Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution

Dominique G Roche[1,2,3], Ilias Berberi[1], Fares Dhane[2], Félix Lauzon[2,4], Sandrine Soeharjono[2], Roslyn Dakin[1], Sandra A Binning[2]

[1] *Department of Biology, Carleton University, Ottawa, Canada, K1S 5B6*
[2] *Département de science biologiques, Université de Montréal, Montréal, Canada, H3C 3J7*
[3] *Institut de Biologie, Université de Neuchâtel, Neuchâtel, Switzerland; 2000*
[4] *Department of Biology, McGill University, Montréal, Canada, H3A 1B1*

Correspondence to: Dominique G. Roche
209 Nesbit Biology Building, Carleton University
1125 Colonel By Drive
Ottawa, Ontario, Canada, K1S 5B6
E: dominique.roche@mail.mcgill.ca
T: 1-438-889-5822

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

Abstract

Many leading journals in evolution and ecology now mandate open data upon publication. Yet, there is very little oversight to ensure the completeness and reusability of archived datasets, and we currently have a poor understanding of the factors associated with high quality (FAIR) data-sharing. We assessed 362 open datasets linked to first- or senior-authored papers published by 100 principal investigators (PIs) in the fields of evolution and ecology over a period of seven years to identify predictors of data completeness and reusability ('data archiving quality'). Datasets scored low on these metrics: 56.4% were complete and 45.9% were reusable. Data reusability, but not completeness, was slightly higher for more recently archived datasets and PIs with less seniority. Journal open data policy, PI gender, and PI corresponding author status were unrelated to data archiving quality. However, PI identity explained a large proportion of the variance in data completeness (27.8%) and reusability (22.0%), indicating consistent inter-individual differences in data sharing practices by PIs across time and contexts. Several PIs consistently shared data of either high or low archiving quality, but most PIs were inconsistent in how well they shared. One explanation for the high intra-individual variation we observed is that PIs often conduct research through students and post-docs, who may be responsible for the data collection, curation and archiving. Levels of data literacy vary among trainees and PIs may not regularly perform quality control over archived files. Our findings suggests that research data management training and culture within a PI's group are likely to be more important determinants of data archiving quality than other factors such as a journal's open data policy. Greater incentives and training for individual researchers at all career stages could improve data sharing practices and enhance data transparency and reusability.

Main text

The debate is over regarding the value of open and FAIR data (Findable, Accessible, Interoperable, and Reusable data [1]). Many journals, funding agencies, and policymakers agree that the societal benefits of publicly sharing (non-sensitive) research data outweigh any perceived or reported costs to individual researchers [2-6]. Making data underlying scientific studies publicly available facilitates exploring, validating, and building on published results [7], with few researchers in evolution and ecology (E&E) reporting negative outcomes from sharing [2]. Not only do open data accelerate scientific discovery, as illustrated during the Covid-19 pandemic [8], but sharing data also promotes an academic value system that is more equitable, diverse, and inclusive [9-11].

Some researchers are reluctant to share the data underlying their published results [6, 12, 13], yet most researchers view open data positively [2, 14, 15]. Since 2010, when a handful of journals began requiring open data in E&E [16], policies encouraging this practice have grown rapidly. Now, 20 percent of journals publishing research in E&E mandate open data [17]. Strong journal policies are effective at ensuring that more datasets are shared [18-20], which is often touted as a win for open science [21]. Yet, problems persist [19, 22, 23]. For instance, more than half of open datasets associated with 100 E&E studies published in 2012 and 2013 were incomplete and/or archived in ways that prevented reuse [24]. Similar issues have been documented in psychology [25] and cognition research [26], pointing to the inherent problem with journals mandating open data without appropriate oversight or quality control [27-29]: datasets get archived but the majority are incomplete and challenging to reuse. Developing effective strategies to promote good data sharing practices requires that we first identify which factors are associated with complete and reusable open data [30].

We assessed the archiving quality (completeness and reusability) of open datasets associated with publications by tenured or tenure-track E&E faculty members (PIs) in biology departments at the 21 highest-ranked universities in Canada. PIs were necessarily first or last author on the publications assessed (see pre-registered methods). Data completeness (availability of data allowing computational reproducibility) and reusability (ease with which data can be reused by third parties) were assessed following Roche et al. [24]. Scores above 3 on two 5-point scales indicate complete or reusable data (Table 1). In total, we examined 362 datasets shared by 44 women and 56 men (Table S1). We ran a Bayesian bivariate mixed-effects linear model with PI identity and their institution of employment as random factors. We tested whether article publication date, journal open data policy, and the seniority, gender and corresponding author status of PIs predicted the archiving quality of their open datasets. A post-hoc exploratory analysis was also carried out to examine the relationship between the complexity of datasets (estimated as the number and size of archived data files) and their completeness and reusability.

The completeness and reusability scores of datasets varied considerably within and among PIs (Fig. 1). Overall, 56.4% of datasets were complete (mean completeness score of 3.4 ± 1.3 SD), and 45.9% were reusable (mean reusability score of 3.1 ± 1.4 SD) (Fig. S1) This represents a moderate improvement of approximately 10% above the completeness and reusability of datasets associated with E&E studies published in 2012 and 2013 [24]. Data completeness and reusability were strongly correlated within (R^2=0.79, 95% CI: 072.–0.85) and among (R^2=0.77, 95% CI: 0.56–0.92) PIs (see [31] for an explanation of among and within individual correlations).

Open data is a relatively recent concept in E&E, having been introduced in earnest a decade ago [5, 32]. As such, PIs who developed their research skills prior to this period may be less likely to have incorporated these principles into their research workflow. We assessed datasets as far back as 2013 in our analysis but included faculty members hired as recently as 2019, using year of first scientific publication as a proxy for PI seniority. Therefore, our study likely includes datasets published by new PIs during their PhD and post-doc years, when they might have had access to various training opportunities. For instance, a growing number of biology departments recognize the value of data science and initiatives such as Data Carpentry (datacarpentry.org) and FOSTER (fosteropenscience.eu) now routinely offer workshops in data management across North America and Europe. We found that PIs with less seniority tended to share slightly more reusable data than PIs with more seniority, suggesting that early training initiatives may be bearing fruit (Figs 2, S2 H). This result is good news because younger researchers tend to be more fearful and reluctant to share their data than senior researchers [33] despite reporting a more favorable attitude towards open data [33-36] [but see 37]. Our study included datasets spanning seven years (2013-2019). We found that datasets associated with more recent studies were slightly more reusable than those of older studies (Figs 2, S2 J). In contrast, dataset completeness was independent of PI seniority and publication date (Fig. 2, S2 G,I). These results suggest that, while there have been slight improvements to data sharing practices through time, these are slow to change. Training and increased exposure to open science practices are no doubt contributing to this slow improvement, but additional work is needed at all career stages to enhance data archiving quality.

