Clim. Past Discuss.,
https://doi.org/10.5194/cp-2017-157-AC1, 2018

Climate of the Past Discussions

Interactive comment on "Technical Note: Open-paleo-data implementation pilot – The PAGES 2k special issue" *by* Darrell Kaufman and PAGES 2k special-issue editorial team

Darrell Kaufman and PAGES 2k special-issue editorial team

darrell.kaufman@nau.edu

Received and published: 21 February 2018

Reply to comment by Karoly and others

D. Kaufman and the PAGES 2k Special Issue Editorial Team (Nerilie Abram, Michael N. Evans, Pierre Francus, Hugues Goosse, Hans Linderholm, Marie-France Loutre, Belen Martrat, Helen V. McGregor, Raphael Neukom, Scott St. George, Chris Turney, and Lucien von Gunten)

We (the PAGES 2k Special Issue Editorial Team) chose the interactive, public-discussion platform of *Climate of the Past* to trial our data stewardship initiative be-

cause we recognize that advancing open-data policies requires community discussion. The comment by David Karoly, Kathryn Allen and Patrick Baker was what we were looking for: a frank debate about translating data policies into practice. We thank the commenters for expressing their views and for exposing the contention that data policies often elicit, and we appreciate their general "*support . . . for open data sharing and transparency in the palaeoclimate sciences.*"

The PAGES 2k special issue was an opportunity, for those who opted to participate, to work toward a higher standard of data stewardship, a goal shared by the journal publisher, who endorsed our enforcement of the data policy. The process was new, and everyone (especially us) learned from the experience, including ways to improve. For this reason, we prepared the Technical Note that describes the procedure we used to pilot a strict and systematic interpretation of the journal's data policy within the framework of the PAGES Data Stewardship Activity[1]. The Technical Note was not meant to reiterate the merits of open-data policies, which are described in articles cited in the Note (e.g., the FAIR Guiding Principles[2]). Nonetheless, we agree with the commenters that, "*it is important to discuss the disadvantages as well as the advantages of open-data sharing.*"

I. The commenters present three disadvantages of an open-data policy:

(1) *"The impact of rigid data policies. . . will generally be negative on early-career researchers who are often working to schedule around their PhD study and cannot as rapidly produce the final products of their work. . . This leaves them vulnerable . . . and may compromise . . . their postgraduate studies. . ."*

Response: We fully agree that the interests of early career researchers (ECRs) must be protected. Whether open-data policies are generally negative for ECRs is debatable: Open data can (and has, based on the experiences of early and mid-career

[1] http://pastglobalchanges.org/ini/int-act/data-stewardship
[2] http://www.nature.com/articles/sdata201618

scientists involved in the PAGES Data Stewardship Activity) lead to new collaborations, greater impact and citation of their findings, and demonstrated compliance with current university and funding-body data policies, which in turn aids the scientist's case for future employment and funding support. Following open-data best practices demonstrates that an ECR values scientific transparency and reproducibility and has the knowledge to advance this scientific trend.

We understand that releasing data relinquishes control, but we stand with the affirmations of professional scientific organizations that reproducible science requires that the underlying data (and essential metadata) be released as part of a peer-reviewed study. The commenters might be referring to the additional time needed to prepare a dataset for delivery to a repository, relative to the short timeline of a PhD program. We contend that the additional work is minor relative to the effort involved in publishing and that is most efficient when done as part of preparing a publication. The process becomes easier as streamlined workflows are developed.

