doi.org/10.22498/pages.26.2.54

Open-data practices and challenges among early-career paleo-researchers

Alexander Koch[1], K.C. Glover[2], B. Zambri[3], E.K. Thomas[4], X. Benito[5] and J.Z. Yang[6]

We conducted a survey on open-data-sharing experiences among early-career researchers (ECRs). While ECRs feel open-data sharing benefits their career, insufficient training in data stewardship presents a substantial challenge to data reusability.

Paleoclimate researchers readily acknowledge the benefits of open data, while identifying the need to improve best practices for data archival and sharing (Kaufman and PAGES 2k special-issue editorial team 2018). Growing data repositories are especially beneficial for ECRs, enabling the pursuit of synthetic, large-scale research questions from the start of their career. Fully implementing open-data practices throughout a project's lifecycle, however, remains time consuming and challenging.

We sought to understand how these challenges relate specifically to ECRs, and summarize here the results from a recent survey. Our survey was designed around the following questions:

- What challenges do ECRs face in following open-data practices?
- Do ECRs perceive open-data practices as advantageous?
- How can open-data practices enable ECRs' long-term scientific objectives?

While open-data practices are overwhelmingly perceived as advantageous for both one's long-term career and the advancement of science, our results highlight that the largest challenges to ECR implementation include unfamiliarity with community norms, and a lack of training and support. This perspective should inform the community's work towards greater standardization and rigor for open-data-sharing practices.

Methods

The anonymous survey consisted of 30 multiple-choice and free-response questions (see Suppl. Information). We wrote questions to target concerns raised in an ECR forum on open-data experiences (PAGES Early-Career Network 2018), and in consideration of the interactive discussion phase of the PAGES 2k Network open-data-implementation-pilot manuscript in the journal *Climate of the Past* (Kaufman and PAGES 2k special-issue editorial team 2018). Here we define ECRs as non-tenured survey respondents, since achieving tenure is unlikely within five years after PhD completion. We used Qualtrics as our survey platform, and disseminated the survey via paleoscience listservers (e.g. ECN-list; pmip-announce; paleoclimate-list; paleolim-list; Ecolog-list), Twitter, and word of mouth. The survey was open for 17 days, from 31 May to 17 June 2018.

Survey results and implications

Demographics

A total of 183 respondents completed the survey, with 163 identifying as non-tenure. The majority of respondents are students (38%) and postdocs (42%) from Europe (55%) and North America (33%; Fig. 1). Most respondents work with terrestrial (37%) or marine records (27%), or numerical models (23%). A larger proportion of respondents primarily collects or generates data (88%), rather than solely reanalyzing existing datasets (11%), for their research. Respondents commonly characterize their work as driven and dependent on quantitative data (60%). We use the survey results from the 20 tenured respondents as a point of comparison throughout the discussion below.

Data-sharing experience, opinions, and challenges

To facilitate reproducible science, Wilkinson et al. (2016) propose that published scientific data should be Findable, Accessible, Interoperable and Reusable (FAIR). Yet most non-tenured respondents (84%) are unfamiliar with the FAIR guiding principles for data management, a substantially higher proportion than in the tenured group (65%).

Tenured and non-tenured respondents equally feel that data (both 100%), metadata (both 90%) and code (e.g. data-analysis scripts; tenured: 65%; non-tenured: 70%) should be made publicly available and the proportion of respondents who regularly archive open data steadily increases from students (20%) to tenured researchers (80%; Fig. S10, supplementary information). More than two-thirds in all response groups most commonly utilize open databases or journal supplements (tenured: 72%; non-tenured: 65%) followed by personal or institutional databases (tenured: 18%; non-tenured: 12%, Fig. S11).

All respondents reported that a lack of metadata, inconsistent formatting, and data that are not centralized, not digitally available, or paywalled remain top challenges (Fig. S8). Yet, our results highlight that this problem may start at the ECR career stage: over half of the non-tenured

