Individual costs and benefits of sharing open data in ecology and evolution

Sandrine Soeharjono[1] and Dominique G. Roche[1,2,3*]

[1] *Département de science biologiques, Université de Montréal, Montréal, Canada*
[2] *Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland*
[3] *Department of Biology, Carleton University, Ottawa, Canada*

*e-mail for correspondence: dominique.roche@mail.mcgill.ca

Key words: data sharing, FAIR data, open science, public data archiving, reproducibility, transparency

Abstract

Open data facilitate reproducibility and accelerate scientific discovery but are hindered by perceptions that researchers bear costs and gain few benefits from publicly sharing their data, with limited empirical evidence to the contrary. We surveyed 140 faculty members working in ecology and evolution across Canada's top 20-ranked universities and found that more researchers report benefits (47.9%) and neutral outcomes (43.6%) than costs (21.4%) from sharing data. Benefits were independent of career stage and gender, but men and early career researchers were more likely to report costs. We outline proposed mechanisms to reduce individual costs of data sharing and increase benefits.

Main text

Open and FAIR (Findable, Accessible, Interoperable, Reusable) data are a cornerstone of transparent, reproducible, and reusable scientific research. Therefore, many funding bodies and scientific journals are adopting policies mandating open data as a condition of funding or publication[1-5]. In ecology and evolution, the first journal began requiring open data in 1999[6], and the number of journals with a mandatory open data policy currently exceeds 60[7]. Despite broad support for data sharing by funders, publishers, and policy makers, many academics resist making their research data publicly available[5,8-11]. One reason might be that, while the collective

benefits of open data are broadly acknowledged[3,12-15], there is less evidence that publicly sharing research data benefits individuals[16-21].

Open data practices also remain controversial among academics due to perceived individual costs in today's hypercompetitive research environment[22-26]. Many researchers feel a sense of ownership over their data and fear that potential individual benefits do not make up for future publications lost by relinquishing priority of access to the data they collected – the fear of being 'scooped'[4,6,13,27]. Researchers have also voiced concerns that sharing their data could result in misinterpretations or misuses by third parties[18] as well as public shaming if errors are identified in their work[4,5]. Currently, apprehensive authors can avoid publishing in journals that mandate open data[16], ignore funder and/or journal policies altogether[28], or share their data in ways that make them difficult or impossible to re-use[29,30], undermining efforts to shore up reproducibility. Understanding and addressing the real and/or perceived costs of data sharing is essential to bolster researcher's engagement in open and FAIR data practices[4].

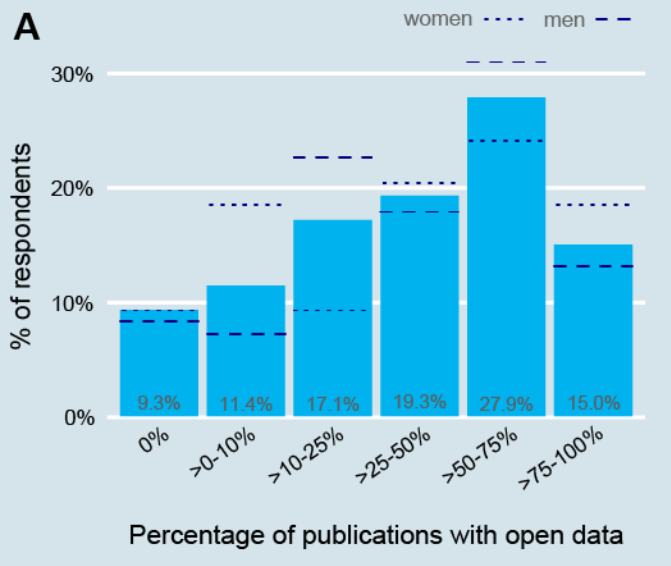
Research to date has centered on understanding the opinions of researchers in order to identify 'motivations and enablers' or 'fears and barriers' to data sharing[5,10,18,26,31-34]. A recent systematic review of researcher's attitudes towards open data identified five factors that strongly influence data sharing: seniority, age, lack of time, loss of control over one's data, and data misappropriation[18]. Although loss of control and data misappropriation are frequently cited by researchers as reasons for not sharing[35], it is unclear whether these fears are perceptions or legitimate concerns. We contacted 351 principal investigators (PIs) self-described as working in the fields of ecology and evolution at the 20 highest-ranked Canadian universities to investigate actual positive and negative outcomes of sharing open data: 140 PIs completed our questionnaire, including 54 women, 84 men, and 2 non-binary persons. (methods on p. 8)

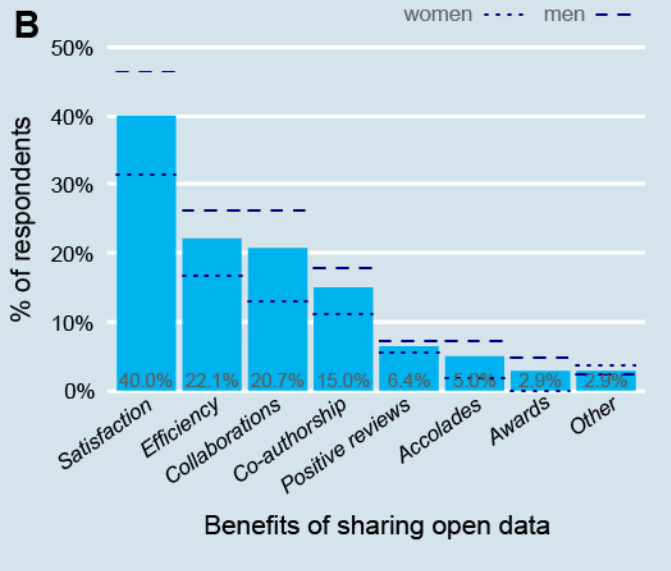
A large majority of survey respondents reported always or occasionally sharing open data alongside their published papers (84.3%), believing that open data are beneficial for society (88.6%), and supporting mandatory open data policies (77.9%) (Table 1; Fig. 1A). Support for mandatory policies was similar among genders ($\chi^2_{(2)}$=3.54, P=0.17) and career stages ($\chi^2_{(2)}$=3.65, P=0.16). These views and practices by PIs in ecology and evolution align with recent reports of growing support for open data initiatives by scientists across disciplines[35].

Table 1. Opinions and experiences of PIs in ecology and evolution regarding open data. The percentage of respondents (N=140) having selected one of three options or answered yes or no to the following questions: **1.** Do you tend to share your data publicly even if open data are not required by the journal in which you have published? **2.** Are you generally supportive of mandatory open data policies by journals and/or funding agencies? **3.** Do you believe that publicly sharing non-sensitive research data is beneficial to society (where non-sensitive data protect the identity of study participants and do not put organisms at risk of exploitation or disturbance)? **4.** Do you believe that mandatory open data requirements by journals or funders impose excessive time demands on researchers? **5.** Do you believe that the time requirements and effort needed to share your research data have decreased over time (e.g. as a result of prior experiences sharing data and/or improved data management practices)? **6.** Have you experienced benefits from publicly sharing your data? **7.** Have you experienced costs from publicly sharing your data? 95% confidence intervals are indicated in light grey. Questions in the table are paraphrased. Note that 35.0% of respondents reported experiencing only benefits from sharing open data, 12.9% experienced both costs and benefits, and 8.6% experienced only costs.