(2) "*A data policy should also respect the fact that . . . funding . . . comes from agencies. . . that expect a return on their investment from the funded group, not from a different group. Rather than a didactic top-down approach from data modellers or synthesizers who clearly require palaeoproxy data for their use, a data policy should be formulated with input from those who generate the records.*"

Response: We agree that data generators should not turn their data over to another group in a way that diminishes the credit received for their contribution. We support – and many of us are among – data generators who are pushing for widespread adoption of data citations[3] as a mechanism for data generators to receive credit on par with a bibliographic citation. A major motivation for the data activity behind the PAGES 2k special issue was to foster the use of data citations. The editorial team worked with authors to help them include data citations in their papers. In the case of the dataset referred

[3] https://www.force11.org/datacitationprinciples

to by the commenters, we reached out to the data generators and offered to *"facilitate the transfer of your data to NOAA or other long-term archive so that you can receive a persistent identifier, which can be used for a proper data citation in your name."* Most of the data generators agreed because this supports, rather than diminishes, the credit that they accrue. The citation, re-use and development of new applications of publicly available results of funded research is in fact generally encouraged, a quantifiable credit to both the investigator and the agency.

(3) ". . . *we would not have included some data sets that were generously shared with us on the basis that they not be made publicly available yet. The group who generated these data sets was working towards their publication as part of an ongoing broader grant-funded project. This situation unnecessarily placed the PhD student writing this paper in an extremely difficult and stressful position.*"

Response: The commenters' paper was more difficult in regards to data stewardship than any other in the special issue because, unlike most every other paper, some of the previously published data used in their study were not available through public repositories. We regret that this process led to a stressful situation. We believe that it could have been avoided through discussions between the data generators and the authors during the data-gathering phase. In this case, it appears that the reluctance of data generators to make their data openly available should have precluded their inclusion into the dataset used in the commenter's paper. We contend that data syntheses need to be based on publicly available data and that studies that rely on unpublished or nonpublic data rightfully risk skepticism, because results cannot be verified. Ideally, a dataset should be vetted through peer review, and stand on its own merits before it is accepted as input to someone else's publication. We recognize that there are cases when it is appropriate for a study to include data from the public domain that lack an associated peer-reviewed publication, especially for relatively simple types of data, or for data that come with sufficient metadata for intelligent reuse and are attributable to the original data generator through a proper data citation. While offering unpublished

data for use in an another study might seem generous, use restrictions attached to the data can make reliance on unpublished data counterproductive. We advocate the use of data-oriented publications, such as *Scientific Data,* as an alternative means for data generators to obtain credit through subsequent citations of their data.

II. The commenters disagree with our interpretation of the journal's data policy:

(1) The journal's data policy "*recommends depositing data . . . in reliable (public) data repositories*" whereas we *"required"* this.

Response: In our view, following a policy requires adopting its recommendations, and the purpose of our activity was explicitly to apply a strict and systematic interpretation of the policy, consistent with the COPDESS statement referred to in the journal's data policy[4]. More important than semantics, the journal policy states, "*If the data are not publicly accessible, a detailed explanation of why this is the case is required.*" We conveyed this option to the data generator and to the commenters, "*If extenuating circumstances preclude the public release of data used in a paper, the reason must be clearly stated.*" In response, the data generator, who was not part of the author team, explained, "*Releasing data from current projects that have yet to finish (or for unfunded projects) runs the real risk of being mined and so diluting my chances of future support."* We did not think that this explanation was appropriate for publication in the data availability section of the article. More importantly, the data in question had already been published and cited as such by the authors, so the issue was the availability of a digital version of data already used in a previous publication, which is difficult to justify withholding. We agree that special circumstances can arise that must be considered on a case-by-case basis and we remained open to such explanations from authors throughout the process.

(2) *"No option was provided by the Team to allow for publicly accessible data sets with a doi through journal supplementary material, which is allowed by the journal."*

[4]https://www.climate-of-the-past.net/about/data$_policy.html$

Response: Copernicus Publication does not offer data-curation services and supplements are archived only on their in-house servers. Unlike articles, supplements are not sent to long-term archives. The journal's data policy instead asks authors to use a data repository that is registered through re3data.org. When asked about this by the authors during the manuscript review, we clarified the policy.

III. The commenters take issue with our procedures and assertions:

(1) The notice of our intention to enforce the journal's data policy was not sent to authors until after some papers had been submitted.