We examined whether data archiving quality was influenced by PIs being corresponding author on the published article (in addition to being first or senior author). We took PI corresponding author status as an indicator that the PIs themselves archived the data. Assuming that PIs are highly competent at managing their research data, we expected open datasets with PIs as corresponding author to be of higher archiving quality, on average, than those archived by presumably less experienced researchers (likely students or post-docs). We found no support for this hypothesis: data archiving quality was unrelated to corresponding author status (Figs 2, S2 E,F). Conventions regarding who is corresponding author on a published study may vary among sub-disciplines and research labs. However, the corresponding author is ultimately responsible for compliance with journal policies, including open data [38]. The fact that corresponding author status had no bearing on data archiving quality is worrying and suggests that PIs do not understand the responsibilities associated with this role, do not have the tools or training to ensure that open data are compliant with journal policies, or simply do not care. Education, capacity-building, and incentives targeted at individuals are needed to address these issues [2, 7, 39, 40].

We had no a priori hypothesis for why PI gender might influence data archiving quality. However, we included this predictor in the model because we believe that gender differences are important to consider in scientific research. We found no evidence to suggest that PI gender influences the quality of open data (Figs 2, S2 A,B). This result is encouraging given that men in E&E self-identify as experiencing more costs than women as a result of sharing open data [2]. Women accounted for almost half of the PIs assessed in our study, yet far fewer than 50% of PIs in biology departments at Canadian institutions identify as women [41].

When journals have a mandatory open data policy, the number of archived datasets underlying published research articles increases [18]. We tested whether such policies also translate into more complete and reusable data. We hypothesized that data archiving quality would be higher for studies published in journals requiring open data. Alternatively, it is also possible that researchers voluntarily archiving datasets in journals without a policy share higher quality data than researchers who are forced to do so. Contrary to this logic, a journal’s open data policy had no bearing on data archiving quality (Fig. 2, S2 C,D), indicating that policies alone do little to ensure that shared data are complete and reusable. Some of the world's largest funding agencies (e.g., ERC, NSF, NERC, Canadian Tri-Council) now require that PIs specify data management and/or sharing plans to obtain funding. However, compliance with these policies is low [22]. Unless resources are invested in training, technical support and policy oversight [2, 15], data risk not being archived or not contributing to advancing knowledge in instances where they are made available.

We assessed multiple datasets by the same researchers, which allowed us to calculate repeatability scores for both data completeness and reusability. Repeatability (R) ranges between 0 and 1 and is the proportion of the total variance in scores attributed to among (or inter-) individual differences: high R values indicate large score differences among individuals and consistent scores within individuals [42]. Data completeness was moderately repeatable with $R_{adj} = 0.28$ (95% CI: 0.16–0.39), and data reuse with $R_{adj} = 0.22$ (95% CI: 0.14–0.34), revealing differences within and among PIs (Fig. 1). This variability reflects several realities that merit discussion. On the one hand, PIs in academia often conduct research through students and post-docs, who may ultimately be responsible for data collection, curation and archiving. Thus, in some cases, the PI may not have performed quality control over the archived files, potentially explaining the considerable within-individual variation in data completeness and reusability scores we observed (variable scores in Fig. 1). In these cases, data archiving quality may be a better reflection of data management not by the PI, but by the person within the PI's research group who was responsible for archiving the dataset. On the other hand, some PIs consistently scored high or low in both data completeness and reuse (low and high scores in Fig. 1). This consistency within research groups suggests both a robust lab culture promoting good research data management or, alternatively, a PI's lack of competence or reluctance to engage in data sharing and student training in this regard. PIs who oppose open data initiatives (e.g., [2, 12]) are unlikely to respond positively to incentives or training opportunities to improve data archiving quality and FAIRness. However, our results suggest that only a minority of PIs potentially fall within this category (approximately 10%; Fig. 1). Rather, most PIs were inconsistent in how they shared data associated with their publications, or consistently shared highly complete and reusable data. We found no indication that dataset complexity influenced archiving quality, suggesting that PI often struggle to share even simple datasets (Fig. S4). These finding points to the importance of facilitating sound research data management practices within research groups to achieve high-quality, FAIR data sharing.

Overall, our data suggest that journal policies are ineffective at ensuring that open data in E&E are complete and reusable. We also found that data archiving quality is slow to improve over time. However, most PIs did share high-quality open data, either consistently or occasionally. Striking variation in data archiving quality within PIs suggests that education, training, and technical support could help raise the bar by enabling good data sharing practices to become the rule rather than the exception [39].

Materials and Methods

Our methods were pre-registered at https://doi.org/10.17605/OSF.IO/A492M. We assessed open datasets from research faculty members in biology departments at the 21 highest-ranked Canadian universities based on the 2019 Times Higher Education World University Rankings. Although we initially planned to select the top 20 Canadian universities, we retained 21 universities due to a three-way tie for rank 19. Our study focused on Canadian academic faculty. However, our findings are likely to be representative of the broader population of PIs in E&E given Canada's diverse academic institutions as well as the high degree of PI mobility in today's globalized academic landscape. Furthermore, many granting agencies in Europe and the USA require that data from funded research be publicly archived within a specified timeframe of publishing. This is not yet the case in Canada: the Tri-council Granting agencies now require a data management plan for grants submitted in 2021 and beyond, but this does not yet include a requirement for open data. This allowed us to assess the effect of journal policies on archiving practices by Canadian PI's independent of requirements from funding agencies on the same practices.

