PROCEEDINGS B

royalsocietypublishing.org/journal/rspb

Biological science practices

Cite this article: Gomes DGE *et al*. 2022 Why don't we share data and code? Perceived barriers and benefits to public archiving practices. *Proc. R. Soc. B* **289**: 20221113. https://doi.org/10.1098/rspb.2022.1113

Received: 8 June 2022
Accepted: 2 November 2022

Subject Category:
Global change and conservation

Subject Areas:
computational biology, ecology, environmental science

Keywords:
open science, data science, reproducibility, transparency, data reuse, code reuse

Author for correspondence:
Dylan G. E. Gomes
e-mail: dylan.ge.gomes@gmail.com

Why don't we share data and code? Perceived barriers and benefits to public archiving practices

Dylan G. E. Gomes[1,2], Patrice Pottier[3,†], Robert Crystal-Ornelas[4,†], Emma J. Hudgins[5], Vivienne Foroughirad[6], Luna L. Sánchez-Reyes[7], Rachel Turba[8], Paula Andrea Martinez[9], David Moreau[10], Michael G. Bertram[11], Cooper A. Smout[12] and Kaitlyn M. Gaynor[13,14]

[1]NRC Research Associate, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, WA 98112, USA
[2]Cooperative Institute for Marine Resources Studies, Hatfield Marine Science Center, Oregon State University, Newport, OR 97365, USA
[3]Evolution & Ecology Research Centre, School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, New South Wales 2052, Australia
[4]Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[5]Department of Biology, Carleton University, Ottawa, Canada, K1S 5B6
[6]Department of Biology, Georgetown University, Washington, DC 20057, USA
[7]School of Natural Sciences, University of California, Merced, 95343 USA
[8]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095-7239, USA
[9]Australian Research Data Commons, The University of Queensland, Brisbane 4072, Australia
[10]School of Psychology and Centre for Brain Research, University of Auckland, Auckland 1010, New Zealand
[11]Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, Umeå, SE-907 36, Sweden
[12]Institute for Globally Distributed Open Research and Education (IGDORE), Brisbane 4001, Australia
[13]Departments of Zoology and Botany, University of British Columbia, Vancouver, Canada, BC V6T 1Z4
[14]National Center for Ecological Analysis and Synthesis, Santa Barbara, CA 93101, USA

DGEG, 0000-0002-2642-3728; PP, 0000-0003-2106-6597; RC-O, 0000-0002-6339-1139; EJH, 0000-0002-8402-5111; VF, 0000-0002-8656-7440; LLS-R, 0000-0001-7668-2528; RT, 0000-0003-3388-4503; PAM, 0000-0002-8990-1985; DM, 0000-0002-1957-1941; MGB, 0000-0001-5320-8444; CAS, 0000-0003-1144-3272; KMG, 0000-0002-5747-0543

The biological sciences community is increasingly recognizing the value of open, reproducible and transparent research practices for science and society at large. Despite this recognition, many researchers fail to share their data and code publicly. This pattern may arise from knowledge barriers about how to archive data and code, concerns about its reuse, and misaligned career incentives. Here, we define, categorize and discuss barriers to data and code sharing that are relevant to many research fields. We explore how real and perceived barriers might be overcome or reframed in the light of the benefits relative to costs. By elucidating these barriers and the contexts in which they arise, we can take steps to mitigate them and align our actions with the goals of open science, both as individual scientists and as a scientific community.

1. Introduction

Science is an iterative process in which our understanding of the world is continually updated with new information. Open, reproducible and transparent science practices allow us to more quickly and reliably evaluate, replicate and integrate studies to advance our knowledge [1–3]. A key component of open science is the publishing of datasets and analytical code used to make scientific

[†]These authors contributed equally.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6296319.

THE ROYAL SOCIETY PUBLISHING

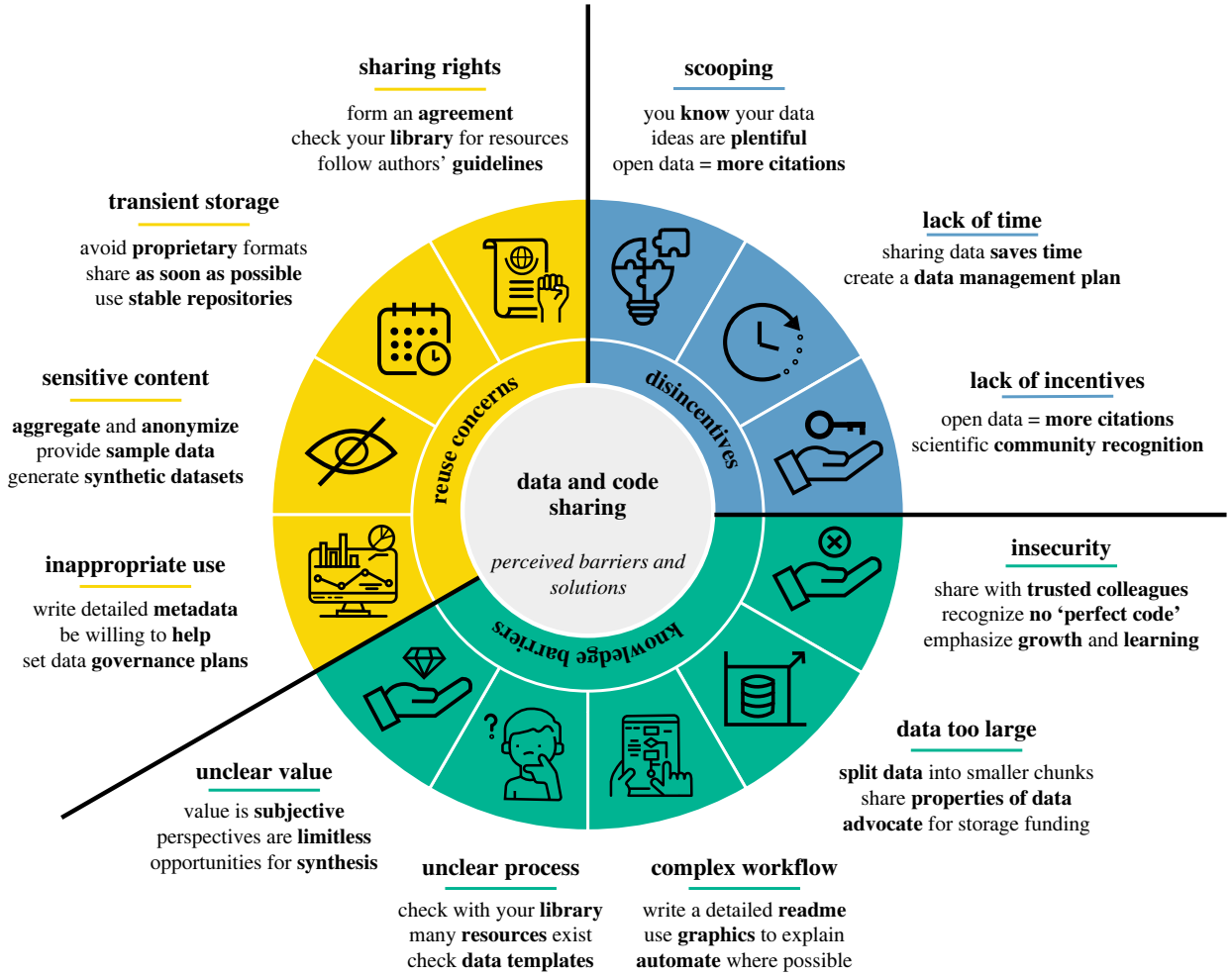

Figure 1. Perceived barriers and solutions to sharing data and code. We highlight 12 distinct barriers (see icons and corresponding underlined titles) to researchers publicly sharing data and code, which can be broken into three larger groups (knowledge barriers, reuse concerns and disincentives; innermost circle). Underneath the section titles, we list a few suggestions for overcoming these barriers (see main text for more details). (Online version in colour.)

inference [4,5]. Given the rapid growth of computational resources to store, process and analyse big data, sharing data and code with the public is easier and more important than ever.

