Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157
Manuscript under review for journal Clim. Past
Discussion started: 19 December 2017

Climate of the Past Discussions
Open Access
EGU

Technical Note:

Open-paleo-data implementation pilot – The PAGES 2k special issue

Darrell S. Kaufman[1] and PAGES 2k special-issue editorial team[*]

[1] School of Earth Sciences & Environmental Sustainability, Northern Arizona University, Flagstaff, USA.

[*] A full list of authors and their affiliations appears at the end of the paper.

Correspondence to: D. Kaufman (darrell.kaufman@nau.edu)

Abstract. Data stewardship is an essential element of the publication process. Knowing how to enact generally described data policies can be difficult, however. Examples are needed to model the implementation of open-data polices in actual studies. Here we explain the procedure used to attain a high and consistent level of data stewardship across a special issue of the journal, *Climate of the Past*. We discuss the challenges related to (1) determining which data are essential for public archival, (2) using previously published data, and (3) understanding how to cite data. We anticipate that open-data sharing in paleo sciences will accelerate as the advantages become more evident and the practice becomes a standard part of publication.

1 Introduction

The benefit of open data for accelerating scientific discovery is widely accepted, but the practice of making data readily available is rarely embraced. The paleo-science community is currently in the early stages of adopting the relatively new open-data policies established by major publishers[1,2], research funders[3,4], and government agencies[5] (citations selected from among many). Few examples are available that model best practices for open paleo data over the entire lifecycle of a research project, including scientific publication, and best practices are evolving rapidly along with the revolution in cyber-enabled data sharing. Considering the vast variety of data types inherent in paleo sciences, knowing how to apply generalized data policies in real situations is not always obvious. Funding agencies and journals can set requirements, but top-down mandates alone are unlikely to foster the necessary cultural changes in scientific communities (Kattge et al., 2014; McNutt et al., 2016). Bottom-up motivation from the research community, including early-career scientists, is needed to drive the open-data revolution (Molloy, 2011; Lemprière, 2017).

Publication is the ideal stage in the progression of a research project for concerted data management and stewardship. Authors striving to enhance the impact and visibility of their study are receptive to input from peer reviewers

[1] https://publications.agu.org/author-resource-center/publication-policies/data-policy/
[2] http://www.nature.com/authors/policies/availability.html
[3] https://www.nsf.gov/bfa/dias/policy/dmp.jsp
[4] https://erc.europa.eu/funding-and-grants/managing-project/open-access
[5] https://www2.usgs.gov/datamanagement/share.php

Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157
Manuscript under review for journal Clim. Past
Discussion started: 19 December 2017

and journal editors. Articles that make the underlying data easily discoverable and reusable have been shown to benefit from a higher rate of citation (Piwowar and Vision, 2013). Without proper curation in a public repository, the ability to locate data from previously published studies diminishes rapidly after publication, even for articles that purported to make associated data available (Vines et al., 2014). Scientists are most familiar with the specific details of their data when submitting a manuscript for publication and therefore are better able to prepare the data for transfer to a public repository at that moment rather than following publication when the familiarity and incentive have faded.

As the paleo-community's data resources continue to grow, compilations that summarize and curate large datasets are becoming increasingly useful (Kaufman, 2017). The Past Global Changes (PAGES) project is an international, community-driven effort that facilitates the development of data-intensive synthesis products in paleo sciences, and promotes long-term care of the community's data resources. PAGES research is primarily conducted by working groups whose primary scientific goals are often based on a major, original data synthesis product (e.g., PAGES 2k Consortium, 2017a,b). Themes that involve multiple working groups are coordinated through PAGES' "integrative activities," of which data stewardship is currently one. It has developed guidelines for open-data sharing[6] and has facilitated the creation of data products that follow the "FAIR Guiding Principles" for data management and stewardship (Findability, Accessibility, Interoperability, and Reusability; Wilkinson et al., 2016)

The purpose of this technical note is to describe the procedure used to implement the PAGES' data-sharing policy and to attain a high and consistent level of data stewardship throughout the special issue of *Climate of the Past*, "Climate of the past 2000 years: regional and trans-regional syntheses[7]," hereafter, the PAGES 2k special issue. This same procedure was subsequently implemented through the PAGES Young Scientists Meeting[8] special issue of the journal. We highlight key challenges and approaches to overcoming the challenges. Rather than discussing the merits of open-data policies, we focus on an example of data stewardship implementation.