We reviewed the biology department website at each of the 21 selected universities in a random order and identified all researchers primarily conducting research in the fields of ecology and/or evolution (E&E) with a rank of assistant, associate or full professor. Adjunct professors and researchers who primarily focus on molecular biology, genetics, genomics, bioinformatics, theoretical biology, comparative physiology and paleontology were excluded given our focus on researchers in E&E. Each researcher’s primary fields of study were determined from public information on the university websites and cross-checked by a minimum of two people (IB, FD, SAB, RD, DGR). This criterion served to limit the scope of the study to E&E and facilitate consistent assessment of datasets, given our shared expertise. To standardize the selection of researchers across universities and avoid bias, we omitted E&E researchers who are primarily affiliated with departments other than biology (e.g., environmental sciences, natural resources, fisheries and ocean sciences, veterinary sciences). In total, we identified 351 researchers that met these criteria Table S1).

To be included in our study, candidate researchers had to have at least two articles containing a data availability statement that were published in a peer-reviewed scientific journal between January 2013 and June 2019. The researcher had to be first or senior (last) author on these articles, ensuring that they were one of the primary intellectual contributors (in E&E, the convention is that the first and last author are primarily responsible for the work). We used Google Scholar and the researchers’ personal and/or institutional websites to identify articles. When researchers did not have a Google Scholar profile, we verified their publication list using Web of Science. Researchers at each university were screened in a random order. Articles for each researcher were manually searched in a reverse chronological order (i.e., starting in 2019, ending in 2013) to determine whether a data availability statement was present, either stated explicitly at the end of the article, or embedded in the main text. If there was an absence of an open data statement but presence of electronic supplementary material (ESM), we looked for evidence of open data in the ESM (i.e., raw or processed data as opposed to summary statistics). Reviews, commentaries, and theoretical or simulation studies were excluded. The article search for every researcher was completed when seven articles containing a data availability statement or open data were identified, or when the reverse-chronological scan reached January 2013. In total, 4,322 articles were examined, 928 of which contained a data availability statement and/or associated open data.

The strength of a journal’s open data policy and date of implementation was determined by reviewing each journal’s author guidelines and relevant editorials. When necessary, we contacted journal editors for clear information on whether open data were required (i.e., mandatory open data) or encouraged (i.e., optional open data) as a condition of publication at the time a paper was published. Journals without an open data policy were categorized as optional open data.

We identified 194 researchers with at least two articles containing a data availability statement and/or open data across the 21 universities (Table S1). Of these, we randomly selected up to four women and four men at each university to evaluate the data archiving quality of their open datasets. We aimed to randomly select three women and three men at each university but some universities had fewer than three researchers per gender (Table S1). The departments of biology at two institutions had no researchers that met our selection criteria. One researcher identified as gender non-binary but was not part of our random sample. We made assumptions about gender based on names and pronouns used in public profiles on university websites or social media. We recognize that gender presentation, names, and pronouns are not necessarily indications of a person's gender and that, in the absence of additional information from the individuals, we may have unintentionally made incorrect assumptions about individuals' genders. In total we assessed 362 datasets published in 97 journals by 100 PIs. We scored the completeness and reusability of shared datasets on a scale from 0 (min score) to 5 (max score) following Roche *et al.* 2015 [24] (Table 1). The number of datasets assessed per researcher ranged from two to five; if a researcher had more than five shared datasets in the period from 2013-2019, we selected the most recent five.

Statistical analysis

We used a Bayesian bivariate mixed-effects regression model (R package MCMCglmm v2.32 [43]) to identify factors influencing data archiving quality and estimate its repeatability (i.e., the proportion of the total variance attributable to differences among individuals) [42, 44, 45]. Data completeness and reusability scores were included as two dependent variables in the model; researcher ID and university were specified as random effects, with researcher nested within university; PI gender, PI seniority (measured as the year of their first peer-reviewed publication, assessed on Google Scholar or Web of Science), PI author status (corresponding author or not), journal open data policy at the time of publication (mandatory, optional), and year of study publication were included as fixed effects. Journal impact factor (JIF) was not included in the model because MCMCglmm does not tolerate missing values in the fixed predictors (this is a deviation from the preregistered methodology doi:10.17605/OSF.IO/A492M).

The two dependent variables were mean-centered and standardized to one standard deviation unit prior to inclusion in the model (i.e., mean=0, standard deviation=1). The numerical (PI seniority, year of study publication) and categorical predictors (gender, corresponding author, journal open data policy) were mean-centered and standardized to two standard deviation units (i.e., mean=0, standard deviation=0.5) following Araya-Ajoy et al. [46]. Categorical predictors were treated as binary variables (values of 0 and 1) to allow centering and standardization. The advantage of mean-centering the predictors is that it ensures model intercepts are estimated for the average value of the predictors, facilitating interpretation of the results. Mean centering allows the estimate of the intercepts to be calculated for the average 'environmental' conditions [44]; the use of two standard deviations for predictor standardization allows for direct comparison of the variance explained by categorical and continuous predictors [46].

We specified a mildly informative inverse-Wishart prior and tested the sensitivity of the model to prior specification by examining how the posterior means and 95% credible intervals changed when specifying a parameter-expanded prior [see 43]. We checked the model by plotting the traces of the parameters, examining autocorrelation among samples drawn by MCMCglmm, and computing the Gelman-Rubin statistic to evaluate convergence (see archived script). Model diagnostics were satisfactory and conclusions were not sensitive to the choice of prior (Fig. S3).

We calculated the adjusted repeatability (R_{adj}) for a researcher's data completeness and data reusability as the proportion of the total variance due to differences among individuals when accounting for fixed and random effects in the statistical model [45]. Within- and among-individual correlations between data completeness and reusability were calculated as outlined in Roche et al. [42].

Following a reviewer suggestion, we conducted an exploratory (i.e., non-registered) analysis to examine whether dataset complexity could explain variation in data archiving quality. The rational for this analysis is that simple datasets (e.g., simple experimental design, few variables, low sample size) might be easier to share in a complete and reusable fashion than complex datasets containing many different experiments or observational studies, a large number of variables, and many measurements. We estimated dataset complexity as the number and size (in KB) of data files and plotted these variables against data completeness and reusability scores for each archived dataset (Fig. S4).