Data and code sharing allows innovative reanalysis with new, improved methods or synthesis with other datasets, potentially leading to new insights [6–8]. Datasets collected for the purpose of answering one particular question can also be valuable assets to future researchers with entirely different questions and goals [9]. As computer programming becomes more necessary and accessible for reproducible data cleaning, processing, model building and statistical analyses, the value of code (e.g. programming scripts or other scientific software) for the scientific community is increasing [5]. While common language can only approximate a useful description of an analytical method, code can guide precise reproduction of the methodology in a research article. Code sharing saves researchers time from 'reinventing the wheel' in future projects and allows others to modify existing code for their own purposes.

In addition to advancing the scientific enterprise, publicly sharing data and code can benefit society at large [10–12]. For example, open science practices led to rapid advancements in our understanding of, and thus ability to combat, the emergence of SARS-CoV-2 [13–16]. However, one does not have to be an altruist to share data and code, as there are also many individual benefits that often outweigh any perceived costs [17]. For example, researchers who practice open science benefit from increased citation rates, visibility, collaboration efficiency and ease of future work [18–20].

Despite the benefits of open science for individual researchers, science and society, many biologists do not publicly share their data and code. We convened a working group at the inaugural (2021) meeting of the Society for Open, Reproducible, and Transparent Ecology and Evolution (SORTEE) to explore barriers to data and code sharing, and this paper is the distillation of our discussion (see description of process in §6 below). Here, we review common reasons for the failure to adopt open data and code practices. We have grouped these reasons into three broad categories: knowledge barriers, reuse concerns and career incentives (figure 1). Our target audience is the individual researcher who is looking to navigate the open science landscape amidst uncertainty and hesitation, and we therefore focus on changes in individual behaviour and offer counterpoints to alleviate the individual researcher's concerns. That said, we recognize the importance of top–down as well as bottom–up change, and we discuss the critical role of journals, funding agencies and research institutions in setting policies to incentivize individual behaviour. We hope that our recommendations empower individuals to both alter

their own behaviour and advocate for top–down change, and we encourage readers in decision-making positions to promote systemic change toward more open biological research.

2. Knowledge barriers

(a) Unsure about the process

Many researchers do not share their data and code simply because they do not know how. The process of archiving data and code is not always straightforward. In one survey of biologists, 46% of respondents were unaware of how to organize data in a presentable and usable way, and 33% reported that not knowing which online hosting service, or repository, to use was a barrier to sharing data and code [21]. The choice of repository can depend on multiple factors, like the type of digital output, science domain, size, national policy, funding agency and access restrictions [22–24], although there are general repositories that capture many forms of digital outputs, such as Zenodo (https://www.zenodo.org), OSF (https://osf.io/), Dryad (https://www.datadryad.org) and Figshare (https://figshare.com).

While data and code sharing can be daunting, there is a growing number of online resources and tips to support individuals and teams of researchers through the process (see [25–28]). For example, editorial support staff and institutional libraries often provide free, but under-used, guidance or assistance in archiving data and code [29]. There are also many data and metadata templates to help standardize data and to ensure that data are reusable by others [30–32]. FAIR principles and practices (**F**indability, **A**ccessibility, **I**nteroperability and **R**euse), for example, provide a framework and a set of guidelines that help researchers understand how to share data and code most effectively [33] (https://www.go-fair.org/fair-principles/). These templates will likely save researchers time in the long term, as data will be more organized and readily usable for their own future work [34]. Ultimately, even if data are shared in a repository that may not be the ideal fit, or if the code is not optimized or fully annotated, some form of data and code sharing is better than none. One of the best ways to gain knowledge about data and code sharing is through experience, so we encourage researchers to use the resources that are available to them, not to shy away from publications that require sharing, and accept that their practices will improve over time.

(b) Complex or manual workflows

When many manual steps are involved in a data workflow, researchers may be unsure about how to share their process in a fully reproducible manner. While the best practice for open science is to process and clean data with reproducible code, researchers have different levels of comfort with programming and some workflows may require manual steps or proprietary software. As a result, some intermediate data products may not be derivable from code alone.

To facilitate reproducibility in these cases, researchers should detail any manual data processing steps or point-and-click selection tools of a workflow in a metadata or readme file that accompanies data and code [35]. These manual workflows include manually summarizing or cleaning data in spreadsheets, cursor-based polygon selections in GIS software or cursor-based acoustic analyses, for example. The manual steps required in between scripts should be described as clearly as possible, and unless the process is highly subjective, the results should be approximately reproducible with sufficient detail. Using non-proprietary documents (e.g. pdf) with embedded images (e.g. screenshots), workflow diagrams, graphical readme files and other explanatory figures as supplemental information to a manuscript can aid the reader in understanding such complex manual steps. It is easy to forget exact steps after data are analysed, thus it is imperative that researchers document these manual steps as carefully as possible throughout the work, starting at the conception of the project.

Of course, seeking ways to reduce manual steps through automation can make for more efficient and reproducible workflows. For example, tools like OpenRefine (https://openrefine.org) can help write 'recipes' from point-and-click data cleaning workflows. Manual tasks can be converted to a coding script retroactively, giving downstream users control over the inputs needed in that stage of the workflow. For example, manual point-and-click selection of polygons in GIS software can be turned into code by defining the selected values (e.g. latitude and longitude) after the process. As another example, cursor-based trimming of acoustic files can be turned into a programmatic command that reads a set of input values (e.g. start time and end time) and carries out a function at those inputs (such as with ffmpeg [36]). While the inputs may have been derived manually, the process can be documented in functional steps with code.

(c) Large data files

Datasets are rapidly growing in size, complexity and quantity, thereby creating logistical barriers to sharing [37,38]. For example, some types of data like remotely sensed satellite imagery or climate model projections can produce terabytes or even petabytes of data per day [38,39]. Even transferring and storing datasets on the order of 1 GB can create challenges arising from file size [40]. Researchers may be wary of the storage space required to publicly share large data files, or unsure of best practices for bundling data into smaller subsets.

As cloud storage capacity grows each year, there are many opportunities for free storage of large research datasets. For example, there is no storage limit at the OSF repository and a 50 GB limit per dataset at Zenodo [22]. In rare cases in which datasets exceed these limitations, dataset managers can bundle smaller datasets for easier upload and downstream data reuse [41]. For example, The Climate Modelling Intercomparison Project (CMIP6) data and associated model runs contain petabytes of data, but are divided into smaller 'file sets' for more efficient storage and download [42]. The CMIP6 creators use consistent data and metadata standards [43,44] to ensure that all file sets are interoperable. Even working with smaller observational, experimental or modelling datasets, this process of bundling data files can make data management easier and more organized. Researchers might consider archiving model input and testing data by a relevant subgroup (e.g. by time period, variable groupings, or spatio-temporal resolution) [41]. As 'big data' continues to grow, funders, journals and research institutions need to offer financial and personnel support for the storage and maintenance of such large datasets.

(d) Insecurity

Insecurity, embarrassment and fear can be powerful emotional barriers to publicly sharing data and code. Publicly exposing the behind-the-scenes details of data management and analysis can feel vulnerable, especially for early-career researchers and novice coders [45]. Some may fear a scenario in which others find inaccuracies or errors in the data or analysis that undermines the results, which can lead to corrections or retractions [46] and weaken trust in the scientist and science in general [47].