2 Procedure for implementing the open-paleo-data policy

2.1 Reaffirm the intent and goals of relevant policies

The editorial team of the PAGES 2k special issue viewed the compendium as an opportunity to work with motivated authors to model best practices of data management and stewardship associated with publication, despite the additional work required for both authors and editors. To begin, the goals and procedures of the data stewardship activity were discussed with the supportive journal publisher, Copernicus Publications. A notice was sent to prospective authors of the special issue to explain the rationale and process for the data stewardship effort and to alert them to the relatively high standard as stated

[6] http://pastglobalchanges.org/ini/int-act/data-stewardship
[7] https://www.clim-past.net/special_issue841.html
[8] https://www.clim-past.net/special_issue912.html

Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157
Manuscript under review for journal Clim. Past
Discussion started: 19 December 2017

in the journal's (*Climate of the Past*)[9] and PAGES'[6] data policies. Submissions that did not include the required data availability statement were returned to authors with guidelines on how to prepare the statement prior to resubmitting.

2.2 Review manuscripts for data accessibility

Once accepted to the Discussion phase, the editorial team reviewed the data contained in each manuscript. Generally, the most important data were presented in figures and tables, providing a focus for the data review. The editorial team identified the most relevant datasets of two general types: (1) data described in previous publications, including data-reanalysis products, and third-party unpublished data ("input data"), and (2) data generated as an outcome of the study ("output data"). The editorial team used the information presented in the data availability statement and elsewhere in the manuscript to locate the datasets as described, sometimes unsuccessfully. Missing or inaccessible data were specified in a data-review comment, which was prepared by the editorial team, then posted as part of the public interactive discussion. Authors were asked to transfer all datasets that were not easily accessible to a community-recognized public data repository, and to cite the dataset using the persistent identifier issued by the repository. Authors were asked to reply to the data-review comment as part of the public interactive discussion, as a matter of record for the publication. The data-review comments and author replies for each Discussion paper of the PAGES 2k special issue are available at the journal's website[7].

2.3 Work with authors to meet the data-stewardship standards

A few authors expressed concerns about releasing data provided to them from other scientists or about making digital versions of their output data available. These views were among those revealed by extensive surveys that explore data-sharing practices and attitudes among scientists (Schmidt et al., 2016; Tenopir et al., 2011). The vast majority of authors were agreeable to making the modifications specified by the editorial team. Once manuscripts were revised to address the data-review comments, they were again evaluated by the editorial team, with the primary focus on the data availability statement and the data citations. Papers were accepted only after all of the input and output data had been properly archived and reviewed for completeness by the editorial team. This required coordination between the author, editor, publisher, and data repository, with somewhat different procedures used for each case. Some repositories allow reviewers to privately access submitted data prior to public release, others can set up a shell to receive the data, and the data can be circulated by the author to reviewers and the repository in preparation for publication. Whatever the exact procedure, the intent is for authors to make their datasets available to reviewers and to transfer their data to a public repository prior to final acceptance of the article.

[9] https://www.climate-of-the-past.net/about/data_policy.html, https://publications.copernicus.org/services/data_policy.html

Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157
Manuscript under review for journal Clim. Past
Discussion started: 19 December 2017

3. Challenges

3.1 Determining which data are "essential"

Deciding which data are important to make available for reuse was not always obvious. Not all of the data presented in a publication are necessarily essential for reproducing the study results, and some might not be useful for future researchers. While there may be unforeseen potential uses of data, requiring authors to archive data that have little bearing on the primary outcome of a study can be onerous and pointless. Instead, the utility of each dataset must be evaluated in context of the unique contribution of the particular study. Discussions among the editorial team, which included disciplinary specialists, were frequently needed to judge the significance of particular datasets and to prioritize the request for data management.

Output data. The most important outcomes of the study were usually clear, and the resulting datasets that comprise those outcomes were easy to identify. They constituted the data that future users might need to compare with other similar data, or to test their sensitivity to different assumptions, or to incorporate in a future data synthesis. Some manuscripts included various renditions of the output data, such as the outcomes of different sensitivity experiments or different data-processing routines, such as levels of smoothing. In these cases, the version that was favoured by the author and the one that most likely would be reused in future research was identified as the top priority for archival. Studies often presented multiple datasets as the primary outcome; in this case, the entire suite of data was submitted to a data repository and a single landing page was used to organize the various datasets as a package under one link. In addition to the data themselves (often time-series or spatial data), essential metadata needed to improve discoverability and facilitate the accurate reuse of the dataset were required to be included. What constitutes essential metadata for paleo data varies for different data types and purposes, and is different for new publications versus legacy data. An effort to advance paleo-data standards is underway through the NSF-EarthCube-supported LinkedEarth project[10] (Khider et al., 2017). Policies and standards for the availability of computer code are less developed, but are being addressed by some modelling-focused publications (e.g., Copernicus' journal, *Geoscientific Model Development*[11])

Input data. The most important input data used in a study was often less obvious. If a particular dataset was necessary to reproduce the primary results of the study, it was considered essential and therefore necessary to make available through a public repository. Most of the results from previous studies that were mentioned as supporting information, but not analysed or featured graphically, were deemed as non-essential to the primary outcome of a study and therefore a bibliographic citation sufficed. Most papers included a discussion and figure that compared the outcome of the study with previously published data. The original publications were cited, but their underlying digital data were not always available. In this case, the editorial team considered whether the implications of the data comparison were central to the conclusions of the study, or

[10] http://linked.earth
[11] https://www.geoscientific-model-development.net/about/code_and_data_policy.html

Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157
Manuscript under review for journal Clim. Past
Discussion started: 19 December 2017

in the case of review papers, whether the data summaries would be valuable for future researchers. If so, the editorial team asked the authors to make the previously published data available through a public repository.