All analyses were done in R version 4.0.3.

Acknowledgements and funding sources

We thank Redouan Bshary for the idea that led to this study and Ellen Bledsoe and the associate editor and two anonymous reviewers for helpful comments on the manuscript. We acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (grant no. UIF-537860–2018) and the Canada Research Chair program. DGR was supported by the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. 838237-OPTIMISE.

Author Contributions

DGR, RD, and SAB designed the study. IB, FD, FL, SS, and DGR collected and managed the data. DGR analyzed the data. DGR and FD made the figures. DGR and SAB wrote the paper. All authors revised and approved the paper.

Competing Interests statement

DGR is a member of Research Data Canada's Policy Committee, a member of the Canadian National Committee for CODATA, and the president of the Society for Open, Reproducible and Transparent Ecology and Evolutionary biology (sortee.org).

Data and code availability

This study was pre-registered (https://doi.org/10.17605/OSF.IO/A492M). The anonymized data and the analysis script to reproduce our results are available on the Open Science Framework (https://doi.org/10.17605/OSF.IO/P2YWM) and were shared with the editors and reviewers upon submission.

Table and Figure captions

Table 1. Scoring system and criteria used to assess data completeness and reusability. Reproduced from Roche et al. (2015) https://doi.org/10.1371/journal.pbio.1002295

Fig. 1 Individual differences in the data archiving quality of open data published by principal investigators (PIs) in ecology and evolution. Each caterpillar plot shows (A) the completeness and (B) the reusability of data from 362 datasets published by 100 PIs. PIs are identified by a vertical grey line and ordered from lowest to highest individual mean score. The colour of the data points indicates the year in which a study was published.

Fig. 2 Only publication year and PI seniority predict data reusability (but not completeness). Posterior means and 95% credible intervals from a Bayesian bivariate mixed-effects model to examine predictors of data completeness and reusability (n = 362 open datasets shared by 100 principal investigators [PIs]). The predictor variables included in the model include: the year of the PI's first publication as a measure of seniority, the year in which the study was published, the PI's gender, the journals' open data policy, and whether the PI was the corresponding author on the published study. Black dots indicate weak relationships and grey dots indicate posteriors that overlap zero.

References

1. Wilkinson M.D., Dumontier M., Aalbersberg I.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L.B., Bourne P.E. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018.
2. Soeharjono S., Roche D.G. 2021 Reported individual costs and benefits of sharing open data among Canadian faculty members in ecology and evolution. *BioScience* **71**, 750–756. (doi:10.1093/biosci/biab024).
3. Culina A., van den Berg I., Evans S., Sánchez-Tójar A. 2020 Low availability of code in ecology: A call for urgent action. *PLoS Biol.* **18**, e3000763. (doi:10.1371/journal.pbio.3000763).
4. Costello M.J. 2009 Motivating online publication of data. *BioScience* **59**, 418-427.
5. Parr C.S., Cummings M.P. 2005 Data sharing in ecology and evolution. *Trends Ecol. Evol.* **20**, 362-363. (doi:10.1016/j.tree.2005.04.023).
6. Roche D.G., Lanfear R., Binning S.A., Haff T.M., Schwanz L.E., Cain K.E., Kokko H., Jennions M.D., Kruuk L.E. 2014 Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol.* **12**, e1001779. (doi:10.1371/journal.pbio.1001779).
7. Westoby M., Falster D.S., Schrader J. 2021 Motivating data contributions via a distinct career currency. *Proc. R. Soc. Biol. Sci. B* **288**, 20202830. (doi:10.1098/rspb.2020.2830).
8. Xu B., Gutierrez B., Mekaru S., Sewalk K., Goodwin L., Loskill A., Cohn E.L., Hswen Y., Hill S.C., Cobo M.M., et al. 2020 Epidemiological data from the COVID-19 outbreak, real-time case information. *Sci. Data* **7**, 106. (doi:10.1038/s41597-020-0448-0).
9. Hood A.S.C., Sutherland W.J. The data-index: An author-level metric that values impactful data and incentivizes data sharing. *Ecology and Evolution* **11**, 14344–14350. (doi:10.1002/ece3.8126).
10. Buxton R., Nyboer E.A., Pigeon K., Raby G.D., Rytwinsky T., Gallagher A.J., Schuster R., Lin H.S., Fahrig L., Bennett J.R., et al. 2021 Avoiding wasted research resources in conservation science. *Conserv. Sci. Prac.* **3**, e329. (doi:10.1111/csp2.329).
11. Roche D.G., O'Dea R.E., Kerr K.A., Rytwinski T., Schuster R., Nguyen V.M., Young N., Bennett J.R., Cooke S.J. Closing the knowledge-action gap in conservation with open science. *Conserv. Biol.* (doi:10.1111/cobi.13835).

12. Mills J.A., Teplitsky C., Arroyo B., Charmantier A., Becker P.H., Birkhead T.R., Bize P., Blumstein D.T., Bonenfant C., Boutin S. 2015 Archiving primary data: solutions for long-term studies. *Trends Ecol. Evol.* **30**, 581-589.

13. Tedersoo L., Küngas R., Oras E., Köster K., Eenmaa H., Leijen Ä., Pedaste M., Raju M., Astapova A., Lukner H., et al. 2021 Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192. (doi:10.1038/s41597-021-00981-0).

14. Tenopir C., Rice N.M., Allard S., Baird L., Borycz J., Christian L., Grant B., Olendorf R., Sandusky R.J. 2020 Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLOS ONE* **15**, e0229003. (doi:10.1371/journal.pone.0229003).

15. Science Digital, Simons N., Goodey G., Hardeman M., Clare C., Gonzales S., Strange D., Smith G., Kipnis D., Iida K., et al. 2021 The State of Open Data 2021. Digital Science. Report. https://doi.org/10.6084/m9.figshare.17061347.v1

16. Moore A.J., McPeek M.A., Rausher M.D., Rieseberg L., Whitlock M.C. 2010 The need for archiving data in evolutionary biology. *J. Evol. Biol.* **23**, 659-660. (doi:10.1111/j.1420-9101.2010.01937.x).