Question	Always (%)		Occasionally (%)		Never (%)	
1. Do you share open data?	37.9	[29.8 - 45.9]	46.4	[38.2 - 54.7]	15.7	[9.7 - 21.7]
	Yes (%)		**Indifferent (%)**		**No (%)**	
2. Do you support mandatory open data?	77.9	[71.0 - 84.7]	10.7	[5.6 - 15.8]	11.4	[6.2 - 16.7]
	Yes (%)				**No (%)**	
3. Are open data beneficial to society?	88.6	[83.3 - 93.8]			11.4	[6.2 - 16.7]
4. Does data sharing require excessive time?	34.3	[26.4 - 42.1]			65.7	[57.9 - 73.6]
5. Have time requirements decreased for you?	40.0	[31.9 - 48.1]			60.0	[51.9 - 68.1]
6. Have you benefited from sharing open data?	47.9	[39.6 - 56.1]			52.1	[43.9 - 60.4]
7. Have you incurred costs from sharing open data?	21.4	[14.6 - 30.0]			78.6	[71.8 - 85.4]

Almost half of respondents (47.9%) reported positive outcomes from sharing open data (Table 1), without notable differences between genders (odds ratio=0.77, z=-1.5, *P*=0.134) and career stages (odds ratio=1.0, z=0.32, *P*=0.75) (Fig 2A). The most common benefit was a personal sense of satisfaction from openly sharing one's data, but many PIs also reported benefits in terms of productivity and career advancement (Fig 1B). 43.6% of PIs reported experiencing neither costs nor benefits.





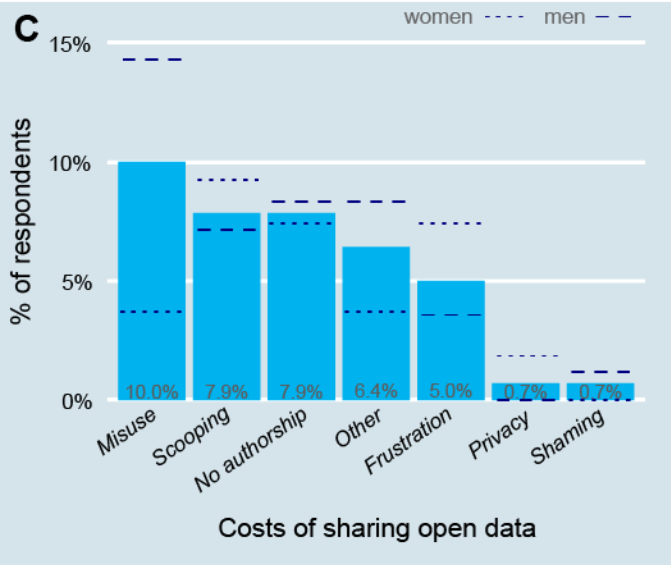
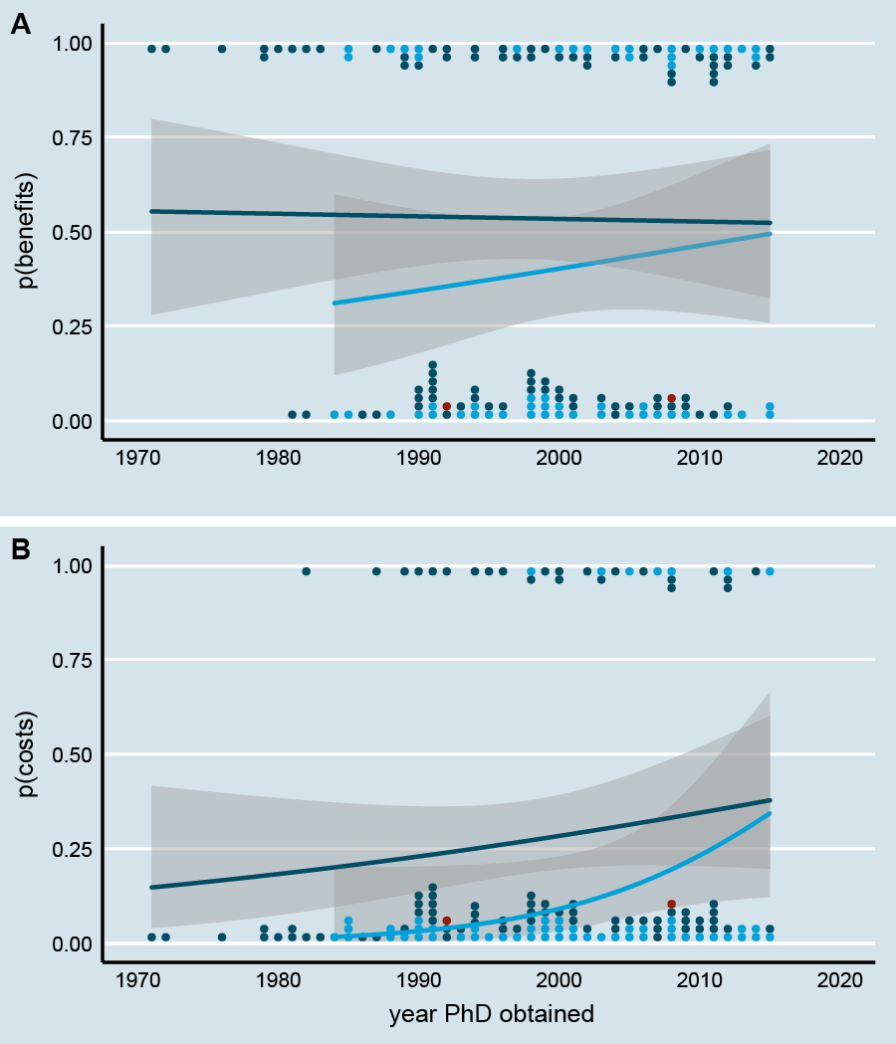


Fig. 1 Costs and benefits of sharing open data in ecology and evolution. The percentage of respondents (N=140: 54 women, 84 men, 2 non-binary) having: **(A)** published empirical studies with associated open data between 2015-2020; **(B)** experienced benefits from open data; and **(C)** experienced costs from open data. Dashed and dotted lines indicate the percentage of men and women having selected each category, respectively. Bars (and lines) do not add up to 100% in panels B and C since respondents could select more than one category. **Benefits of open data include**: a personal sense of satisfaction from having published a research product other than a scientific paper; greater efficiency from adopting better data management practices; new collaborations (whether to reuse the open data or not); offers of co-authorship from researchers having re-used the data; positive reviews on applications for grants, fellowships or manuscripts; accolades from peers; awards or promotions, other benefits. **Costs of open data include:** data misuse or misinterpretation by a third party; being 'scooped' on a paper; lack of warranted co-authorship on a study reusing open data; other costs; frustration from having to share data that researchers would have preferred to keep private; putting humans or other species at risk of privacy infringement or disturbance/exploitation; and public shaming or reputational damage from errors being identified in a researcher's work.

