesting their STRUCTURE input files. Here, we examine how often these requests led to us obtaining the data. We did not e-mail the authors of articles where the data were already available online. A detailed description of our data request process appears on Dryad (doi: 10.5061/dryad.6bs31), but we essentially contacted corresponding and senior authors of each article up to 3 times over a 3-wk period, and recorded if and when the data were received.

We obtained data directly from the authors for 7 of the 12 eligible articles in *BMC Evolutionary Biology*, and 27 datasets from 45 articles from *PLoS One* (Table 1). All seven of the *BMC Evolutionary Biology* datasets arrived between 8 and 14 d after our initial request. Ten of the *PLoS One* datasets came within 1 wk, 13 came between 8 and 14 d, and 4 arrived between 15 and 21 d. Unlike the online data, which could generally be obtained within a few minutes, the requested datasets took a mean of 7.7 d to arrive, with one author responding that the dataset had been lost in the year since publication. More than one e-mail had to be sent to the corresponding and/or senior author for 53% of papers, and the authors of 29% of the papers did not respond to any of our requests. No data were received >21 d after our initial request. We also note that requesting data *via* e-mail did upset some authors, particularly when they were reminded of the journal's data archiving policy or when multiple e-mails were sent.

Our average return of 59% in an average of 7.7 d is markedly better than has been reported in similar studies: Wicherts *et al.* (8) received only 26% of requested datasets after 6 mo of effort with authors of 141 psychology articles, and Savage and Vickers (9) received only 1 of 10

datasets requested from articles in *PLoS Medicine* and *PLoS Clinical Trials*. In a 1999 study, Leberg and Neigel (10) e-mailed the authors of 30 articles that contained an incomplete description of their sequence dataset, but received the requested data from just one of them. Since the latter study and ours both involved the evolutionary biology community, it appears that attitudes to data sharing have improved dramatically over the past decade. However, the two more recent studies that used human data still had low success rates, perhaps because privacy and consent issues are a significant impediment to data sharing in these fields.

CONCLUSIONS

Our results demonstrate that journal-based data archiving policies can be very effective in ensuring that research data are available to the scientific community, especially when journals require that a data accessibility statement appear in the manuscript. The "recommend archiving" group of journals encompassed the broadest spread of policy types, yet as a whole only had 10 of 89 datasets available. The policies range from a simple "Submission . . . implies that . . . all relevant raw data, will be freely available to any scientist wishing to use them for noncommercial purposes" at *BMC Evolutionary Biology* to an endorsement of the full JDAP at *Journal of Heredity*. However, none of these policies led to >23% of the data being available online (at *BJLS*), and there was no significant difference between the success of this policy type and having no policy at all.

Interestingly, *PLoS One*'s very comprehensive policy, which is >1000 words long and contains statements such as "data should be provided in an open access institutional repository, a general data repository such as Dryad, or as Supporting Information files with the published paper" was only marginally more effective than *BMC Evolutionary Biology*'s simple request that the data be freely available, with 11 and 7% of the data online, respectively.

The difference between *PLoS One* and the "mandate archiving" journals may arise because the wide breadth of subject areas in *PLoS One* precludes having a policy with the bald simplicity of the JDAP: "[Journal X] requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive." Even though the portion of *PLoS One*'s author community that uses STRUCTURE broadly overlaps with the authors of the papers in the JDAP journals, it may be that the lack of a single strong statement leads to much lower compliance. One simple remedy for this situation might be the introduction of a mandatory data accessibility statement in all manuscripts. For fields where archiving is not (yet) standard practice, this could state that the data were available from the authors, but in fields where archiving is expected the authors would indicate where their data were available online.

More broadly, a study by Piwowar and Chapman (11) on 397 microarray datasets from 20 journals also found that having a "strong" (*i.e.*, close to mandatory) data archiving policy led to a high proportion (>50%) of the datasets being available online. Journals that had a "weak" policy (*i.e.*, recommended archiving) had just over 30% of microarray datasets available, and journals with no policy had only about 20% availability. Furthermore, the researchers also found that a journal with an impact factor (IF) of 15 was 4.5 times more likely to have the microarray data online than a journal with an IF of 5. We find a similar effect in our data: using the 2010 IFs, we were 3.2 times more likely to find the data online for a journal with an IF of 5.0 (the average IF of the JDAP journals) compared to those with an IF of 2.2 (the average IF of the "no policy" journals); details of this analysis are available online (doi: 10.5061/dryad.6bs31). We are able to exclude higher IF as the primary cause of the high rate of data archiving in the JDAP journals: in 2010 (before the mandatory archiving policy was introduced), none of the 27 eligible articles in the *Journal of Evolutionary Biology, Evolution,* or *Heredity* had archived their data, even though their IFs were essentially the same in 2010 and 2011 (*Molecular Ecology* recommended archiving in 2010 and was excluded from this comparison). This result suggests that the introduction of the JDAP policy in 2011 was primarily responsible for the abrupt rise in the proportion of articles in these three journals that archived their data. However, it is possible that IF still plays a role, as only journals with a high IF may feel able to introduce stringent archiving policies. The positive effects of a strongly worded data archiving statement were also confirmed by a much larger study involving 11,603 microarray datasets (12).