17. Berberi I., Roche D.G. 2021 Living database of journal data policies in E&E. Available on the Open Science Framework: https://doi.org/10.17605/OSF.IO/D6SP3.

18. 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. 2013 Mandated data archiving greatly improves access to research data. *FASEB J.* **27**, 1304-1308.

19. Federer L.M., Belter C.W., Joubert D.J., Livinski A., Lu Y.-L., Snyders L.N., Thompson H. 2018 Data sharing in PLOS ONE: An analysis of data availability statements. *PloS One* **13**, e0194768.

20. Roche D.G., Raby G.D., Norin T., Ern R., Scheuffele H., Skeeles M., Morgan R., Andreassen A.H., Clements J.C., Louissaint S., et al. 2022 Paths towards greater consensus building in experimental biology. *J. Exp. Biol.* **225**, jeb243559. (doi:10.1242/jeb.243559).

21. O'Dea R.E., Parker T.H., Chee Y.E., Culina A., Drobniak S.M., Duncan D.H., Fidler F., Gould E., Ihle M., Kelly C.D., et al. 2021 Towards open, reliable, and transparent

ecology and evolutionary biology. *BMC Biology* **19**, 68. (doi:10.1186/s12915-021-01006-3).

22. Couture J.L., Blake R.E., McDonald G., Ward C.L. 2018 A funder-imposed data publication requirement seldom inspired data sharing. *Plos One* **13**, e0199789. (doi:10.1371/journal.pone.0199789).
23. Vines T.H. 2014 Scientific community: Journals must boost data sharing. *Nature* **508**, 44-44.
24. Roche D.G., Kruuk L.E., Lanfear R., Binning S.A. 2015 Public data archiving in ecology and evolution: how well are we doing? *PLoS Biol.* **13**, e1002295.
25. Towse J., Ellis D.A., Towse A.S. 2020 Opening Pandora's Box: Peeking inside psychology's data sharing practices, and recommendations for change. *Beh. Res. Meth.* (doi:10.31234/osf.io/k6rux).
26. Hardwicke T.E., Mathur M.B., MacDonald K., Nilsonne G., Banks G.C., Kidwell M.C., Hofelich Mohr A., Clayton E., Yoon E.J., Henry Tessler M. 2018 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.
27. Sholler D., Ram K., Boettiger C., Katz D.S. 2019 Enforcing public data archiving policies in academic publishing: A study of ecology journals. *Big Data Soc.*, 1-18. (doi:10.1177/2053951719836258).
28. Christian T.-M., Gooch A., Vision T., Hull E. 2020 Journal data policies: Exploring how the understanding of editors and authors corresponds to the policies themselves. *PLOS One* **15**, e0230281. (doi:10.1371/journal.pone.0230281).
29. Roche D.G. 2016 Open data: policies need policing. *Nature* **538**, 41.
30. Wass M.N., Ray L., Michaelis M. 2019 Understanding of researcher behavior is required to improve data reliability. *GigaScience* **8**, 1-8. (doi:10.1093/gigascience/giz017).
31. Careau V., Wilson R.S. 2017 Of Uberfleas and Krakens: Detecting Trade-offs Using Mixed Models. *Integrative and Comparative Biology* **57**, 362-371. (doi:10.1093/icb/icx015).
32. Fairbairn D.J. 2011 The advent of mandatory data archiving. *Evolution* **65**, 1-2.
33. Tenopir C., Dalton E.D., Allard S., Frame M., Pjesivac I., Birch B., Pollock D., Dorsett K. 2015 Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS One* **10**, e0134826.

34. Tenopir C., Allard S., Douglass K., Aydinoglu A.U., Wu L., Read E., Manoff M., Frame M. 2011 Data sharing by scientists: practices and perceptions. *PLoS One* **6**, e21101. (doi:10.1371/journal.pone.0021101).

35. Chawinga W.D., Zinn S. 2019 Global perspectives of research data sharing: A systematic literature review. *Libr. Inf. Sci. Res.* **41**, 109-122.

36. Abele-Brehm A.E., Gollwitzer M., Steinberg U., Schönbrodt F.D.J.S.P. 2019 Attitudes toward Open Science and public data sharing: A survey among members of the German Psychological Society. *Soc. Psychol.* **50**, 252-260. (doi:10.1027/1864-9335/a000384).

37. Campbell H.A., Micheli-Campbell M.A., Udyawer V. 2019 Early career researchers embrace data sharing. *Trends Ecol. Evol.* **34**, 95-98.

38. McNutt M.K., Bradford M., Drazen J.M., Hanson B., Howard B., Jamieson K.H., Kiermer V., Marcus E., Pope B.K., Schekman R., et al. 2018 Transparency in authors' contributions and responsibilities to promote integrity in scientific publication. *Proc. Natl. Acad. Sci.* **115**, 2557-2560. (doi:10.1073/pnas.1715374115).

39. Mons B. 2020 Invest 5% of research funds in ensuring data are reusable. *Nature* **578**, 491.

40. Renaut S., Budden A.E., Gravel D., Poisot T., Peres-Neto P. 2018 Management, archiving, and sharing for biologists and the role of research institutions in the technology-oriented age. *BioScience* **68**, 400-411. (doi:10.1093/biosci/biy038).

41. Natural Sciences and Engineering Research Council of Canada. 2017 Women in Science and Engineering in Canada. Available at: https://www.nserc-crsng.gc.ca/_doc/Reports-Rapports/WISE2017_e.pdf.

42. Roche D.G., Careau V., Binning S.A. 2016 Demystifying animal 'personality' (or not): why individual variation matters to experimental biologists. *J. Exp. Biol.* **219**, 3832-3843. (doi:10.1242/jeb.146712).

43. Hadfield J.D. 2010 MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1-22.

44. Dingemanse N.J., Dochtermann N.A. 2013 Quantifying individual variation in behaviour: mixed-effect modelling approaches. *J. Anim. Ecol.* **82**, 39-54.

45. Nakagawa S., Schielzeth H. 2010 Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol. Rev.* **85**, 935-956.

46. Araya-Ajoy Y.G., Mathot K.J., Dingemanse N.J. 2015 An approach to estimate short-term, long-term and reaction norm repeatability. *Methods Ecol. Evol.* **6**, 1462-1473.