Approximately one fifth of PIs (21.4%) reported having previously experienced some form of negative outcome from sharing open data (Table 1). The probability of experiencing negative outcomes was greater for men than women (odds ratio=5.85, z=2.21, *P*=0.027) and tended to be higher for researchers at an early career stage (odds ratio=1.04, z=1.93, *P*=0.054) (Fig 2B). The effect of career stage was independent of gender (z=-1.28, *P*=0.200). Why men experience costs more frequently than women is unclear. One possible explanation is that costs are more readily perceived by men than women, which translates into higher self-reported costs among male respondents. Psychology research has consistently found that men tend to be more competitive than women[36]. For example, men in our study might have perceived that their data were 'misused' more often than women if they considered the reuser to be a competitor. In contrast, our finding that early career researchers (ECRs) tend to experience more costs than their senior colleagues aligns with concerns expressed by others[27]. The few studies having examined the effect of age and seniority on data sharing attitudes report that, while younger researchers think more favorably about data sharing and reuse[32,33], ECRs tend to be more fearful and reluctant to share their data than senior researchers[18,26,32,33] (but see[11]). The most prevailing costs experienced by PIs of all genders and career stages were related to data misuse or misinterpretation; only one respondent reported experiencing reputational damage from errors being identified in their work (Fig 1C).

Fig. 2 The influence of gender and career stage on the probability of researchers in ecology and evolution experiencing (A) benefits and (B) costs from sharing open data. Men (N=84) are shown as dark blue dots, women (N=54) as light blue dots, and non-binary persons (N=2) as green dots. Predictions from binomial generalized linear models are indicated as dark blue (men) and light blue (women) lines, along with 95% confidence intervals.



One in three respondents felt that that mandatory open data requirements by journals or funders imposed excessive time demands on researchers (Table 1). This was echoed in several free-form comments by PIs arguing that publicly sharing data in a reusable format can be '*extremely time consuming*' and felt little value in doing so given a '*sense that they* [the data] *are very rarely used*'. Indeed, while the societal benefits of sharing 'big data' are frequently acknowledged[37,38], we lack empirical evidence that 'long-tail' research data (data from small) are frequently reused, at least presently[39]. Similarly, several researchers expressed support for open data but felt that time demands to clean and annotate data (and sometimes the associated code or script) were '*not excessive but non-trivial*', '*definitely significant*', or '*huge'*, depending on the type and amount of data to be shared. Some PIs expressed that data sharing expectations promote the adoption of

good data curation early on, ultimately reducing time demands for data management and research more generally. This sentiment was reflected in the fact that nearly half (47.5%) of PIs who had previously published open data reported a reduction in the time and effort needed to share their data. The absence of time gains among other respondents (52.5%) could result from several PIs having shared few datasets (47.8% shared data with less than 50% of their papers; Fig. 1A) and observations by PIs that frequent differences among studies and datasets mean that "*each data submission is de novo*". A recurring comment was that research groups seldom have qualified personnel to assist with research data management (RDM) and that funders and universities must dedicate resources for capacity building to help researchers comply with new RDM and open data policies.

When asked what initiatives by journals, funding agencies, and universities would increase their participation in open data, PIs formulated extensive and specific recommendations centred on three themes. The comprehensive list of anonymous comments in the free-form sections of the questionnaire is provided as supplementary information (Table S1).

***Better guidelines, standards, and training for data sharing*.** Editorial open data policies are often unclear[2,40], resulting in what one respondent characterized as '*clunky and uncertain processes, especially around sensitive data*'. Respondents emphasized the need for clearer guidelines on what, where and how data should be shared. Compounding this issue is a lack of well-established, domain-specific standards to facilitate sharing (and reusing) data in many subdisciplines of ecology and evolution[41]. The need for adequate training to familiarize students and PIs with good data management and curation practices was repeatedly highlighted as critical.

Greater support for RDM and equity in sharing practices. Two leading concerns of PIs with open data were financial and time requirements, resulting in what several PIs described as inequitable requirements on labs with different levels of funding and personnel. Some PIs suggested small grants to organize data files for public access; however many insisted that a sustainable approach to open data requires funding agencies to mandate, check, and provide adequate financial and technical resources to support data sharing.

Better incentives for sharing and better protection from potential negative outcomes. PIs clearly communicated that a greater adoption of data sharing practices requires greater

recognition of open data in research funding and career advancement decisions. A popular suggestion was that funders and universities ask applicants to list open datasets and consider these in their evaluation of researchers as is currently done for student/personnel training and, increasingly, for contributions to Equity, Diversity and Inclusion. In parallel, one suggestion was that journals reduce page, colour figure or open access fees to incentivize data sharing. Beyond increasing benefits, PIs also stressed the need to formalize mechanisms for avoiding negative outcomes, such as agreed-upon best practices for reusing open data and attributing much needed credit to the data collectors.

Overall, most comments by PIs were highly constructive with very few stating they would never make their data open (i.e., '*when hell freezes over*'). Some recommendations by PIs were impractical for advancing open data (e.g., data should be available on request and contingent on sharing a research proposal) and some pointed to developing mechanisms that already exist (e.g., trusted repositories[42], embargo options[16], citeable DOIs for datasets[43], access control for sensitive data[44]). The latter highlights a lack of familiarity with current open data practices among some PIs, which is not surprising since open data are a relatively new expectation in most subdisciplines of ecology and evolution.

In conclusion, our findings indicate strong support for mandatory open data policies among PIs in ecology and evolution and a willingness to publicly share research data. The practical recommendations highlighted here can readily be implemented by funders, universities and journals to incentivize participation in data sharing and reduce potential costs to researchers.

Methods

A questionnaire was sent to all faculty members in the biology departments (or equivalent) of the 20 highest-ranked Canadian universities (based on the 2019 Times Higher Education World University Rankings) who were self-described researchers in the fields of ecology and evolution (N=351). The biology department website of each university was reviewed to identify all researchers conducting research in ecology and/or evolution (E&E) with a rank of assistant, associate or full professor. We did not contact adjunct professors and faculty members whose

research primarily focused on molecular biology, genetics, genomics, bioinformatics, theoretical biology, comparative physiology, and paleontology.

Researchers were contacted via a personalized e-mail sent to their institutional inbox on January 20th, 2020, to request their participation in our study. Two reminders were sent at 8-day intervals. Participants were informed of the aim of our study prior to accessing the survey, debriefed once the survey was completed, and asked for their written, informed consent allowing their answers to be included in the study. Respondents were given until the end of the survey period, on February 29th, 2020, to answer, modify or withdraw their answers. Researchers who agreed to participate in our study were directed to a questionnaire comprising thirteen multiple choice questions (Supplementary Information Table S2). For each question, respondents had the option of leaving comments or expanding on their answers. Respondent's gender and year of PhD completion were self-reported as part of the questionnaire. The questionnaire was developed using the formr survey software (https://formr.org), which integrates with the R statistical software for data analysis[45]. Respondents' answers were anonymized such that we were blind to their identity during analysis.