To reduce the insecurity associated with the public sharing of data and code, researchers can first share materials with trusted co-authors or peers in a safe environment, like laboratory meetings or code sharing clubs. Pre-print servers can also provide a lower-stakes venue for soliciting feedback on code and analyses prior to formal peer-review (e.g. *bioRxiv*). At the time of manuscript submission, data and code should be shared with the journal peer-reviewers to improve the quality of the manuscript and supporting materials [48–50]. Getting feedback on code and documentation at all of these stages can improve its efficiency, clarity and utility beyond an individual project. Many repositories (e.g. Open Science Framework and Dryad) allow for a private 'peer-review' data and code sharing link if authors do not wish to make their products available to the public until after the review process is complete. It is important that peer-reviewers assess the quality of the data and code products themselves and report on whether there are sufficient metadata to understand such products [51,52]. If authors do not submit data and code for peer-review, we suggest that reviewers and editors recommend authors upload such products before final acceptance. Many peer-reviewers may not have the expertise to review such products. If this is the case, we encourage peer-reviewers to explain this to journal editors, who would benefit by explicitly soliciting peer-review of the data and code itself. Once published, code usage will generate additional feedback that will improve functionality and fix errors. It is also important to recognize that there is no such thing as 'perfect code'. There will always be trade-offs (e.g. among clarity, efficiency, ease and longevity) and there are diverse opinions about best practices for scientific software [53–56].

Furthermore, the process of cleaning and reviewing data and code for publication will usually reveal errors to the author before they are exposed publicly, which leads to higher-quality results than if data and code were not going to be published. If someone identifies a mistake in your data or code, this can easily be updated in the submission of a new 'version' (e.g. via Zenodo, Dryad or Figshare) of data and code. If this mistake changes the results of your published article, there is precedent for gracefully issuing a correction or, much more rarely, a retraction [46]. As a scientific community, we should continue to applaud those who acknowledge and correct human errors. By fostering a more inclusive, kind environment that emphasizes growth and learning over criticism and shame, we will reduce individual insecurities and fear associated with publicly sharing data and code [57,58].

(e) Do not see the value

Researchers may not envision that anyone else would be interested in their data or code and, therefore, do not see the value in sharing it. This may be particularly common when there are low sample sizes, a limited scope of data collection (e.g. in terms of geography, taxonomy and time), large amounts of uncertainty or error, and/or relatively simple or straightforward scripts. A review by Perrier *et al*. [59] suggests that regardless of the reason why people place low value on data sharing, this value judgement is an inherently subjective rather than an objective decision.

Uncertainty about potential reuse should not present a barrier to sharing, as there is a multitude of ways that a given set of data or code could be used by future generations of scientists, which is one reason why major funding agencies and many journals are now requiring open data and code products. Advances in science and technology allow for data reuse that the original data collectors never could have imagined [9,60,61]. In other cases, a dataset of poor quality or limited sampling may represent the only set of data on a particular subject, and its rarity may increase its value despite its shortcomings, as in the case of data on endangered species [62]. Moreover, data are often useful in novel synthesis analyses that may explore research questions entirely unrelated to the original motivation of the data collection. The open science movement is value-driven in pursuit of improved science, and by sharing data and code, we might contribute to interdisciplinary knowledge integration [63]. For example, open collaboration is exemplified by the growth of open projects on public platforms such as GitHub, where collaborators can add value to existing code by integrating their own ideas and knowledge [64]. The more information we leave for future researchers, the better they will be able to progress our understanding of the world around us.

3. Reuse concerns

(a) Inappropriate Use

Many scientists worry that, if they share their science openly, others will misinterpret their data or use their data and code inappropriately [1]. Those who are less familiar with the nuances of a particular data collection or analytical approach may overlook confounding factors and assumptions or draw erroneous and misleading conclusions through reuse.

Fortunately, researchers can take steps to reduce or avoid the inappropriate use of data and code. Data and code can be published alongside detailed metadata information, or with a data paper in an indexed, peer-reviewed journal, including a thorough description of datasets and processes, terms and considerations for reuse, and any limitations, assumptions, caveats, and shortcomings [65]. When one is accustomed to the nuances or assumptions of methods that they frequently use, it can be easy to forget to include important information that would allow others to replicate the study. Thus, metadata descriptions ideally would be looked over by someone other than the original researchers (e.g. peer-reviewers, friendly colleagues, etc.), who might more easily catch these omissions. Dryad data repository will review submitted metadata to some degree (while most free repositories do not), but this process could benefit from an explicit call to review metadata during journal peer-review.

Yet, open data that include thorough metadata can still omit important information that only the original data collectors had access to (e.g. idiosyncrasies of specific field sites or sampling years). Thus, researchers should also include

contact information and an invitation for others to collaborate and/or reach out for assistance in interpreting and using the data and code. Being open to helping others reuse data and code is the best way to avoid misinterpretation, and it may also create opportunities for new collaborations in research areas the original researchers would have never thought to pursue. Yet, it can often be unclear as to who the primary contact for a dataset should be and it is worth considering the longevity of such information. Early-career researchers lead most research [66–68] (but see [69]), yet high turnover in positions [70–72] means that institutional emails frequently become outdated, and principal investigators may not have the capacity or knowledge to respond to inquiries. To ensure continuity in contact information, research groups may consider establishing a shared email address that can persist despite personnel turnover. At the institutional level, financial support from government and funding agencies can further help research programs to maintain continuity, for example by hiring database managers who ensure that data and code products and metadata are well documented and available for future use.

One may also opt to publish data and code in repositories that allow the contributor to set the permissions and rights of access and reuse. For example, certain licences will require acknowledgement, prevent data being used for commercial purposes or being modified without the permission of the owner (see §3b below). Researchers can contact authors or journals to solicit a correction when there is data misuse, or publish a response to ensure the community is aware of it, although this process can be lengthy and complex and may not ultimately change the scientific record or narrative. Lastly, it is important to note that all forms of scientific products can be misused [73,74]. One can just as easily cite previous work erroneously or misinterpret findings in an article's discussion and/or data presented in its figures and tables as one can in data and code shared with the public—this is no reason not to publish these scientific products.

(b) Rights

Researchers may understandably feel a sense of ownership over data and code that they generate and may be hesitant to give up their exclusive right to use them. Furthermore, data and code may have complex ownership involving multiple people and institutions, complicating sharing efforts [75–77]. For example, research may have been conducted collaboratively, data may legally belong to an institution or funder rather than an individual researcher or data may be derived or synthesized from other primary datasets with different owners. Ownership of data and code may be further complicated after the publication of the original research article. Some publishers require a copyright assignment to the journal at the time of submission of a manuscript, which might include data and code products.

It is important to remember that we do not often have exclusive ownership of data to begin with. In cases where research is funded by federal government or public agencies (including, for example, the National Science Foundation and National Institutes of Health in the United States), researchers are obligated to publicly share research products that were generated through public funds for the benefit of society [11]. Institutional libraries and offices dedicated to copyright, open science and commercialization provide support and resources that can help researchers navigate the legal and ethical aspects of ownership and rights [29]. Data and code licences that define terms and conditions of reuse exist to protect researcher rights. In the context of collaborations, sharing agreements made early in the research process can specify the plans for ultimately sharing data, derived data products and code. When dealing with institutional or journal claims to research outputs, researchers should be aware of relevant policies and seek help clarifying the legal implications of institutional partnerships. When data and code are uploaded directly to a journal, those data and code products may be subjected to the same paywall as the article itself. Instead, considering the more general open repositories (listed in §2a above) can lead to increased accessibility and longevity (see §3d below) of data and code.

(c) Sensitive content

There are some situations in which publicizing data may not serve the best interest of science and society, and it should instead remain private [78]. This is sometimes referred to as the 'dual-use' dilemma [79], originally coined to describe the potential for biological data to be usurped for the purpose of bioterrorism. Within biology, notable scenarios that invoke the dual-use dilemma are sharing the location data for species under threat of poaching, capture for the pet trade [80], significant harassment or disturbance to species or their habitat from their whereabouts being exposed [81], private information about individuals [82] or individual interviews that are not meant to be public [83–86].

Researchers, communities and institutions, where appropriate, should weigh the benefits and costs (to individuals, local communities and society at large) of publishing data. In some cases, aggregating, generalizing or anonymizing data can be used to remove sensitive information. In the context of biodiversity conservation [87], there are guidelines regarding the assessment of the sensitivity of the species and the choice of appropriate levels of generalization and masking (either of the species' identity or location) using resources such as those provided by the Global Biodiversity Information Facility (GBIF; [88]). Sharing detailed metadata with a limited subsample of the data can help inform other scientists or stakeholders of the existence and utility of the data you possess [89]. This public-facing data description can include reliable correspondence information and an invitation to request the data privately (although sharing data privately rather than publicly should be done sparingly) [90]. These incomplete datasets can also allow users to test the operation of the accompanying analysis code, without jeopardizing the sensitive information found within the data.