SUPPLEMENTARY INFORMATION

for: *Roche DG, Berberi I, Dhane F, Lauzon F, Soeharjono S, Dakin R, Binning SA (in revision) Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution (in revision)*

Table S1. The number of ecologists and evolutionary biologists in Canada's top-21 ranked universities according to the 2019 THE World University Rankings. Universities are ordered alphabetically. Indicated are the number of PIs in ecology and evolution at each university (E&E PIs), PIs with at least two journal articles published between Jan 2013-June 2019 containing a data availability statement and/or associated open data (Open data PIs), and PIs randomly selected for analysis in this study (Selected PIs).

University	E&E PIs	Open data PIs	Selected PIs	
			Women	Men
Carleton University	11	9	3	3
Dalhousie University	12	8	1	4
Laval University	18	8	1	4
McGill University	17	16	3	3
McMaster University	7	0	0	0
Memorial University	13	5	2	2
Queen's University	12	5	1	4
Simon Fraser University	18	13	3	3
University of Alberta	23	11	3	3
University of British Columbia	41	21	3	3
University of Calgary	17	8	3	3
University of Guelph	19	8	2	3
University of Manitoba	14	4	0	2
University of Montreal	16	9	2	4
University of Ottawa	13	9	3	3
University of Saskatchewan	14	4	2	2
University of Toronto	35	31	3	3
University of Victoria	11	7	3	3
University of Waterloo	8	2	0	0
Western University	18	11	2	3
York University	14	5	4	1
Total	351	194	44	56

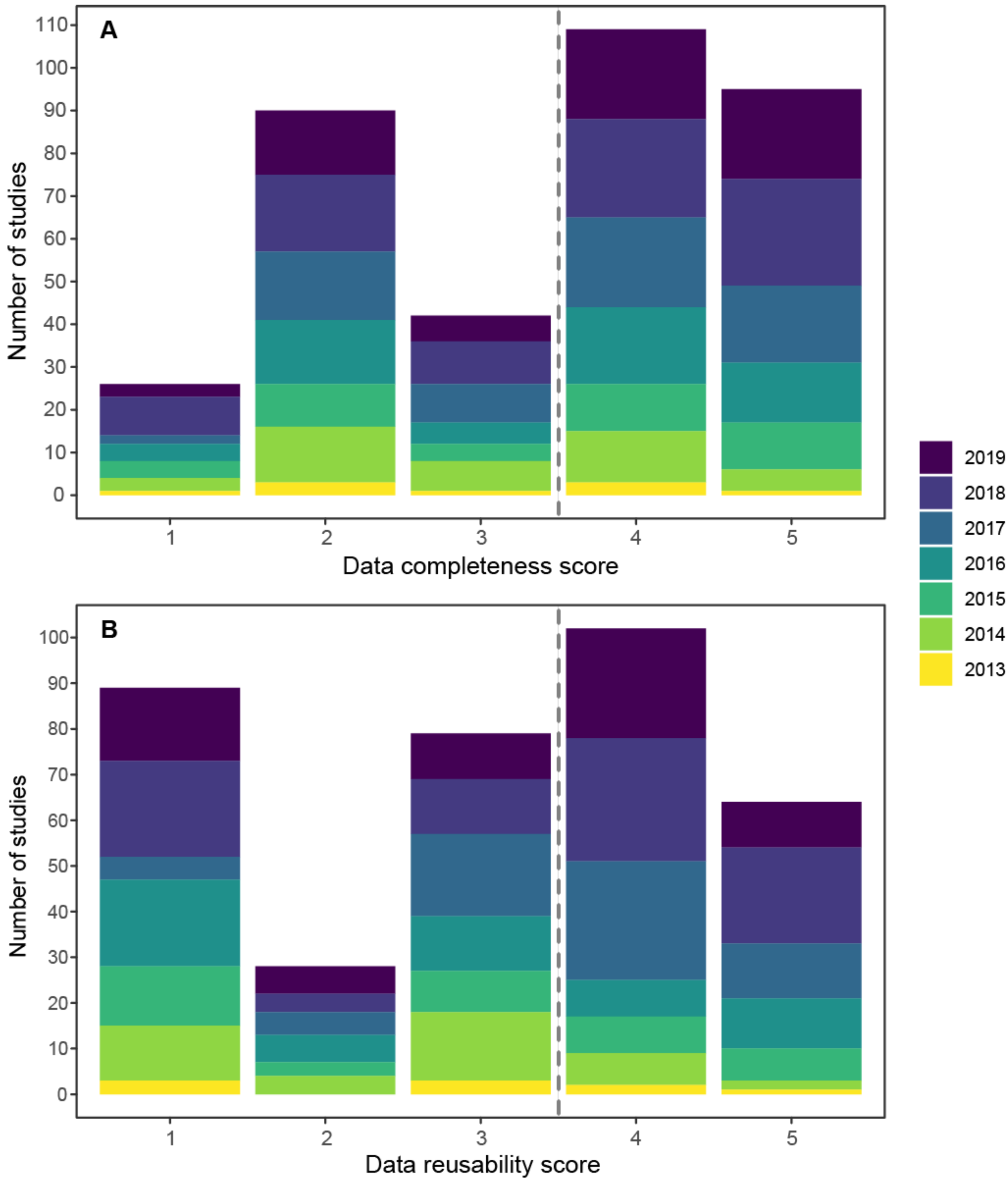

Fig. S1 Frequency distribution of the (A) completeness and (B) reusability scores for open datasets associated with 362 studies shared by 100 researchers between 2013-2019. A score of 5 indicates exemplary archiving, and a score of 1 indicates poor archiving (Table 1). Studies with scores of 3 or lower (left of the grey dashed lines) are incomplete or difficult to reuse.