If respondents indicated experiencing benefits from sharing open data, they were directed to a multiple-choice question providing options of possible benefits. One option was 'other benefits', and respondents were asked to specify the nature of these benefits if they selected this option. The same procedure applied when respondents were asked about experiencing costs from sharing open data. 'Other benefits' mentioned by respondents included having a backup of their data in the cloud, being contacted by students and professors at other universities where their data were being used for teaching, and more visits to websites aimed at informing policymakers. 'Other costs' mentioned by respondents included delays in publication associated with finding a repository that would accept a large dataset, mistaken concerns raised with journal editors by other researchers, and a corrigendum resulting from a data error (a time cost to the PI but a benefit to science).

Statistical analysis. The two respondents who identified as non-binary were included in the descriptive statistics but not in the statistical analyses; they did not report experiencing either costs or benefits from sharing open data. We used two generalized linear models (GLM) with a binomial distribution of error terms to examine the effect of gender and career stage (year of PhD

obtention) on whether respondents had experienced costs and benefits from sharing open data (N=138). We used a multinomial logistic regression (multinom function in the R package nnet[46]) to examine the effect of gender and career stage on the respondent's level of support for mandatory open data policies (N=138). We treated the predictor variables 'year of PhD completion' and gender as continuous and categorical variables, respectively. GLMs were verified with diagnostic tests using functions in the R package DHARMa[47]. We used likelihood ratio tests and type III sum of squares to extract p-values from the multinomial model. There was no interaction between gender and career stage in any of the three models (all *P*s > 0.12) and the interaction term was removed to extract accurate summary outputs for main effects. All analyses were done in Rv3.6.3[48].

Data and code availability: The anonymized data and the analysis script for this study can be accessed on the Open Science Framework (https://osf.io/dq7zm/?view_only=eb605e8297d14b1faf46f9a97bb15d9c).

Author Contributions: DGR designed the study, SJ collected the data, SJ and DGR analyzed the data, SJ and DGR wrote the paper.

Competing Interests: The authors declare no competing interests.

Acknowledgments: We thank Sandra Binning, Ilias Berberri, Fares Dhane, Félix Lauzon, and Jennifer McClung for assistance with various aspects of this study. We acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (UIF-537860-2018). DGR was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 838237-OPTIMISE.

References

1. Culina, A., van den Berg, I., Evans, S. & Sánchez-Tójar, A. Low availability of code in ecology: A call for urgent action. *PLoS Biol.* **18**, e3000763, doi:10.1371/journal.pbio.3000763 (2020).
2. Sholler, D., Ram, K., Boettiger, C. & Katz, D. S. Enforcing public data archiving policies in academic publishing: A study of ecology journals. *Big Data & Society* **6**, 2053951719836258, doi:10.1177/2053951719836258 (2019).
3. Whitlock, M. C. Data archiving in ecology and evolution: best practices. *Trends Ecol. Evol.* **26**, 61-65, doi:10.1016/j.tree.2010.11.006 (2011).
4. Costello, M. J. Motivating online publication of data. *BioScience* **59**, 418-427 (2009).
5. Houtkoop, B. L. *et al.* Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science* **1**, 70-85 (2018).
6. Moore, A. J., McPeek, M. A., Rausher, M. D., Rieseberg, L. & Whitlock, M. C. The need for archiving data in evolutionary biology. *J. Evol. Biol.* **23**, 659-660, doi:10.1111/j.1420-9101.2010.01937.x (2010).
7. Berberi, I. & Roche, D. G. Unpublished data. (2020).
8. Fecher, B., Friesike, S. & Hebing, M. What Drives Academic Data Sharing? *PLoS ONE* **10**, e0118053, doi:10.1371/journal.pone.0118053 (2015).
9. Milia, N. *et al.* Mine, yours, ours? Sharing data on human genetic variation. *PLoS One* **7**, e37552, doi:10.1371/journal.pone.0037552 (2012).
10. Piwowar, H. A. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. *PLoS One* **6**, doi:e18657

10.1371/journal.pone.0018657 (2011).

11. Campbell, H. A., Micheli-Campbell, M. A. & Udyawer, V. Early career researchers embrace data sharing. *Trends Ecol. Evol.* **34**, 95-98 (2019).
12. Piwowar, H. A., Vision, T. J. & Whitlock, M. C. Data archiving is a good investment. *Nature* **473**, 285 (2011).
13. Reichman, O., Jones, M. B. & Schildhauer, M. P. Challenges and opportunities of open data in ecology. *Science* **331**, 703-705, doi:10.1126/science.1197962 (2011).
14. Vision, T. J. Open data and the social contract of scientific publishing. *BioScience* **60**, 330-331, doi:10.1525/bio.2010.60.5.2 (2010).

15. Wicherts, J. M. & Bakker, M. Publish (your data) or (let the data) perish! Why not publish your data too? *Intelligence* **40**, 73-76, doi:https://doi.org/10.1016/j.intell.2012.01.004 (2012).
16. Roche, D. G. *et al.* Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol.* **12**, e1001779, doi:10.1371/journal.pbio.1001779 (2014).
17. Hunt, L. T. The life-changing magic of sharing your data. *Nature Human Behaviour*, doi:10.1038/s41562-019-0560-3 (2019).
18. Chawinga, W. D., Zinn, S. J. L. & Research, I. S. Global perspectives of research data sharing: A systematic literature review. (2019).
19. Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K. & McGillivray, B. The citation advantage of linking publications to research data. *PLOS ONE* **15**, e0230416, doi:10.1371/journal.pone.0230416 (2020).
20. Piwowar, H. A. & Vision, T. J. Data reuse and the open data citation advantage. *PeerJ* **1**, e175 (2013).
21. Popkin, G. Data sharing and how it can benefit your scientific career. *Nature* **569** (2019).
22. Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain* **13**, 24, doi:10.1186/s13041-020-0552-2 (2020).
23. Mills, J. A. *et al.* Archiving primary data: solutions for long-term studies. *Trends Ecol. Evol.* **30**, 581-589 (2015).
24. Longo, D. L. & Drazen, J. M. Data sharing. *N. Engl. J. Med.* **374**, 276-277, doi:doi:10.1056/NEJMe1516564 (2016).
25. Edwards, M. A. & Siddhartha, R. Academic Research in the 21st Century: Maintaining Scientific Integrity in a Climate of Perverse Incentives and Hypercompetition. *Environ. Eng. Sci.* **34**, 51-61, doi:10.1089/ees.2016.0223 (2017).
26. Abele-Brehm, A. E., Gollwitzer, M., Steinberg, U. & Schönbrodt, F. D. J. S. P. Attitudes toward Open Science and public data sharing: A survey among members of the German Psychological Society. **50**, 252 (2019).
27. Gewin, V. Data sharing: An open mind on open data. *Nature* **529**, 117-119 (2016).
28. Couture, J. L., Blake, R. E., McDonald, G. & Ward, C. L. A funder-imposed data publication requirement seldom inspired data sharing. *PLOS ONE* **13**, e0199789, doi:10.1371/journal.pone.0199789 (2018).