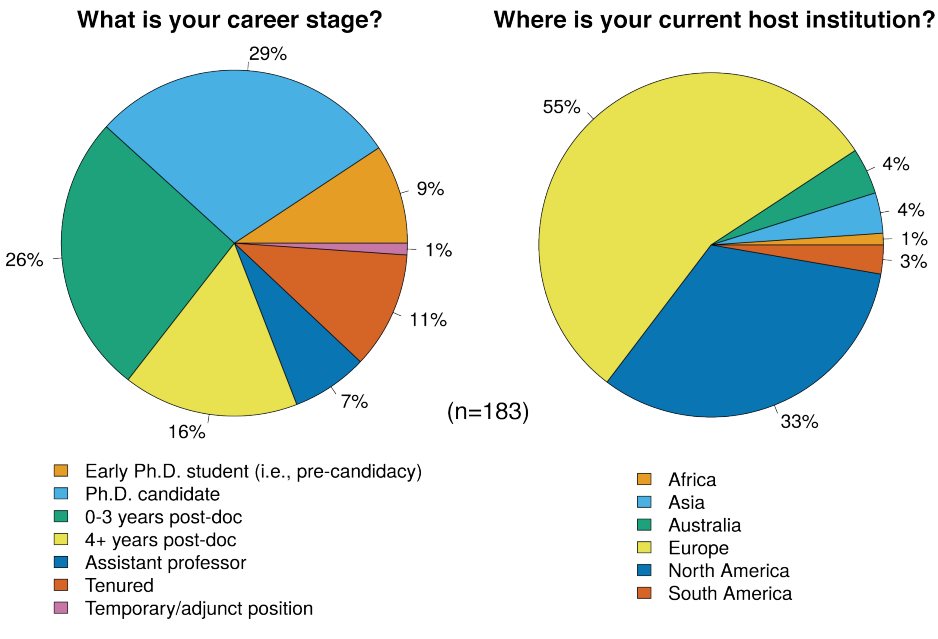

Figure 1: Selected survey demographics.

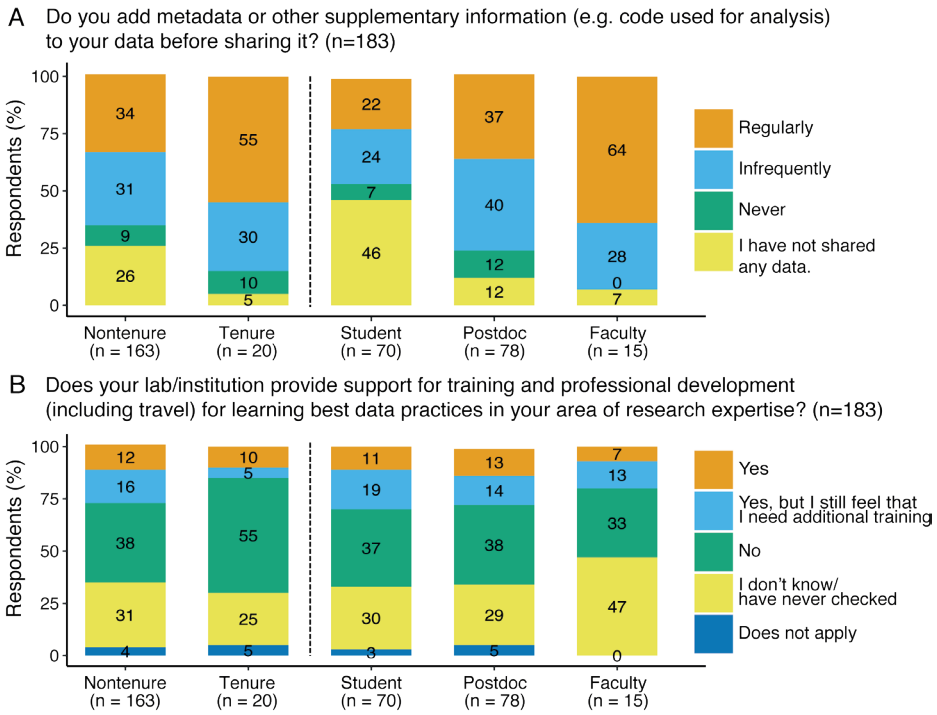

Figure 2: Selected survey responses (%; n=183) grouped by research experience. Non-tenure encompasses student, postdoc and faculty. Results for all survey questions available in the supplement.

respondents indicated "never" (12%) or "infrequently" (45%) adding metadata and code of their own to datasets, compared to 42% tenured respondents (Fig. S12). Our question on data-archival experience (Fig. S12) also reflected this split between ECR stages. If we eliminate respondents who answered "none of the above" because they had not yet published data, students were the largest group to report that the data-archiving process was difficult and the data archive they used lacked metadata templates, tutorials and upload scripts (63%). By comparison, tenured and later-stage ECRs noting this lack of guidance were less (22% each). Thus, unfamiliarity with metadata conventions and data-sharing standards may perpetuate the very problems that respondents identified in existing open datasets.