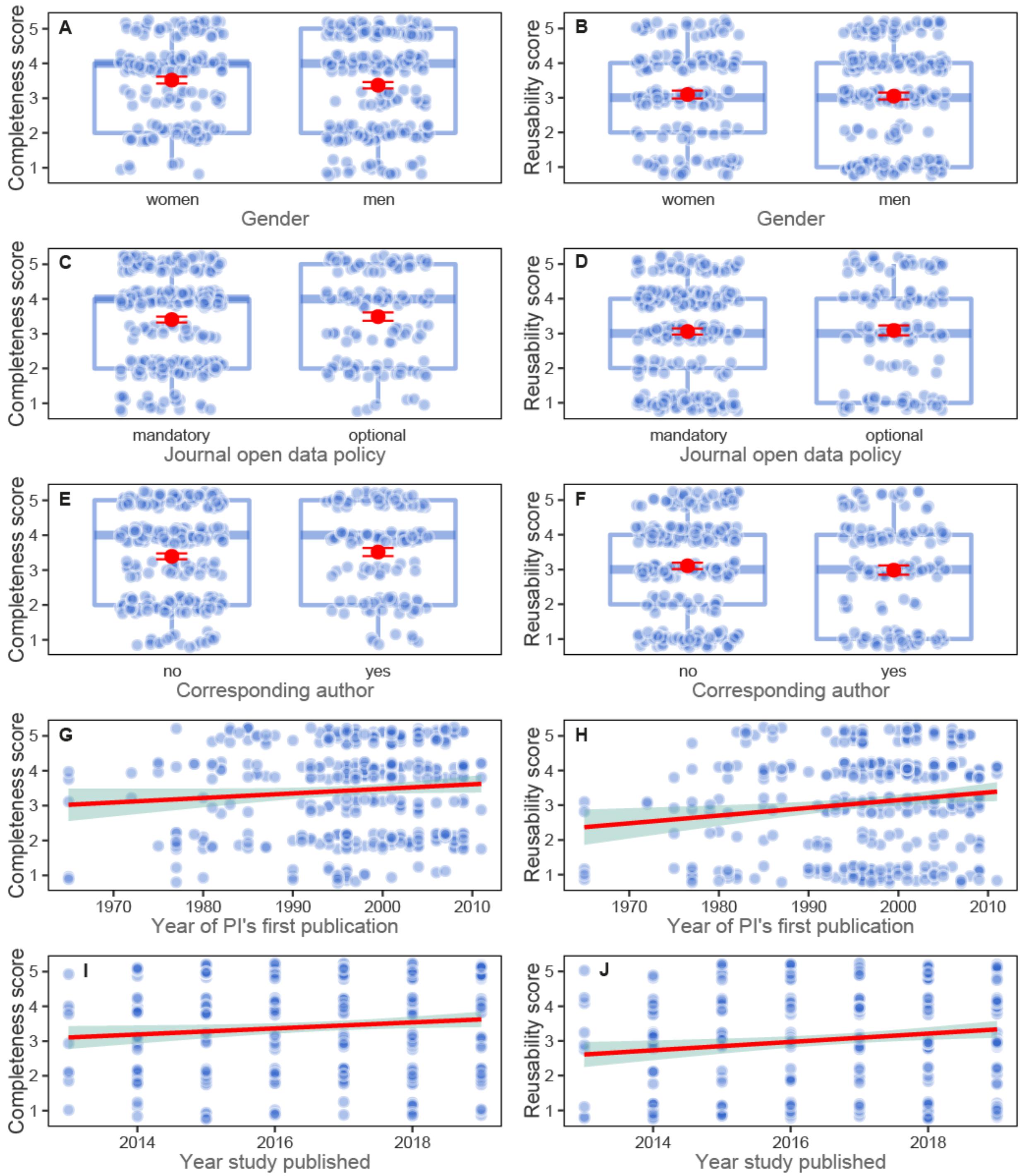

Fig. S2 The relationship between five independent variables and the completeness (A, C, E, G, I) and reusability score (B, D, F, H, J) of 362 open datasets shared by 100 principal investigators (PIs). Independent variables include: the gender of the PI, the journals' open data policy, whether the PI was the corresponding author on the associated paper, the year of the PI's first publication (i.e., seniority), and the year in which the study was published. Red dots are means and error bars represent 95% confidence intervals (CIs). Red lines are least square regressions and shaded areas represent 95% CIs. Note that these graphs are included for visualization purposes only to show the raw data. The relationships depicted do not control for other predictors in the analysis, nor do they control for repeated measurements (i.e., the random effects PI and university), which are included in the main analysis reported in the main text. For posterior means and 95% credible intervals from the Bayesian bivariate mixed-effects model, see Fig. 2.