29. Roche, D. G., Kruuk, L. E., Lanfear, R. & Binning, S. A. Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* **13**, e1002295 (2015).
30. Hardwicke, T. E. *et al.* Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *R. Soc. Open Sci.* **5**, 180448 (2018).
31. Van den Eynden, V. *et al.* Survey of Wellcome researchers and their attitudes to open research. figshare: https://doi.org/10.6084/m9.figshare.4055448.v1. (2016).
32. Tenopir, C. *et al.* Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one* **10**, e0134826 (2015).
33. Tenopir, C. *et al.* Data sharing by scientists: practices and perceptions. *PLoS One* **6**, e21101, doi:10.1371/journal.pone.0021101 (2011).
34. Schmidt, B., Gemeinholzer, B. & Treloar, A. J. P. o. Open data in global environmental research: The Belmont Forum's open data survey. **11**, e0146695 (2016).
35. Digital Science & Figshare. The State of Open Data Report 2019 [online]: Available from https://www.digital-science.com/resources/portfolio-reports/the-state-of-open-data-2019/. (2019).
36. Niederle, M. & Vesterlund, L. Gender and Competition. *Annual Review of Economics* **3**, 601-630, doi:10.1146/annurev-economics-111809-125122 (2011).
37. Hampton, S. E. *et al.* Big data and the future of ecology. *Front Ecol Environ* **11**, 156-162, doi:10.1890/120103 (2013).
38. May, M. Life science technologies: Big biological impacts from big data. *Science* **344**, 1298-1300 (2014).
39. Evans, S. R. Gauging the Purported Costs of Public Data Archiving for Long-Term Population Studies. *PLoS Biol.* **14**, e1002432, doi:10.1371/journal.pbio.1002432 (2016).
40. Weber, N. M., Piwowar, H. A. & Vision, T. J. Evaluating data citation and sharing policies in the environmental sciences. *Proceedings of the American Society for Information Science and Technology* **47**, 1-2 (2010).
41. Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D. & Peres-Neto, P. Ecological data should not be so hard to find and reuse. *Trends Ecol. Evol.* **34**, 494-496, doi:10.1016/j.tree.2019.04.005 (2019).

42. Lin, D. A primer on the certifications of a Trusted Digital Repository (TDR). DataScience@NIH Blog. Avaliable: https://datascience.nih.gov/trusted_digital_repository. (2017).
43. Silvello, G. Theory and practice of data citation. *J. Assoc. Inf. Sci. Technol.* **69**, 6-20 (2018).
44. Lennox, R. J. *et al.* A novel framework to protect animal data in a world of ecosurveillance. *BioScience* **70**, 468-476, doi:10.1093/biosci/biaa035 (2020).
45. Arslan, R. C., Tata, C. S. & Walther, M. P. formr: A study framework allowing for automated feedback generation and complex longitudinal experience sampling studies using R. (version v0.18.3). DOI: 10.5281/zenodo.594223. (2018).
46. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. Fourth Edition edn, (Springer, 2002).
47. Hartig, F. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. R package version 0.3.0. Available: https://CRAN.R-project.org/package=DHARMa. (2020).
48. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: https://www.R-project.org/. (2020).

SUPPLEMENTARY INFORMATION

Individual costs and benefits of sharing open data in ecology and evolution

Sandrine Soeharjono[1] and Dominique G. Roche[1,2,3*]

[1] *Département de science biologiques, Université de Montréal, Montréal, Canada*

[2] *Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland*

[3] *Department of Biology, Carleton University, Ottawa, Canada*

Table S1. Comprehensive list of answers in the free-form sections of the questionnaires.

Over the last 5 years, approximately what percentage of your published (empirical) papers have associated open data?

- I would say 75-90%
- This would be greater if there weren't so many demands on my time
- The majority of the papers that I have published have not required the data to be made open, but in many of the publications, much of my data have been published in tables.
- This refers to entire datasets. Some papers (not counted above) also may have publicly available components (e.g., DNA sequences in GENBANK).
- A lot of the datasets associated with recent papers include confidential household survey data, where we stipulated in our ethics certification that we would not share the data. (no more space to elaborate)
- My data are all open archived. I have not used data collected by others
- We're happy to share any data from published papers upon request. However, little of the data is made publicly available.
- I am now retired and have little in the way of research funds. I still publish regularly but cannot afford to venture into open access publishing because the funds would have to be paid personally.
- 100% from my lab; My colleagues with whom I sometimes publish are not so keen.
- It has been standard in genomics research for over a decade.
- 90% is more accurate.
- I mostly work with other researchers' data sets, so I do not control whether I can make the data available.
- We upload associated data sets when requested or required. Not all journals provide this opportunity.
- I have not deposited data in any specialize data bank, but I have shared my data with everyone who asked for it
- I mainly use publicly available data. I do not re-post the data.
- Where is the limit between open data and data in Supplementary info? I put a lot of data in SI, without putting all the raw data of the paper... So 0% for "real raw open data" but close to 100% for SI

- It was not clear to me what you meant by "open data". My answer concerns data that were publicly available and not data that became public because they were included in an appendix of our papers.
- Actually 75-100% (i.e. over 75%) but that category not available.
- Probably more like 75-100%, but that wasn't listed as an option.
- This is rapidly increasing and in past year, most of them have had open data

Are you generally supportive of mandatory open data policies by journals and/or funding agencies? These typically require that authors publicly share the data necessary to reproduce the results/figures presented in a scientific paper.

- Field based studies and high levels of taxonomic work are not given credit when others reuse an available dataset.
- I think that anyone who wants to use data should simply contact the author.
- with publication embargoes and in some cases data sharing arrangements (i don't usually use these but can easily envision when they would be important)
- Some data need to be kept secure.
- In principle I'm OK with it, but it does add more complexity and effort to the publication process. I am also somewhat concerned about losing control of my own data.
- Except that it may change the types of data that we collect or where we publish as some data are confidential and cannot be shared.
- Data from any publication should be available upon request; it is not necessary to publicly post it all.
- I am happy with a subset of data or a clause stating that I'd be happy to share the data on demand. I do not believe in blanket access because people would be able to do other things with the study.
- NSERC grant amounts have not increased with the page charge amounts now required.
- The big publishers are making lots of money at the expense of researchers and grant funds. Scientists need to take back this territory and share information freely (not using for-profit publishers).
- I'm hoping you'll get to this in the survey, but just in case not: open source data can be discriminatory. Given the bias in grants etc.... forcing all to make data open would advantage white men; if data are freely available and white men have faster times to publishing, more grants, more resources, we will simply be providing more publishing, more grants to white men.
- What is "data" is a key problem. Much of what is 'shared" or expected to be shared is metadata, i.e., interpreted. It is NOT raw data.
- Yes, in theory. In practice, it is much more difficult, posting a spreadsheet of observations is easier, but less useful than posting the original underwater video.
- My issues concern 1) timelines - often not enough time to publish on collected data, and 2) who gets recognition for the data; however, I share openly on collaborations.
- Access to data for meta-analyses, etc., is often contingent on permission from the data owner. As the data user, I do not always have permission to distributed data further.
- I agree with the policies in general, but we need help in order to make the data fully available. An Excel file from a M.Sc. or a Ph.D. student is not really useful to anybody else.