Additionally, generating synthetic data can be used to provide proof of concept without violating ethics of sharing sensitive data. Within the biomedical field, technical solutions have been developed for sharing synthetic data that capture the statistical properties of the original dataset, including sequential data synthesis using regression and classification trees [91] and software frameworks like statistical health information release (SHARE; [92]). These methods are being developed and generalized toward fields outside of biomedicine, including accessible resources like the *synthpop* R package for synthetic data generation [93].

Importantly, it is necessary to consider how individuals and communities will be impacted by the publishing of certain

information. Sharing interview data, for example, without explicit consent from the interviewee is unethical, and reuse of this information out of context of the framing of the interview and questions can be problematic. In particular, many communities distrust science due to historic and ongoing harm, and special sensitivity is warranted in these cases. For example, Indigenous peoples and their data have been exploited and their natural resources abused by governments and commercial interests globally [83–86,94]. As we collectively move toward open data practices, we (as individuals, institutions, journals and funders) need to recognize the continued injustices to marginalized groups and advocate for data sovereignty. While open data is an important goal for advancing science, it must never perpetuate harm, and there are therefore circumstances in which data are best left unshared.

(d) Transient storage

Researchers may be reluctant to spend time making data and code publicly available if they are unsure of the usability of such products over the long term. Data and code may not be available indefinitely, given the lack of infrastructure for long-term storage facilities, proprietary storage formats and evolving software. Short-term storage options, such as GitHub and cloud-based storage (e.g. Google Drive, OneDrive and Dropbox), offer no promise of permanency as accounts (and thus data and code) can be deleted at any time by the user. Similarly, promises such as 'The raw data/analysis code supporting the conclusions of this article will be made available by the authors, without undue reservation' cannot be fulfilled if those authors lose hard drives, change email accounts, leave academia, or are deceased [90,95].

Researchers should archive their data in repositories that have the greatest likelihood of permanent support and maintenance, which are rarely the journals themselves. Some long-term generic storage infrastructure, such as Dryad and Zenodo, assign digital object identifiers (DOIs) and will retain all files for the lifetime of the repository. Some organizations or academic journals cover costs of long-term data archiving (e.g. CERN with Zenodo and The Royal Society journals with Dryad, respectively), while some funding agencies provide funding for long-term storage costs to their grantees (e.g. Wellcome Trust [96] and NIH [97]). Ultimately, securing funding to ensure long-term storage and usability of code is a community-driven goal that will require research institutions, funders and publishers to work together [98,99].

Additionally, researchers should avoid proprietary file formats and software, such as Microsoft suite (e.g. .doc and .xlsx formats), SAS or SPSS data formats [100]. These products are subject to the stability and consistency of these programs (and any required packages and dependencies) and the continued support for older file formats. To the extent possible, researchers should use stable, non-proprietary file formats (e.g. comma separated value, .csv, for data and plain text files, .txt, for documentation, provenance and metadata files). Another benefit of providing source code is that it can still be examined visually to reproduce past work, even if the code no longer runs properly due to different running environments, versioning issues or a lack of continued availability of software dependencies [101,102].

Researchers can make use of tools that promote backwards compatibility and portability of software and packages within different operating systems. These tools include software containers, which store all packages used alongside the code [103] (e.g. Docker, originally designed for app developers, 'renv' for R [104] and 'conda' for Python (https://docs.conda.io/en/latest/)). Bindr (https://mybinder.org) allows users to interactively run code (e.g. R, Python, Julia, etc.) on Jupyter notebooks, which might be stored remotely on a GitHub repository (for example: https://github.com/geo-yrao/esip-ml-tutorials).

4. Disincentives

(a) Scooping

One of the major barriers to data and code sharing is a fear of being 'scooped'. Scooping in this context colloquially refers to a situation in which a researcher performs analyses on publicly shared data that the original data collector had planned, but not yet completed themselves [105,106]. Potential code sharers may also fear that freely sharing code will reduce opportunities for collaboration and co-authorship with other researchers who may be interested in using their code. Furthermore, long-term datasets may not result in papers immediately, and researchers may be concerned that releasing data too early may compromise their ability to publish. The potential loss of future publications represents a cost in the context of today's scientific landscape, where publications are a key metric used to assess research productivity amidst competition for grants and positions [78,107].

Getting scooped is less likely than one might imagine, given that ideas are plentiful and diverse, and that those who collect data and develop code remain best positioned to undertake future analyses [108,109]. Researchers publish most papers using their own datasets within 2 years of original publication, while papers that cite open datasets peak at 5 years after data publication [18]. Additionally, pre-print servers offer the ability to make first claim to a research project through rapid dissemination of one's work and ideas [110]. Sharing how one collected open datasets along with any preliminary analyses or visualizations can alleviate concerns of being scooped when researchers do not immediately have time to go through the entire peer-review process. In these cases, pre-printed articles are already citable (with a DOI) and benefit from increased viewership, citation rates and collaborations [110–112] (but see [113] for concerns regarding pre-printing sensitive information, and §3c above).

If scooping is a major concern, there are ways to communicate expectations about how data should be used (e.g. see §3d above). In general, however, individual careers and scientific progress are advanced when we take a cooperative, collaborative approach [18,57,114], and data sharing will increase, rather than decrease, opportunities for collaboration. Institutions and funding agencies can alleviate scooping concerns and promote open science practices by viewing shared datasets and code as products that can be, in themselves, just as valuable as publications (see §4c below). Researchers who have spent their time, energy and finances on long-term datasets should receive appropriate credit (e.g. via promotion and future funding opportunities) for collecting such important data, regardless of whether these same researchers have led any scientific publications using the datasets. Giving disproportionate credit to new analyses, rather than new data collection efforts is limiting our knowledge and collective willingness to be open with

our work. As a community, we should be more inclusive of those who generate the data we use in our research. Those who have collected data are instrumental to a research project, and their participation in the development of a publication should be thoughtfully considered. At a minimum, care should be taken to follow the appropriate permissions and rights of access and reuse, and data should be properly cited (see §3 above).

(b) Lack of time

Researchers may be reluctant to share their data and code because of the perceived short- and long-term time commitments required to do so [21]. In the short term, it can take significant time to clean, prepare and annotate data, code and metadata for archiving, especially if these were not well organized from the beginning of the research project (for some guidance on that, see §2a above). In the long term, researchers may be reluctant to commit to ongoing curatorial support of others who try to reuse their data or code (see §4c below).

Despite the upfront time required, sharing research data and code can ultimately save time for individual researchers and their collaborators, as well as for others who want to reuse it. A researcher's most important collaborator is their future self [57], and the practice of annotating and organizing data and code is ultimately most useful to oneself. For example, archiving data in a long-term repository (e.g. Dryad, Figshare and Zenodo) ensures that users always have access to their own data and code files regardless of switching institutions or computers. Beginning a research project with the understanding that data and code will eventually be shared can generally lead to better standards, workflows and documentation throughout, and can reduce the time required for editing and cleaning once the project is complete [35]. The preparation of data and code should be considered as important as other publication tasks like managing citations and editing manuscript grammar, and should be prioritized in project management and delegation of roles within a team [90,98]. Research institutions can support this work by hiring designated data management teams that work with individual researchers, likely housed within institutional libraries [29]. Finally, the creation of supporting documents like descriptive metadata and readme files that include data and code version information [56] can help to ensure that the files are reusable in the long-term without further time commitment from the researcher.

(c) Lack of incentives

In addition to all of the perceived costs of sharing data and code outlined above, there is also a lack of perceived benefit among many researchers [59]. There have historically been few apparent career incentives to making one's data and code publicly available [115]. However, as discussed in the sections above, there are actually more career benefits to sharing data and code than one might realize.