Response: We agree that an earlier reminder of the journal's data policy and our data stewardship initiative for this special issue would have been better, and will include this lesson learnt in our revisions of the Technical Note. Of the 17 papers accepted to the special issue, two had been submitted prior to our message conveying explicit instructions to corresponding authors. We assumed that the open-data policies for the special issue were clear from the outset because, beginning with the first PAGES 2k Circular[5] in 2010, and nearly every year since, the PAGES 2k community has recognized the need to make all data used in its products available publicly, a guideline that applies to all PAGES working groups. More importantly, we were unaware that the commenters were not in agreement with our instructions that "*all of the essential data used in the analyses and generated as results must be made available through a public data repository where they can be tracked using a Data Citation.*" In fact, in their reply[6] to our comments, the authors stated that they would "add all data citations" linked to the digital data used in their study. This was reassuring and, in the end, was done by the authors.

(2) The terms "*top-down*" and "*bottom-up*" were misused.

Response: We agree that, as editors of the special issue, our role was to require

[5]http://www.pages-igbp.org/ini/wg/2k-network/intro

[6]https://www.clim-past-discuss.net/cp-2017-28/cp-2017-28-AC3.pdf

authors to follow through on their replies to reviewer's comments, as well as to follow the intent of the journal's and PAGES' data stewardship statements. On the other hand, the PAGES 2k principles of data archiving that we enforced have been developed over many years and in consultation with a large interdisciplinary group of paleo-data generators, and they are shared by the global scientific community. Nonetheless, in the revisions to the Technical Note, we will replace the ambiguous terms with more descriptive text.

(3) "*. . .manuscripts would have to be withdrawn if they did not comply. . .*" with the data policy.

Response: We gave authors of the special issue papers the option of complying with the data policy (as applied by the PAGES 2k special issue), transferring the manuscript out of the special issue, or withdrawing their paper. The one paper that was not able to meet the strict open-data requirements being piloted in the special issue was simply transferred out of the special issue and into a regular issue of the journal, and little time was lost.

(4) Our statement that "*the practice of making data readily available is rarely embraced*" neglected to recognize the major efforts to secure our community's data resources.

Response: We agree; however, on the basis of the commenters' concerns, it's fair to say that making data readily availability is "not always" embraced. We will revise the statement.

IV. Key points

It's important to keep the two issues separate: (1) whether new data should be made available at the time of publication, and (2) whether previously published but not-yet-publicly-archived data should be used in a synthesis study. The datasets referred to by the commenters relate to the second issue. The data had been attributed to previous publications; they were forwarded to the authors for their synthesis study with the

provision that they not be released. For many years, the PAGES 2k community has reaffirmed the requirement that its data products must be based on publicly available data. The authors should not have accepted publicly withheld data into their PAGES 2k data compilation if the data generators were unwilling to publicly archive the dataset.

Regardless of one's views on open-data sharing, reproducibility is fundamental to scientific integrity and progress. It is only possible if the data used to generate a result are publicly available. One goal of the PAGES 2k special issue was to apply a rigorous interpretation of the journal's data policy, an activity agreed to by the participants as evidenced by their response to our comments in the open discussion phase of the paper reviews. Any authors who preferred to not follow the recommended policies could have moved their paper to a regular issue of the journal. Lacking publicly available data, they would not have been considered part of the PAGES product. While we do not endorse looser data polices, we understand that our approach to implementing them is not yet standard practice in paleoscience.

The intent of our implementation pilot and its documentation in the Technical Note is to help move the community in the direction of open data. The intent was not to extract data from our colleagues, but to create a product that gives explicit credit to the data producers, with greater exposure and citation of their work. We believe that this approach will promote transparency and reproducibility and will help accelerate scientific discovery through the advancement of science in emergent transdisciplinary directions. We stand with Copernicus Publications and the many other high-level international signatories in their commitment[7] to data stewardship as articulated by the Coalition on Publishing Data in the Earth and Space Sciences (COPDESS). Our community's "*data are special resources, critical for advancing science and addressing societal challenges.*" We concur with COPDESS that "*scholarly publication is a key high-value entry point in making data available, open, discoverable, and usable.*"

[7] http://www.copdess.org/statement-of-commitment/