- Yes, after some kind of embargo period to let the data gathers publish their research first.

Do you believe that publicly sharing non-sensitive research data is beneficial to society? Non-sensitive data are data that protect the identity of study participants and do not put organisms at risk of exploitation or disturbance.

- These data are sometimes used for meta-analyses from which the data are imperfect
- Depends on the situation
- Data can be available upon request, without being in a public database
- I don't believe yes or no is an appropriate answer here. Your survey forced me to answer this question.
- Such data can be shared openly without fees being involved. Scientists have lost control of their valuable and precious resources.
- It's beneficial, but for the most part, nobody cares. Open data are mainly used for data mining and meta-analyses.
- Having to publish data IMMEDIATELY when a paper comes out isn't beneficial. Many other projects can usually be done with a given Dataset and we give away that ability.
- We live in an era when facts, and data, no longer seem to matter in public discourse. To push for evidence-based decision-making, we should publish our data.
- I think it is mainly useful to other researchers.
- In principle yes but it carries risks of relying on data that may be faulty or one does not understand.
- Only if it is used correctly and includes all the metadata necessary to understand how the data were collected and what the limitations are.

Do you tend to share your data publicly even if open data are not required by the journal in which you have published?

- I have always shared my data with government agencies and any researcher who request them.
- Much more recently. Before I may not have taken the time.
- I have shared data with researchers without requiring authorship/collaboration, but not associated with a journal. I am not opposed but have not bothered where it hasn't been required.
- Usually components (e.g. gene sequences) rather than entire datasets.
- Always. There is no good reason not to share data.
- It's a lot of work to make data ready to be shared publicly. But even if we don't upload to a public repository, often we share privately when requested
- much of my previous work has been for a sponsoring First Nation, often with data protocols in place
- I share freely with anyone who asks.
- Every journal I publish in requires open data, so this question is moot for me.
- For Genomics work always. For others, on occasion.
- If it is my data, yes.
- I have shared data with anyone who has asked for them
- ArcticNet, for example, requires metadata entries into the Polar Data Catalogue.

- Almost always when my lab generated the data.

Have you experienced positive outcomes (benefits) from sharing your data publicly? Examples include a personal sense of satisfaction, greater efficiency from adopting better data management practices, new collaborations, awards, accolades, etc.

- I suppose I do get some satisfaction, but it's pretty minor.
- Just the good feelings associated with sharing, and being a generous person, which I how I like to view myself.
- I have not done it.
- I like having the data archived somewhere other than my computer or university.
- and I have been royally screwed by it too. I personally am job-secure and hyper collaborative. I CHOOSE to share my data. Forcing is another thing entirely.
- Specifically, a positive assessment on my NSERC Discovery grant whereby several reviewers commended me for my data sharing history and commitment to open science policies.
- Part of my cv for applying for high-level position that involves leadership in all things data.
- Not aware of any benefits
- It’s not something I have thought about.
- When others have used data that my lab generated, that produces added value for effort (and the public funding) that went into producing the data.
- They have been used for teaching in other universities (students or profs wrote to me).
- Being invited to co-author papers on open access initiatives.
- More hits on our website where we are working to inform policymakers.
- I have generally shared my data when requested, but no nothing particular stands out as positive or negative about the process.

Have you experienced negative outcomes (costs) from sharing your data publicly? Examples include being 'scooped' on a paper, having your data misused or misinterpreted by others, experiencing public shaming from someone identifying errors in your data, analyses or conclusions, etc.

- A govt collaborator tried to renegotiate on his agreement to share the data.
- Not yet
- One journal in our field mandates open data PRIOR to acceptance, hence referees have access to data for a paper that has not been accepted. This is unacceptable.
- It took over a year to find an archive that would take my data, which delayed publication by 1 year. Archives able to handle large datasets of this type generally do not exist.
- Time cost in getting data ready to share - I don't mind doing it, but its wasted if no one uses the data.
- Publicly sharing data in a format that is useful for others is extremely time consuming
- It is relatively time consuming to make data publicly available in a useful way and I sense that they are very rarely used.
- I have not done it.

- I spent considerable time going back to old disks and computers to find the original data and update the software so that it could be shared and I have no idea whether it was ever used.
- No, because I share with co-authors/collaborators and get recognition for the data collection.
- On more than one occasion, I have given a data to another researcher who was intent on proving that my conclusions were incorrect. I don't believe that I have suffered any harm from so doing.
- Mistaken concerns raised with journal from researchers that we had shared their data (we had only shared data we had gathered). Large time cost of putting data online for some journals.
- Data error found but corrigendum submitted.
- We are careful about when and what we release.

Do you believe that mandatory open data requirements by journals / funders impose excessive time demands on researchers?

- Not excessive, just unnecessary.
- That will depend. Some of the papers recently are very large geodatabases
- but it does require 1-2 full days to clean and annotate data and associated R scripts for publication, so maybe not excessive but definitely significant.
- Some journals have clunky and uncertain processes, especially around sensitive data.
- I expect this to be the case, hence the reason I have not done it so far.
- somewhat
- I think it is valuable to share data, but certainly takes more time.
- While it is a huge time demand, the result has been worth it.
- These questions would be better with a Likert scale, they are not well-suited to yes/no answers.
- I don’t think journals should require it; I think funding agencies should require it.
- There are circumstances where authors may not be able to provide these data. *Mandatory* open access (like mandatory anything) is rather silly and forgets the human dimensions of scientific research.
- It is incredible to me that scientists do all of the work, including reviewing) and ultimately must pay for the privilege of making their information available.
- My datasets are relatively simple. My colleagues that publish huge datasets have complained some about open data requirements
- I can only speak for my own situation and the situation of close colleagues who, like me, mostly collect simple small datasets.
- It's not about time, it's about equity.
- And costs. It is expensive to collect data and metadata. Free access to it presumes others without resources to do so can now conduct research free of charge and the effort of grant writing, etc.
- There is no question that this takes more time in the short term. However, in the long term, there are obviously benefits to having the data archived and well described by metadata.
- if the expectation from the start is that data will be shared at the end, then the time spent to share goes down because good data curation and archive practices are adopted early on.
- I would say "possibly". Because I've never done open data, I don't know what the time requirements would look like.

- If you manage your own data properly, you have the data ready for a public archive like GenBank or Dryad.
- We need to find an easy way to do it.
- it will require extra work, hard to tell if this is excessive
- Especially when computer codes are required - can get very complicated to share.
- Not all labs have qualified personnel for data management. Resources should be dedicated to this in grants and institutions.
- Yes and no; Yes, because funding agencies want it yet they do not provide the extra funds to make it happen. Sharing data costs time and money.
- Although some journals cannot house large datasets.
- Not excessive, but the time commitment is non-trivial.
- Getting datasets in a form that someone else can read and understand is very time-consuming.