Sharing data and code can increase visibility and recognition of a researcher within the scientific community, which may initiate new collaborations between data sharers and data reusers [116]. It can also help develop open science habits that increase efficiency, and contribute to a better understanding of one's own data and code (e.g. by providing descriptive metadata for files or commenting code). Research papers that include an access link to the primary data are cited significantly more often (25–69% more often) than papers that do not provide access to their data [18,20,117–119]. Data and software journals more frequently publish data and code with their own DOI, which allows data and code to be persistent, searchable, findable and formally cited. Thus, data and code uploads can be cited themselves, or included in a more comprehensive, stand-alone data paper (see https://www.gbif.org/data-papers) that is also citable—and at times to a high degree (https://www.earth-system-science-data.net/).

Increasingly, data and code sharing are being incentivized or even required by funders and publishers of scientific research [95,120,121]. Over the past decade, funding institutions have been acknowledging the importance of public data sharing in accelerating scientific discovery and advancement. Many recent recommendations for public funding agencies require that data and software generated with public funds be provided freely (e.g. OECD Council (https://www.oecd.org/); the U.S. White House [11]), and that funders should consider the value and impact of all research outputs (including data and software) in addition to publications (e.g. San Francisco Declaration on Research Assessment (DORA) (https://sfdora.org/read/); [122]). It is now common for funding institutions to require data sharing statements or data management plans in grant proposals, and to use those as part of funding allocation decisions. In the United States, the National Institutes of Health (NIH) has required data sharing since 2003 [97,123] and the National Science Foundation (NSF) [124] has required a data management plan for grant proposals since 2011. The NSF explicitly expects grantees to share primary data, and failure to comply with data management plans may negatively influence future funding opportunities, or result in the withholding or adjustment of funds [125]. Similarly, many scientific journals now require or strongly encourage data and code to be published alongside manuscripts [126]. The policing of such policies, however, could use strengthening.

Employers and academic institutions have been slower to incentivize data and code sharing with either rewards or punishments, but some institutions are beginning to value these practices among their employees. For example, in 2021, NASA launched their 'Transform to Open Science' initiative in which they proposed a number of incentives to reward and recognize data sharing actions. As part of this initiative, they are establishing an Open Source Science Award Program and aiming to incorporate open science activities into their reviews system. Professional societies are also granting awards to practitioners of open science, including SORTEE.

We hope that as more researchers recognize the value of open science, the publication of data and code will be considered in hiring, tenure and promotion [127]. Indeed, we are not alone in this desire, as the DORA begins: 'There is a pressing need to improve the ways in which the output of scientific research is evaluated by funding agencies, academic institutions and other parties' [128]. As of 31 August 2022, 22 081 individuals and organizations across 159 countries have signed the declaration. DORA outlines the importance of data and software products in individual outputs and makes specific recommendations for funding agencies, institutions, publishers and individual researchers (https://sfdora.org/read/).

5. Conclusion

We recognize that there are many real and perceived costs and barriers to sharing scientific data and code (figure 1). In many cases, on an individual level, perceived barriers may be relatively easily overcome (e.g. lack of knowledge) or may not actually present insurmountable obstacles (e.g. large file sizes). In other cases, the associated downstream benefits to research efficiency, productivity and collaboration may ultimately outweigh costs (e.g. time investment, fear of scooping). It is our hope that by outlining the above barriers to data and code sharing, we have enabled researchers to reflect on their own experiences and practices in order to recognize and mitigate the most salient barriers that they face.

As individuals, we should all make an effort to share well-documented data and code with clear and open lines of communication, which will reduce risks of data misuse while advancing the scientific enterprise. As members of our scientific communities, we should foster a culture that celebrates open science practices by our peers and advocate for incentives to share data and code in the context of research funding, publication and career evaluation. That is, data and code products are useful contributions to science on their own and should be valued as such. As journal editors and reviewers, for example, we can request that authors include data and code with their papers for peer-review—whether or not we have the skills or time to also review those products. Yet, we should be open about this lack of knowledge and journal editors (and authors) should explicitly solicit review of data and code products. Open science has great potential to advance the pace of scientific discovery while fostering a more collaborative and cooperative research environment, and publicly sharing data and code is a critical step towards these goals.

6. Process and authors' contributions

The initial discussion took two hours, was open to any who wanted to join, and was freely available via the SORTEE organisation and conference programme. In the discussion, we collaboratively brainstormed barriers to open data and code, drawing first from our own experiences as individual researchers. All SORTEE conference participants were invited to follow up to write this paper distilling our initial discussion about why we think individuals are reluctant to share data and code and to refine some counter points to these arguments. Those who had opted in participated in three follow-up discussions focused on consolidating and fleshing out the final list of barriers based on our experiences as well as the published literature. We all collaboratively compiled information and references for these arguments and counter-arguments. Each of us then drafted an individual section, followed by group edits. D.G.E.G., R.C.-O. and K.M.G. made final edits for consistency and clarity and P.P. made the central figure with feedback from all authors.

Data accessibility. The supplementary notes are provided in the electronic supplementary material [129].

Authors' contributions. D.G.E.G.: conceptualization, investigation, project administration, supervision, writing—original draft and writing—review and editing; P.P.: investigation, visualization, writing—original draft and writing—review and editing; R.C.-O.: investigation, writing—original draft and writing—review and editing; E.J.H.: investigation, writing—original draft and writing—review and editing; V.F.: investigation, writing—original draft and writing—review and editing; L.L.S.-R.: investigation, writing—original draft and writing—review and editing; R.T.: investigation, writing—original draft and writing—review and editing; P.A.M.: investigation, writing—original draft and writing—review and editing; D.M.: investigation and writing—review and editing; M.G.B.: investigation, writing—original draft and writing—review and editing; C.A.S.: investigation and writing—review and editing; K.M.G.: investigation, writing—original draft and writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. D.G.E.G. was funded by the National Academy of Sciences' NRC Research Associateship Program. P.P. was supported by a UNSW Scientia Doctoral Scholarship. R.C.-O. was funded through the ESS-DIVE repository by the U.S. DOE's Office of Science Biological and Environmental Research (grant no. DE-AC02-05CH11231). E.J.H. was supported by a Fonds de Recherche du Québec - Nature et Technologies B3X Postdoctoral Fellowship. L.L.S.-R. was funded by the National Science Foundation (NSF) grant no. ABI-1759846 to the Open Tree of Life project. R.T. was funded by the National Council for Scientific and Technological Development of Brazil (CNPq) under Grant No. 209261/2014-5. M.G.B. was funded by the Swedish Research Council Formas (grant no. 2020-02293) and the Kempe Foundations (grant nos. SMK-1954 and SMK21-0069). K.M.G. was supported by the National Center for Ecological Analysis and Synthesis Director's Postdoctoral Fellowship.

Acknowledgements. We are grateful to the SORTEE for hosting the conference that brought this group together and generated conversations around this topic. We thank Juliane Gaviraghi Mussoi and the other anonymous SORTEE attendees who participated in the initial workshop discussion but did not opt-in to write the paper with us. We thank Kyle Shannon, Vigdis Vandvik, and an anonymous reviewer for feedback on an earlier version of this manuscript.