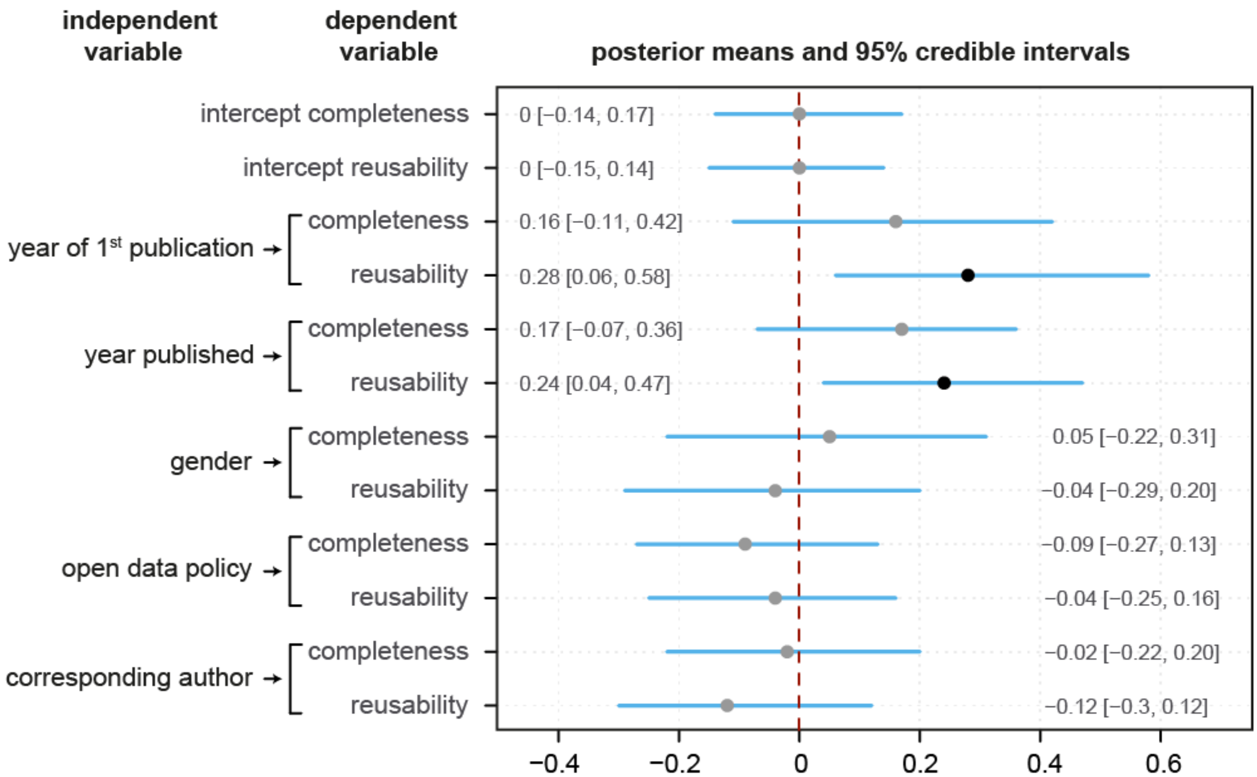

Fig. S3 Posterior means and 95% credible intervals from a Bayesian bivariate mixed-effects model investigating the effect of five independent variables on the completeness and reusability scores of 362 open datasets shared by 100 principal investigators (PIs). The prior was specified as a parameter-expanded prior (in contrast to an inverse-Wishart prior; see Fig. 2). The predictor variables included in the model include: the year of the PI's first publication as a measure of seniority, the year in which the study was published, the PI's gender, the journals' open data policy, and whether the PI was the corresponding author on the published study. Black dots indicate weak relationships and grey dots indicate posteriors that overlap zero.

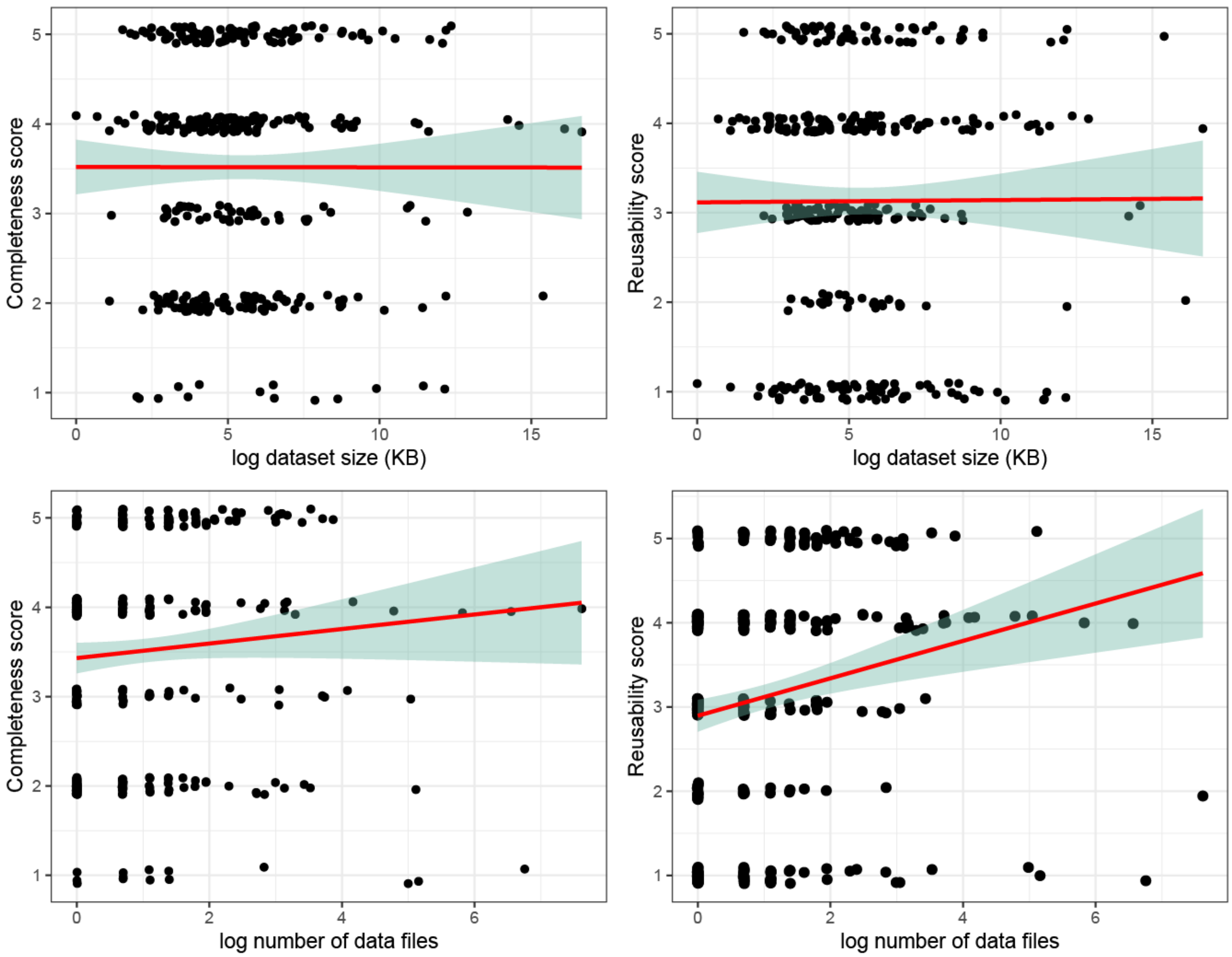

Fig. S4 The relationship between the completeness and reusability scores of open datasets and their complexity as estimated by the number and size (in KB) of archived files. Note that both variables used to estimate dataset complexity should be interpreted with caution as datasets were not archived using a standard file type (some file types [e.g. .xlsx] are inherently larger than others [e.g., .txt) and some files may contain multiple spreadsheets (e.g., .xlsx files can contain multiple tabs or spreadsheets as opposed to .txt or .csv files, which only contain one).