Do you believe that the time requirements and effort needed to share your research data have decreased over time? For example, as a result of prior experiences sharing data and/or improved data management practices.

- Each data submission is de novo.
- It's new enough that it's a bit difficult to say. I do think the system is still in flux.
- I have shared my data on the side, by choice.
- I now have arrangements with at least 2 archives that will take large datasets of this type.
- I answered no because I've never done data sharing....
- It has become longer and more difficult with sequencing data. GenBank used to be a pain (in the 90's), now it's easy.
- better organization of my data and better means of sharing it.
- It has not increased either. An honest answer would have been: "indifferent".
- Our work becomes more complex with every study, so there is little prior experience.

What measures or changes would likely increase your participation in open data? These could be implemented by journals, funding agencies and/or universities. This question is optional.

- in open access journals, reduced publishing fees. Unique DOIs and citable references for data independent of the initial manuscript that used it.
- Authorship agreements; payment (e.g. free colour figures or open access fees waived).
- Option for allowing embargo periods before open data is released that extend for up to 3 years following publication to allow time to publish additional papers before getting scooped
- I really wish they wouldn't ask for it until acceptance - most journals don't but some do. You can embargo it but I dislike this practice and it discourages me from publishing in some journals that do this.
- More journals requiring it, greater value in tenure or annual performance reviews.
- It will likely require mandatory submission of underlying data with each paper, as while I can see the value to having data broadly accessible, when given the choice, I'll move on to some other task every time.
- When hell freezes over.

- Reduce costs
- I would like a data embargo or requirement that people inform original authors and /or discuss further analysis with the original authors
- If we get additional credit for having them as open data.
- credit from NSERC or other funding/granting/promotion agencies. right now, it's on a volunteer basis and i do it mostly for my future students so i can easily refer them to past data and analyses
- Journals paying for archive.
  Clearer and safer (for sensitive data) rules.
- If journals and universities supported it financially, by paying for access to databases like Dryad.
- It becomes mandatory.
- Any measures or changes that make open data sharing easier (e.g., journal partnerships with dryad, partnerships that allow datasets to be shared during the review process). I would also like improved checks on datasets that are shared - too often researchers share summary or incomplete data that could not be used to check/confirm their results.
- Make it mandatory; possibly with a sunset period under specific circumstances
- More clearly getting some sort of publication credit would help.
  Also better guarantees that the data won't be used to scoop me!
  Also, as indicated above, the system is in flux. More confidence that the data will remain available would be nice.
- If it becomes mandatory.
- More credit for all the work one invests in collecting the data.
- Nothing. All data produced in my lab goes on-line at submission of first paper to use the data. All data are paid for by the public and therefore should be available to anyone. We must force everyone to comply as older generations are too slow to change. I am particularly shocked at how difficult it is to obtain data from federal and provincial government agencies. Many people in these agencies also have stupid requirements to be added as an author on any paper that uses data collected in their department. Such practices should be made public, so the public knows how their tax money is being mis-used.
- Better funding for the existence and maintenance for archives is essential. Many funding agencies and journals require data archiving, but they do not provide the archives, and archives for large and complex data sets either don't exist in some field or are restricted (e.g. to certain agencies).
- Small grants to organize data files for public access.
- Requirements from funding agencies.
- I'd be happy to share my data with individuals who contact me and send me a proposal of what they'd like to do with it. If it is a substantial amount of data or conflicting with my own interests, I should have the control over access of my data and data sharing
- my participation is reluctantly already very high. It is typically required.
- Funding (e.g., and additional 10% from an NSERC DG) explicitly linked to data archiving.
  Departmental / university support for this type of additional work that honestly tends to come at the very end of the research process.
  Training for students on how to manage and curate their own data
  Equivalent training for PIs
  Less haughty dogmatism regarding "open data" (a tricky one, I know)

- Scientists getting together to create their own outlets without the intervention of big corporations.
- More evidence that open data are used by others; so far it seems more of a philosophical than real advantage, at least for my type of research which relates to conservation biology and human-wildlife conflict
- If my university mandated that all data were open, my participation in open data would increase. Otherwise, I make all data in from published papers available on Dryad.
- If it was made mandatory
- Resources to support Open Science and open data initiatives and requirements are still inadequate. Policies still need to be developed to allow for full adoption (with few exceptions) for open data requirements. This should include more training for undergraduate and graduate students too.
- The entire system needs changing before I will actively support a mandatory open data policy.
  We need to remove bias in all things that lead to publishing success before I would allow white men to have access to more data. I know that this sounds nasty but OMG I am so tired of 'innovating and collaborating' without thinking of who we are really serving.
- There is a cost (mostly time) associated to publishing the data and to maintain it afterwards (e.g. answering questions about the data) that currently no one is paying for except the researcher. This cost is tangible, but the benefits are a bit intangible, yes I believe sharing data (mostly expertise in fact, but data often comes with it) has led to collaborations, but it is hard to quantify; the time spent preparing it, publishing it, learning the tools required to do so and then answering question is real
- Everything sensible is expensive. First, open access and open data currently has not financial support - journals charge +5K US to publish open access and are now considering storage charges for big data sets. Where does this money come from in a fixed funding envelope? Secondly, what is data? Raw data, access to the same samples, my interpreted data, etc.? This needs to be clarified before we determine what should be shared. For example, I work on vertebrate morphology. Is the raw data the fossil skull, the uCT scan raw segments of the skull, the "cleaned up" segments, or the skull reconstruction? I would argue the raw data is the actual skull. That is always accessible at a museum. The terabyte cost of storing for open access, the raw segmented data is enormous. Who bears this expense? There are many, many unanswered questions.
- make it as easy as possible to do so
- standardised ranking of papers by journals indicating whether all data are freely accessible and associated with the paper, available from other data sources, available on request, not available. These could be reported to funding agencies.
- ready access
- NA - I already share all published data.
- Funding agencies could evaluate researchers on data openness (just as NSERC evaluate researchers on HQP and now EDI)
- Greater recognition (by funding bodies, university administration) of archived data as a scientific contribution in its own right beyond the contribution made by the published article.
- Implementation by journals
- Free archiving