Requesting data directly from authors can also provide access to research data, but this approach can be hampered by delays and the potential for disagreement between requester and the authors. Furthermore, the availability of datasets directly from authors will only decrease as time since publication increases. This is particularly true when researchers leave science or when data that are stored on laboratory computers or websites get misplaced (13, 14).

Even though our results strongly emphasize the value of public databases for archiving scientific data, these databases do require ongoing financial support; this money may come from funding agencies, journal publishers, libraries, or even individual researchers. A recent study put the cost of running the Dryad database at around $400,000/yr; these costs include the maintenance of their archive and the addition and curation of an extra 10,000 datasets/yr. For comparison, the same amount spent by a funding agency on basic research would generate ~16 new publications (15). Given that the long-term availability of these data allows for meta-analyses, the checking of previous results, and not collecting the same data again, money spent on data archiving is extremely cost effective. In light of all these advantages, we believe that journal-based mandatory data archiving policies and data accessibility statements should be more widely adopted. FJ

The authors thank Heather Piwowar, Loren Rieseberg, Phil Davis, and Mike Whitlock for comments on an earlier version of the manuscript, and Arianne Albert for help with the statistics. The authors also express gratitude to the many authors who shared their data.

REFERENCES

1. Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., and Moore, A. J. (2010) Data archiving. *Am. Nat.* **175,** 145–146
2. Wolkovich, E. M., Regetz, J., and O'Connor, M. I. (2012) Advances in global change research require open science by individual researchers. *Global Change Biol.* **18,** 2102–2110
3. Peng, R. D. (2011) Reproducible research in computational science. *Science* **334,** 1226
4. Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M. A., and Ioannidis, J. P. A. (2011) Public availability of published research data in high-impact journals. *PLoS One* **6,** e24357
5. Gilbert, K. J., Andrew, R. L., Bock, D. G., Franklin, M. T., Kane, N. C., Moore, J.-S., Moyers, B. T., Renaut, S., Rennison, D. J., Veen, T., and Vines, T. H. (2012) Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program STRUCTURE. *Mol. Ecol.* **21,** 4925–4930
6. Pritchard, J. K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959
7. Vision, T. (2010) Open data and the social contract of scientific publishing. *Bioscience* **60,** 330–331
8. Wicherts, J. M., Borsboom, D., Kats, J., and Molenaar, D. (2006) The poor availability of psychological research data for reanalysis. *Am. Psychol.* **61,** 726–728
9. Savage, C. J., and Vickers, A. J. (2009) Empirical study of data sharing by authors publishing in PLoS journals. *PLoS One* **4,** e7078
10. Leberg, P. L., and Neigel, J. E. (1999) Enhancing the retrievability of population genetic survey data? An assessment of animal mitochondrial DNA studies. *Evolution* **53,** 1961–1965
11. Piwowar, H. A., and Chapman, W. W. (2010) Public sharing of research datasets: a pilot study of associations. *J. Informetrics* **4,** 148–156
12. Piwowar, H. A. (2011) Who shares? Who doesn't? Factors associated with openly archiving raw research. *PLoS One* **6,** e18657
13. Wren, J. D., Grissom, J. E., and Conway, T. (2006) E-mail decay rates among corresponding authors in MEDLINE. The ability to communicate with and request materials from authors is being eroded by the expiration of e-mail addresses. *EMBO Rep.* **7,** 122–127
14. Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., and Stafford, S. G. (1997) Nongeospatial metadata for the ecological sciences. *Ecol. Appl.* **7,** 330–342
15. Piwowar, H. A., Vision, T. J., and Whitlock, M. C. (2011) Data archiving is a good investment. *Nature* **473,** 285–285

Received for publication August 20, 2012.
Accepted for publication December 18, 2012.