References

1. Ramachandran R, Bugbee K, Murphy K. 2021 From open data to open science. *Earth Space Sci.* **8**, e2020EA001562. (doi:10.1029/2020EA001562)
2. Maskey M, Alemohammad H, Murphy KJ, Ramachandran R. 2020 Advancing AI for Earth science: a data systems perspective. *Eos* **101**, 25. (doi:10.1029/2020EO151245)
3. Lowndes JSS, Froehlich HE, Horst A, Jayasundara N, Pinsky ML, Stier AC, Therkildsen NO, Wood CL. 2019 Supercharge your research: a ten-week plan for open data science. *Nature*. (doi:10.1038/d41586-019-03335-4)
4. Poisot T, Bruneau A, Gonzalez A, Gravel D, Peres-Neto P. 2019 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)
5. 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)
6. Field AP, Gillett R. 2010 How to do a meta-analysis. *Br. J. Math. Stat. Psychol.* **63**, 665–694. (doi:10.1348/000711010X502733)
7. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. 2021 *Introduction to meta-analysis*. New York, NY: John Wiley & Sons.
8. Hedges LV. 1992 Meta-analysis. *J. Educ. Stat.* **17**, 279–296. (doi:10.3102/10769986017004279)
9. Berman JJ. 2015 *Repurposing legacy data: innovative case studies*. Amsterdam, The Netherlands: Elsevier.
10. Armeni K *et al*. 2021 Towards wide-scale adoption of open science practices: the role of open science communities. *Sci. Public Policy* **48**, 605–611. (doi:10.1093/scipol/scab039)

11. Obama B. 2013 Making open and machine readable the new default for government information. *White House*. Executive Order 13642 of May 9, 2013.
12. Kadakia KT, Beckman AL, Ross JS, Krumholz HM. 2021 Leveraging open science to accelerate research. *New Engl. J. Med.* **384**, e61. (doi:10.1056/NEJMp2034518)
13. Besançon L, Peiffer-Smadja N, Segalas C, Jiang H, Masuzzo P, Smout C, Billy E, Deforet M, Leyrat C. 2021 Open science saves lives: lessons from the COVID-19 pandemic. *BMC Med. Res. Methodol.* **21**, 1–18. (doi:10.1186/s12874-021-01304-y)
14. Edwin GT, Klug DM, Todd MH. 2020 Open science approaches to COVID-19. *F1000Research* **9**, 1043. (doi:10.12688/f1000research.26084.1)
15. Homolak J, Kodvanj I, Virag D. 2020 Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. *Scientometrics* **124**, 2687–2701. (doi:10.1007/s11192-020-03587-2)
16. Zuo X, Chen Y, Ohno-Machado L, Xu H. 2021 How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles. *Brief. Bioinform.* **22**, 800–811. (doi:10.1093/bib/bbaa331)
17. Allen C, Mehler DM. 2019 Open science challenges, benefits and tips in early career and beyond. *PLoS Biol.* **17**, e3000246. (doi:10.1371/journal.pbio.3000246)
18. Piwowar HA, Vision TJ. 2013 Data reuse and the open data citation advantage. *PeerJ* **1**, e175. (doi:10.7717/peerj.175)
19. Lortie CJ. 2021 The early bird gets the return: the benefits of publishing your data sooner. *Ecol. Evol.* **11**, 10736. (doi:10.1002/ece3.7853)
20. Piwowar HA, Day RS, Fridsma DB. 2007 Sharing detailed research data is associated with increased citation rate. *PLoS ONE* **2**, e308. (doi:10.1371/journal.pone.0000308)
21. Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Lucraft M, Astell M. 2018 Whitepaper: Practical challenges for researchers in data sharing. (doi:10.6084/m9.figshare.5996786)
22. Stall S *et al*. 2020 Generalist repository comparison chart. (doi:10.5281/ZENODO.3946720)
23. Pampel H *et al.* 2013 Making research data repositories visible: the re3data. org registry. *PLoS ONE* **8**, e78080. (doi:10.1371/journal.pone.0078080)
24. 2022 Data repository guidance. *Nat. Policies*. See https://www.nature.com/sdata/policies/repositories.
25. Soderberg CK. 2018 Using OSF to share data: a step-by-step guide. *Adv. Methods Practices Psychol. Sci.* **1**, 115–120. (doi:10.1177/2515245918757689)
26. Brown AV, Campbell JD, Assefa T, Grant D, Nelson RT, Weeks NT, Cannon SB. 2018 Ten quick tips for sharing open genomic data. *PLoS Comput. Biol.* **14**, e1006472. (doi:10.1371/journal.pcbi.1006472)
27. Wilson SL, Way GP, Bittremieux W, Armache JP, Haendel MA, Hoffman MM. 2021 Sharing biological data: why, when, and how. *FEBS Lett.* **595**, 847–863. (doi:10.1002/1873-3468.14067)
28. Eglen SJ *et al*. 2017 Toward standard practices for sharing computer code and programs in neuroscience. *Nat. Neurosci.* **20**, 770–773. (doi:10.1038/nn.4550)
29. Ayre LB, Craner J. 2017 Open data: what it is and why you should care. *Public Library Q.* **36**, 173–184. (doi:10.1080/01616846.2017.1313045)
30. Bond-Lamberty B, Christianson DS, Crystal-Ornelas R, Mathes K, Pennington SC. 2021 A reporting format for field measurements of soil respiration. *Ecol. Informatics* **62**, 101280. (doi:10.1016/j.ecoinf.2021.101280)
31. Ely KS *et al*. 2021 A reporting format for leaf-level gas exchange data and metadata. *Ecol. Informatics* **61**, 101232. (doi:10.1016/j.ecoinf.2021.101232)
32. Fowler D, Barratt J, Walsh P. 2017 Frictionless data: making research data quality visible. *Int. J. Digital Curation* **12**. (doi:10.2218/ijdc.v12i2.577)
33. Wilkinson MD *et al*. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 1–9. (doi:10.1038/sdata.2016.18)
34. Toczydlowski RH *et al*. 2021 Poor data stewardship will hinder global genetic diversity surveillance. *Proc. Natl Acad. Sci. USA* **118**, e2107934118. (doi:10.1073/pnas.2107934118)
35. Stoudt S, Vásquez VN, Martinez CC. 2021 Principles for data analysis workflows. *PLoS Comput. Biol.* **17**, e1008770. (doi:10.1371/journal.pcbi.1008770)
36. FFmpeg Developers. 2018 FFmpeg. (https://ffmpeg.org/)
37. Farley SS, Dawson A, Goring SJ, Williams JW. 2018 Situating ecology as a big-data science: current advances, challenges, and solutions. *BioScience* **68**, 563–576. (doi:10.1093/biosci/biy068)
38. Stephens ZD *et al*. 2015 Big data: astronomical or genomical? *PLoS Biol.* **13**, e1002195. (doi:10.1371/journal.pbio.1002195)
39. Agapiou A. 2017 Remote sensing heritage in a petabyte-scale: satellite data and heritage Earth Engine \copyright applications. *Int. J. Digital Earth* **10**, 85–102. (doi:10.1080/17538947.2016.1250829)
40. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N. 2019 Deep learning and process understanding for data-driven Earth system science. *Nature* **566**, 195–204. (doi:10.1038/s41586-019-0912-1)
41. Simmonds M *et al*. 2022 Guidelines for publicly archiving terrestrial model data to enhance usability, intercomparison, and synthesis. *Data Sci. J.* **21**, 3. (doi:10.5334/dsj-2022-003)
42. Eggleton F, Winfield K. 2020 Open data challenges in climate science. *Data Sci. J.* **19**, 52. (doi:10.5334/dsj-2020-052)
43. Gregory J. 2003 The CF metadata standard. *CLIVAR Exchanges* **8**, 4.
44. Hassell D, Gregory J, Blower J, Lawrence BN, Taylor KE. 2017 A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2. 1). *Geosci. Model Dev.* **10**, 4619–4646. (doi:10.5194/gmd-10-4619-2017)
45. Hampton SE *et al.* 2015 The Tao of open science for ecology. *Ecosphere* **6**, 1–13. (doi:10.1890/ES14-00402.1)
46. Fox J. 2020 The history of retractions from ecology and evolution journals. See https://dynamicecology.wordpress.com/2020/02/24/the-history-ofretractions-in-ecology-and-evolution/. (Posted on 24 February 2020.)
47. Pennisi E. 2020 Spider biologist denies suspicions of widespread data fraud in his animal personality research. *Sci. Insider* (doi:10.1126/science.abb1258)
48. Fernández-Juricic E. 2021 Why sharing data and code during peer review can enhance behavioral ecology research. *Behav. Ecol. Sociobiol.* **75**, 1–5. (doi:10.1007/s00265-021-03036-x)
49. Powers SM, Hampton SE. 2019 Open science, reproducibility, and transparency in ecology. *Ecol. Appl.* **29**, e01822. (doi:10.1002/eap.1822)
50. Squazzoni F *et al*. 2020 Unlock ways to share data on peer review. *Nature* **578**, 512–514. (doi:10.1038/d41586-020-00500-y)
51. MacLeod L, Greiler M, Storey MA, Bird C, Czerwonka J. 2017 Code reviewing in the trenches: challenges and best practices. *IEEE Softw.* **35**, 34–42. (doi:10.1109/MS.2017.265100500)
52. Kratz JE, Strasser C. 2015 Researcher perspectives on publication and peer review of data. *PLoS ONE* **10**, e0117619. (doi:10.1371/journal.pone.0117619)
53. Barnes N. 2010 Publish your computer code: It is good enough. *Nature* **467**, 753. (doi:10.1038/467753a)
54. Wilson G *et al*. 2014 Best practices for scientific computing. *PLoS Biol.* **12**, e1001745. (doi:10.1371/journal.pbio.1001745)
55. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK. 2017 Good enough practices in scientific computing. *PLoS Comput. Biol.* **13**, e1005510. (doi:10.1371/journal.pcbi.1005510)
56. Lee G *et al*. 2021 Barely sufficient practices in scientific computing. *Patterns* **2**, 100206. (doi:10.1016/j.patter.2021.100206)
57. Lowndes JSS, Best BD, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, Jiang N, Halpern BS. 2017 Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.* **1**, 1–7. (doi:10.1038/s41559-016-0001)
58. Gownaris NJ, Vermeir K, Bittner MI, Gunawardena L, Kaur-Ghumaan S, Lepenies R, Ntsefong GN, Zakari IS. 2022 Barriers to full participation in the open science life cycle among Early Career Researchers. *Data Sci. J.* **21**, 2. (doi:10.5334/dsj-2022-002)
59. Perrier L, Blondal E, MacDonald H. 2020 The views, perspectives, and experiences of academic researchers with data sharing and reuse: a meta-