- Funding needed!!
- If government funding agencies would require, check, and provide resources to make data open, it would provide the best incentive.
- One single, optimized data sharing platform supported by my university.
- Let us publish our data 7 years after the paper is out. It lets the owners retain the ability to make the most out of their own data before sharing it.
- Explicit instructions for what metadata to include; ability to define limits on how data would be used or how data would be acknowledged/credited; best practices for how to deal with data obtained from other sources (not collected de novo); recognition that long term monitoring data may "belong" to other entities besides the authors
- More journals linking to Dryad and paying the $100 fee. Recognition every time your open data is used - not a co-authorship but some other indicator that would be part of cv; A specific section in grant applications to list open data repositories; Grant agencies requiring proof that any research-supported data has been archived; Embargo policies of repositories that allow researchers conducting long-term studies to meet data archiving requirements without costs to their careers.
- The society of which I am president has adopted a policy of providing free uploading to Dryad for data used in our publications (The Auk, and The Condor).
- Implementing stronger policy on data sharing, particularly citation of the source, and making this compulsory across the board and for everyone. I believe this would make everything more transparent and will facilitate timely publications.
- If we could reduce the payment for open-sourcing, I believe more scientists would participate. Not everyone has access to large grants that allow for the ability to publish open source. This is something that might be useful for NSERC to offset.
- The platforms for publicly posting data are not free, but I am very reluctant to post my data on a site run or administered by a for-profit journal. Having publicly funded data archives would help.
- Co-authorship requirements that recognize the work of the data "owner" in terms of obtaining grants to collect the data, as well as those who actually collect the data.
- make it mandatory and always offer embargo for PhD data
- We need help from our institutions, to find formats more useful to users outside the laboratory and we need workshop to teach students how to set up data files properly.
- Perhaps journals could be intermediaries, certainly not the government as they inevitably impose rigid procedures, protocols, and other useless hindrances
- as a requirement for publishing
- As much latitude as possible on the format of uploaded data...
- mandatory by all journals
- I am concerned about making datasets available when I plan to do more analyses on the data. Or when I may think of new ways to analyze the data later.
- Free technical support by a data manager and free training of students in data management
- Implement a single, standardized open data reporting standard
- Institutional support for data management, and cooperation by students to actually provide data in a form that could be shared
- (1) Funding agencies to use that as selection criterion for funding. (2) Funding agencies to provide researchers with extra $ for researchers to make it happen (because sharing data

costs time and money).

- More support from journals for speciality datasets. Journals should implement data checking processes - this would benefit field.
- clean guidelines about authorship - when is using the data crucial to a publication and should it be acknowledged?
- Data systems that collate similar types of data are key. Short-term research projects once published should all be placed in collaborative data systems of similar data types. But there are programs that are long-term for which the questions asked will evolve over time. Those datasets should not be required to be public because they are the heart and soul of a researcher's program.
- Making data repositories more accessible and training students to organize and store raw data correctly.
- Give some flexibility as to how much is shared.

Table S2. Questionnaire

---

1. When (in what year) did you obtain your PhD?

A box will be provided for respondents to write their response.

---

2. What is your gender? (please skip this question if you do not identify as male or female).

Note that an email was sent subsequently to the study participants apologizing for not explicitly including an option to identify as gender non-binary and to indicate "non-binary" in the optional free-form section of this question, if applicable.

- ○ Female
- ○ Male

---

3. Have you ever publicly shared (archived) 'raw' data associated with a published study (as supplementary material or in an online data repository)? 'Raw' data correspond to data used in the statistical analyses or underlying the results (e.g. figures) presented in a published study.

- ○ Yes
- ○ No

---

4. Over the last 5 years, approximately what percentage of your published (empirical) papers have associated open data?

- ○ < 10 %
- ○ < 25%
- ○ 25-50 %
- ○ 50-75 %

- ○ 75-99%
- ○ 100%

5. Are you generally supportive of mandatory open data policies by journals and/or funding agencies? These policies typically require that authors publicly share the data necessary to reproduce the results and figures presented in a scientific paper.

- ○ Yes
- ○ No
- ○ I am indifferent

6. Do you believe that publicly sharing non-sensitive research data is beneficial for society? Non-sensitive data are data that protect the identity of study participants and do not put organisms at risk of exploitation or disturbance.

- ○ Yes
- ○ No

7. Do you tend to share your data publicly even if open data are not required by the journal in which you have published?

- ○ Almost always
- ○ On occasion
- ○ Never

8. Have you experienced positive outcomes (benefits) from sharing your data publicly? Examples of benefits include a personal sense of satisfaction, greater efficiency from adopting better data management practices, new collaborations, awards, accolades, etc.

- ○ Yes

- ○ No

If the respondent answers no, the next question is skipped.

---

9. What benefits have you experienced?

- ○ Personal sense of satisfaction from having shared/published a research product other than a scientific paper
- ○ Greater efficiency from adopting better data management practices when conducting your research (e.g., when having to revise analyses, send your data to collaborators, or share data as a condition of publication)
- ○ New collaborations with researchers having contacted you because of your shared data (whether to reuse these data or not)
- ○ Offers of co-authorship from other researchers wanting to re-use or re-analyse your data
- ○ Awards or a promotion (e.g. from a journal, society, government or academic institution such as a university)
- ○ Accolades (e.g. praise from colleagues, your dean, head of department, or on social media)
- ○ Positive reviews on a grant, fellowship or journal submission
- ○ Other (please specify)

---

10. Have you experienced negative outcomes (costs) from sharing your data publicly?
Examples of costs include, being 'scooped' on a paper, having your data misused or misinterpreted by others, experiencing public shaming from someone identifying errors in your data, analyses or conclusions, etc.

- ○ Yes
- ○ No

If the respondent answers no, the next question is skipped.

11. What costs have you experienced?

- ○ Frustration from having to share data that you would have preferred to keep private
- ○ Being ‘scooped’ on a paper (i.e. a third party reused your data and published research that you had intended to carry out before you had time to do so)
- ○ Your data were misused or misinterpreted by a third party (for e.g., your data are used to address a research question for which your data are not appropriate)
- ○ Your data put humans or other species at risk of privacy infringement, disturbance or exploitation
- ○ You were not offered co-authorship on a study reusing your data and you believe that co-authorship was warranted
- ○ You experienced public shaming or reputational damage from a third party identifying errors in your work as a result of open data
- ○ Other (please specify)

12. Do you believe that mandatory open data requirements by journals/funders impose excessive time demands on researchers?

- ○ Yes
- ○ No

13. Do you believe that the time requirements and effort needed to share your research data have decreased over time (for e.g., as a result of prior experiences sharing data and/or improved data management practices)?

- ○ Yes
- ○ No

14. What measures or changes (by journals, funding agencies and/or universities) would likely increase your participation in open data? (This question is optional)

A box will be provided for respondents to write their response.

Debrief and informed consent

Thank you for your participation in this survey. Our research aims to evaluate the prevalence of potential costs and benefits of open data in the field of ecology and evolution. In addition, some of your answers will assist us in determining whether costs/benefits differ among men and women as well as among researchers at different career stages. Answers to the last (optional) question will help us formulate recommendations for improving editorial and funder policies on open data.

We intend to publish a peer-reviewed paper based on the information that you and other respondents have provided. The confidentiality of your personal data will be guaranteed by the removal of all identifying information. You may withdraw your answers from being included in this research at any time until December 15, 2019 by e-mailing sandrine.soeharjono@umontreal.ca

If you agree to these terms, your answers will be recorded. If you do not agree to these terms, your answers will be deleted.

Do you consent to the use of your anonymized responses in our study?

- ○ I consent
- ○ I do not consent

End of Questionnaire