Data-sharing resources and training

The most common resources allocated to data sharing are time (tenured: 36%, non-tenured 19%) and staff help (tenured: 12%; non-tenured: 16%; Fig. S18). Over a third of the respondents that work in a lab (tenured: 36%; non-tenured: 48%) report that their lab is working towards standard operating procedures (SOPs), suggesting that labs do recognize a need for SOPs for data formatting and sharing. This is particularly important as our survey results signal that the most widespread issue may be related to labs without such SOPs (non-tenured: 89%, tenured: 78%; Fig. S17). More tenured (80%) than non-tenured (69%) respondents work in labs or institutions that offer no support for learning best practices for data sharing, or are not aware whether such support is available (Fig. 2b). Additionally, of the respondents who received training (26%), more than half feel that they need additional training.

Summary and recommendations

It is clear that the community recognizes the positive outcomes of an open-data culture: 95% of all non-tenured respondents and 90% of all tenured respondents feel that data sharing is advantageous to their career. However, equally pervasive are the difficulties surrounding open-access data preparation and publication as well as obtaining metadata-supported data (open-access or otherwise). Specifically, the lack of SOPs and institutional support paired with the unfamiliarity of best practices such as the FAIR guiding principles pose a challenge to data reusability. These benefits and challenges were widespread at all career stages.

Our survey targeting ECR practices and concerns highlighted that open-data usage tends to expand with career progression. We attribute that to researchers becoming more habituated to data-sharing procedures as they advance in their PhD programs, and career. Yet, we also found challenges unique to the ECR career stage:

- steep learning curve for new practitioners;
- widespread unfamiliarity with alternative data-sharing options such as data embargoes.

What can our community do to address these challenges for ECRs, and better promote open-data norms? ECRs working for senior (tenured) researchers may be in the position where their mentor is unfamiliar with the latest data-stewardship best practices, and thus either simply follow their mentor's practices, or must independently find other resources to support good data-sharing practices in their own work. Our survey results, however, suggest that data-management training initiatives (e.g. those offered by the Belmont Forum and Data Tree) are not widely used nor known. We therefore recommend dedicated community-led efforts to raise awareness and promote available training in data stewardship. Additionally, a continued discussion within the community regarding ways to motivate senior researchers and institutions to embrace community-wide data-sharing practices and SOPs will be key for establishing a culture of training ECRs in good data stewardship.

We therefore offer the following recommendations:

(1) Highlight existing resources, including FAIR, embargoes, and training available to ECRs (and other researchers).

(2) Encourage community efforts to the use of best practices in data stewardship and SOPs among ECRs, senior researchers and institutes.

We believe that the PAGES Early-Career Network (pastglobalchanges.org/ecn)can play an integral role in this movement by providing a platform for discourse within the community and a resource for data-stewardship training initiatives.

ACKNOWLEDGEMENTS

Our questionnaire was generated using Qualtrics software, Version May, 2018. Qualtrics and all other Qualtrics product or service names are trademarks of Qualtrics, Provo, USA.

SUPPLEMENTARY INFORMATION

Access the whole survey summary here: doi.org/10.22498/pages.26.2.54

AFFILIATIONS

[1]Department of Geography, University College London, UK
[2]Climate Change Institute, University of Maine, Orono, USA
[3]Department of Environmental Sciences, Rutgers University, New Brunswick, USA
[4]Department of Geology, University at Buffalo, USA
[5]National Socio-Environmental Synthesis Center (SESYNC), University of Maryland, Annapolis, USA
[6]Department of Communication, University at Buffalo, USA

CONTACT

Alexander Koch: alexander.koch.14@ucl.ac.uk

REFERENCES

Belmont Forum (2018) Retrieved August 6, from bfe-inf.org/action-theme-4-capacity-building-human-dimensions

Data Tree (2016) Retrieved August 6 from datatree.org.uk

Kaufman D, PAGES 2k Special Issue Editorial Team (2018) Clim Past 14: 593-600

PAGES Early-Career Network (2018) Retrieved July 6, from groups.google.com/forum/#!topic/pages-early-career-network/rOp6Hc7J6fc

Wilkinson MD et al. (2016) Sci Data 3: 160018