synthesis. *PLoS ONE* **15**, e0229182. (doi:10.1371/journal.pone.0229182)

60. Winerman L. 2004 New uses for old data. Data sharing and large-scale databases can provide a wealth of information and save researchers both time and money. *American Psychological Association Career Center*. See https://www.apa.org/gradpsych/2004/09/newuses.
61. Van de Sandt S, Lavasa A, Dallmeier-Tiessen S, Petras V. 2019 The definition of reuse. *Data Sci. J.* **18**, 22. (doi:10.5334/dsj-2019-022)
62. Enquist BJ *et al*. 2019 The commonness of rarity: global and future distribution of rarity across land plants. *Sci. Adv.* **5**, eaaz0414. (doi:10.1126/sciadv.aaz0414)
63. Cohoon J, Howison J. 2021 Norms and open systems in open science. *Information Cult.* **56**, 115–137. (doi:10.7560/IC56201)
64. Mergel I. 2015 Open collaboration in the public sector: the case of social coding on GitHub. *Govern. Information Q.* **32**, 464–472. (doi:10.1016/j.giq.2015.09.004)
65. Candela L, Castelli D, Manghi P, Tani A. 2015 Data journals: a survey. *J. Assoc. Information Sci. Technol.* **66**, 1747–1762. (doi:10.1002/asi.23358)
66. Shin J, Cummings W. 2010 Multilevel analysis of academic publishing across disciplines: research preference, collaboration, and time on research. *Scientometrics* **85**, 581–594. (doi:10.1007/s11192-010-0236-2)
67. Costas R, Van Leeuwen TN, Bordons M. 2010 A bibliometric classificatory approach for the study and assessment of research performance at the individual level: the effects of age on productivity and impact. *J. Am. Soc. Information Sci. Technol.* **61**, 1564–1581. (doi:10.1002/asi.21244)
68. Lissoni F, Mairesse J, Montobbio F, Pezzoni M. 2011 Scientific productivity and academic promotion: a study on French and Italian physicists. *Indust. Corp. Change* **20**, 253–294. (doi:10.1093/icc/dtq073)
69. Abramo G, D'Angelo CA, Di Costa F. 2011 Research productivity: are higher academic ranks more productive than lower ones? *Scientometrics* **88**, 915–928. (doi:10.1007/s11192-011-0426-6)
70. Heffernan TA, Heffernan A. 2019 The academic exodus: the role of institutional support in academics leaving universities and the academy. *Prof. Dev. Educ.* **45**, 102–113. (doi:10.1080/19415257.2018.1474491)
71. Kelsky K. 2015 *The professor is in: the essential guide to turning your PhD into a job*. New York, NY: Three Rivers Press.
72. Gould J. 2019 Working Scientist podcast: too many PhDs, too few research positions. *Nat. Careers Podcast* (doi:10.1038/d41586-019-03439-x)
73. West JD, Bergstrom CT. 2021 Misinformation in and about science. *Proc. Natl Acad. Sci. USA* **118**, e1912444117. (doi:10.1073/pnas.1912444117)
74. Todd PA, Guest JR, Lu J, Chou LM. 2010 One in four citations in marine biology papers is inappropriate. *Mar. Ecol. Prog. Ser.* **408**, 299–303. (doi:10.3354/meps08587)
75. Asswad J, Marx Gómez J. 2021 Data ownership: a survey. *Information* **12**, 465. (doi:10.3390/info12110465)
76. 2018 Data ownership, rights and controls: Reaching a common understanding. Discussions at The British Academy, The Royal Society and techUK seminar on 3 October 2018. See https://royalsociety.org/-/media/policy/projects/datagovernance/data-ownership-rights-and-controls-October-2018.pdf.
77. Office of Research Integrity. 2022 Data Ownership. *Responsible Conduct in Data Management*. See https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dotopic.html.
78. Duke CS, Porter JH. 2013 The ethics of data sharing and reuse in biology. *BioScience* **63**, 483–489. (doi:10.1525/bio.2013.63.6.10)
79. Somerville MA, Atlas RM. 2005 Ethics: a weapon to counter bioterrorism. *Science* **307**, 1881–1882. (doi:10.1126/science.1109279)
80. Stuart BL, Rhodin AG, Grismer LL, Hansel T. 2006 Scientific description can imperil species. *Science* **312**, 1137. (doi:10.1126/science.312.5777.1137b)
81. Larson CL, Reed SE, Merenlender AM, Crooks KR. 2016 Effects of recreation on animals revealed as widespread through a global systematic review. *PLoS ONE* **11**, e0167259. (doi:10.1371/journal.pone.0167259)
82. Annas GJ. 2003 HIPAA regulations: a new era of medical-record privacy? *New Engl. J. Med.* **348**, 1486. (doi:10.1056/NEJMlim035027)
83. Walter M, Suina M. 2019 Indigenous data, indigenous methodologies and indigenous data sovereignty. *Int. J. Soc. Res. Methodol.* **22**, 233–243. (doi:10.1080/13645579.2018.1531228)
84. Rainie SC, Kukutai T, Walter M, Figueroa-Rodríguez OL, Walker J, Axelsson P. 2019 Indigenous data sovereignty. In *The State of Open Data: Histories and Horizons* (eds T Davies, SB Walker, M Rubinstein, F Perini), pp. 300–319. Cape Town, South Africa: African Minds and the International Development Research Centre (IDRC).
85. Lovett R, Lee V, Kukutai T, Cormack D, Rainie SC, Walker J. 2019 Good data practices for Indigenous data sovereignty and governance. In *Good data* (eds A Daly, SK Devitt, M Mann), pp. 26–36. Amsterdam, The Netherlands: Institute of Network Cultures.
86. Kukutai T, Taylor J. 2016 *Indigenous data sovereignty: toward an agenda*. Canberra, Australia: ANU Press.
87. Tulloch AI *et al*. 2018 A decision tree for assessing the risks and benefits of publishing biodiversity data. *Nat. Ecol. Evol.* **2**, 1209–1217. (doi:10.1038/s41559-018-0608-1)
88. Chapman AD. 2020 Current best practices for generalizing sensitive species occurrence data, version 1. Copenhagen, Denmark: GBIF.
89. Michener WK. 2015 Ecological data sharing. *Ecol. Informatics* **29**, 33–44. (doi:10.1016/j.ecoinf.2015.06.010)
90. Tedersoo L *et al*. 2021 Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 1–11. (doi:10.1038/s41597-021-00981-0)
91. Drechsler J, Reiter JP. 2011 An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput. Stat. Data Analysis* **55**, 3232–3243. (doi:10.1016/j.csda.2011.06.006)
92. Gardner J, Xiong L, Xiao Y, Gao J, Post AR, Jiang X, Ohno-Machado L. 2013 SHARE: system design and case studies for statistical health information release. *J. Am. Med. Inform. Assoc.* **20**, 109–116. (doi:10.1136/amiajnl-2012-001032)
93. Quintana DS. 2020 A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife* **9**, e53275. (doi:10.7554/eLife.53275)
94. Smith LT. 2021 *Decolonizing methodologies: research and indigenous peoples*. New York, NY: Bloomsbury Publishing.
95. Stodden V, Seiler J, Ma Z. 2018 An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl Acad. Sci. USA* **115**, 2584–2589. (doi:10.1073/pnas.1708290115)
96. Wellcome Fund. 2017 Data, software and materials management and sharing policy. See https://wellcome.org/grantfunding/guidance/data-software-materials-management-and-sharing-policy.
97. National Institutes of Health. 2020 Final NIH policy for data management and sharing. NOT-OD-21-013. Vol NOT-OD-21-013. NIH Grants & Funding. Bethesda, MD: Office of The Director, National Institutes of Health. See https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html.
98. Mons B. 2020 Invest 5% of research funds in ensuring data are reusable. *Nature* **578**, 491. (doi:10.1038/d41586-020-00505-7)
99. Knowles R, Mateen BA, Yehudi Y. 2021 We need to talk about the lack of investment in digital research infrastructure. *Nat. Comput. Sci.* **1**, 169–171. (doi:10.1038/s43588-021-00048-5)
100. Briney KA, Coates HL, Goben A. 2020 Foundational practices of research data management. *Rio* **6**, e56508. (doi:10.3897/rio.6.e56508)
101. Di Cosmo R, Zacchiroli S. 2017 Software heritage: Why and how to preserve software source code. In *iPRES 2017—14th Int. Conf. on Digital Preservation, Kyoto, Japan September 2017*, pp. 1–10. See https://hal.archives-ouvertes.fr/hal-01590958.
102. Shamir L, Wallin JF, Allen A, Berriman B, Teuben P, Nemiroff RJ, Mink J, Hanisch RJ, DuPrie K. 2013 Practices in source code sharing in astrophysics. *Astro. Comput.* **1**, 54–58. (doi:10.1016/j.ascom.2013.04.001)
103. Wiebels K, Moreau D. 2021 Leveraging containers for reproducible psychological research. *Adv. Methods Pract. Psychol. Sci.* **4**, 25152459211017853. (doi:10.1177/25152459211017853)
104. Ushey K. 2022 Renv: Project Environments. R package version 0.16.0. See https://rstudio.github.io/renv/.

105. Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, Kokko H, Jennions MD, Kruuk LE. 2014 Troubleshooting public data archiving: suggestions to increase participation. *PLoS Biol.* **12**, e1001779. (doi:10.1371/journal.pbio.1001779)
106. Brown CT. 2013 The Cost of Open Science. *Living in an Ivory Basement*. See http://ivory.idyll.org/blog/the-cost-ofopen-science.html.
107. Stodden V, Borwein J, Bailey DH. 2013 Setting the default to reproducible. *Computational science research. SIAM News* **46**, 4–6.
108. Silberzahn R *et al*. 2018 Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Practices Psychol. Sci.* **1**, 337–356. (doi:10.1177/2515245917747646)
109. Knorr-Cetina KD. 1983 The ethnographic study of scientific work: towards a constructivist interpretation of science. In *Science Observed: Perspectives on the Social Study of Science* (ed. KD Knorr-Cetina), pp. 115–140. London, UK: Sage.
110. Sarabipour S, Debat HJ, Emmott E, Burgess SJ, Schwessinger B, Hensel Z. 2019 On the value of preprints: an early career researcher perspective. *PLoS Biol.* **17**, e3000151. (doi:10.1371/journal.pbio.3000151)
111. Xie B, Shen Z, Wang K. 2021. *arXiv:* Is preprint the future of science? A thirty year journey of online preprint services. *arXiv preprint*, 2102.09066. (doi:10.48550/arXiv.2102.09066)
112. Bourne PE, Polka JK, Vale RD, Kiley R. 2017 Ten simple rules to consider regarding preprint submission. *PLoS Comput. Biol.* **13**, e1005473. (doi:10.1371/journal.pcbi.1005473)
113. Flanagin A, Fontanarosa PB, Bauchner H. 2020 Preprints involving medical research—Do the benefits outweigh the challenges? *JAMA* **324**, 1840–1843. (doi:10.1001/jama.2020.20674)
114. McKiernan EC *et al*. 2016 Point of view: how open science helps researchers succeed. *elife* **5**, e16800. (doi:10.7554/eLife.16800)
115. Soeharjono S, Roche DG. 2021 Reported individual costs and benefits of sharing open data among Canadian Academic Faculty in ecology and evolution. *BioScience* **71**, 750–756. (doi:10.1093/biosci/biab024)
116. Ding WW, Levin SG, Stephan PE, Winkler AE. 2010 The impact of information technology on academic scientists' productivity and collaboration patterns. *Manage. Sci.* **56**, 1439–1461. (doi:10.1287/mnsc.1100.1195)
117. Sears JR. 2011 Data sharing effect on article citation rate in paleoceanography. American Geophysical Union, Fall Meeting 2011, abstract id. IN53B-1628.
118. Christensen G, Dafoe A, Miguel E, Moore DA, Rose AK. 2019 A study of the impact of data sharing on article citations using journal policies as a natural experiment. *PLoS ONE* **14**, e0225883. (doi:10.1371/journal.pone.0225883)
119. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. 2020 The citation advantage of linking publications to research data. *PLoS ONE* **15**, e0230416. (doi:10.1371/journal.pone.0230416)
120. Cadwallader L, Papin JA, Mac Gabhann F, Kirk R. 2021 Collaborating with our community to increase code sharing. *PLoS Comput. Biol.* **17**, e1008867. (doi:10.1371/journal.pcbi.1008867)
121. Nature. 2014 Code share. *Nature* **514**, 536. (doi:10.1038/514536a)
122. Akhmerov A, Cruz M, Drost N, Hof CH, Knapen T, Kuzak M, Martinez-Ortiz C, Turkyilmaz-van der Velden Y, van Werkhoven B. 2019 Raising the profile of research software. Recommendations for funding agencies and research institutions. *Zenodo*. (doi:10.5281/zenodo.3378572)
123. Brickley P. 2003 NIH says scientists should share. *Genome Biol.* **4**, 1–3. (doi:10.1186/gb-2003-4-2-p1)
124. National Science Foundation. 2020 Proposal & award policies & procedure guide. See https://www.nsf.gov/pubs/policydocs/pappg22_1/.
125. National Science Foundation. 2015 NSF's public access plan: today's data, tomorrow's discoveries. Increasing access to the results of research funded by the National Science Foundation. See https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf.
126. Grant R, Hrynaszkiewicz I. 2018 The impact on authors and editors of introducing Data Availability Statements at Nature journals. *Nature* **13**. (doi:10.2218/ijdc.v13i1.614)
127. Schmidt R, Curry S, Hatch A. 2021 Research culture: creating SPACE to evolve academic assessment. *Elife* **10**, e70929. (doi:10.7554/eLife.70929)
128. Cagan R. 2013 San Francisco declaration on research assessment. *Dis. Models Mech.* **6**, 869–870. (doi:10.1242/dmm.012955)
129. Gomes DGE *et al*. 2022 Why don't we share data and code? Perceived barriers and benefits to public archiving practices. Figshare. (doi:10.6084/m9.figshare.c.6296319)