3.2 Using data published by other researchers

Using data that are already available through a public repository is straightforward; relying on non-archived data, in contrast, hinders transparency, reproducibility and reusability. This is common practice in paleo sciences because only some paleo data are presently available through public repositories. The studies describing the data are published and therefore the data are technically public, at least in graphical form. Rather than digitizing the data from the original publication and archiving the second-hand, degraded version, the editorial team asked authors to be stewards of the data assets that they used in their study by working with the original data generators to transfer the data and associated metadata to a public repository, with the goal of enhancing their discoverability and intelligent reuse by downstream scientists. In a few cases, especially those involving early-career lead authors and reluctant data generators, the editorial team assisted with the data transfer.

In one example of authors engaging data generators, Shuman et al. (2017a) synthesized data from 93 previously published proxy records of North American hydroclimatology, over half of which were not available through public archives. The digital data had been provided by the original data generators for the purpose of the synthesis. The authors compiled the data along with essential metadata gleaned from the published articles describing the datasets. They input the (meta)data using the Linked PaleoData[12] structure (LiPD; McKay and Emile-Geay, 2016). Once contained in LiPD format, the (meta)data could be automatically translated to the template used by the World Data Service (WDS) for Paleoclimatology (hosted by NOAA). These files were then sent to the original data generators for validation before transferring them to NOAA-WDS Paleoclimatology for long-term curation. Each dataset received a unique and persistent identifier, which was cited to credit the original data generators. In addition, as part of the contribution of their synthesis, Shuman et al. (2017b) compiled all 93 proxy records, including the metadata needed for accurate and consistent reuse of the data, into a single package of machine-readable files, including several serializations. These were posted on the landing page[13] along with the primary outcome of the synthesis.

Another approach to engaging data generators is a data publication. These focus on the data compilation itself, including a description of the dataset and the procedures used to assemble it, rather than scientific interpretation of the data. Authorship on such products can be inclusive of all those who contribute and validate the data and metadata. The data product can then be cited to credit data generators, and it can be used as the basis for addressing research questions by the community. The PAGES 2k Consortium (2017a) temperature dataset is a recent example.

[12] http://lipd.net
[13] https://www.ncdc.noaa.gov/paleo/study/22732 (*note to reviewers: link will be active soon*)

Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157
Manuscript under review for journal Clim. Past
Discussion started: 19 December 2017

3.3 Understanding how to cite data

Data are fundamental products of research; citing their source, like all other sources of information, is good research practice (CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; Data Citation Synthesis Group, 2014; Starr et al., 2015). Data citations facilitate open-data sharing and they assign credit to the data generator, which might be someone other than the author of the associated article. While data citations are gaining traction in the research community, most paleo-scientists have not used them and most paleo-oriented journals have not yet established specific instructions to authors for their style and use. For the PAGES 2k special issue, authors were asked to include two citations to credit previously published essential input data: (1) a bibliographic reference where the data are described, and (2) a data citation that points to the version of the dataset as it is stored in a public online server. For datasets that relied on data-processing tools to access, plot or analyse the data from a data repository, authors were also asked to include a citation to the code or software. Authors were also asked to provide a data citation for the primary outcome of their study.

A data citation is different than a traditional bibliographic reference, which cites the article that describes the dataset. Instead, a data citation refers to the location of the original data as lodged in a pubic repository (e.g., PAGES 2k Consortium, 2017b). At the core of a data citation is a persistent identifier, which is machine actionable and globally unique (Fenner et al., 2016). The identifier is typically a DOI number, or a URL link issued by the NOAA-WDS Paleoclimatology. Other URL links that reference data or data-access tools on the Internet often brake after a short time (Dellavalle et al., 2003) and should not be used. Data contained in supplementary information attached to an article does not suffice as a data citation because it might require a journal subscription to access, and publishing companies are not usually recognized as secure long-term data repositories (e.g., a member of the ICSU World Data System[14]). Sometimes, more than one article describes the same dataset, in which case a data citation is needed to resolve the providence of the data. Sometimes, datasets evolve as additional information is gathered and the original data are upgraded or modified. The major paleo-data repositories offer a means for versioning of datasets, and the version can be specified as part of the data citation, although this practice is and its conventions are not yet well described.

Authors chose different approaches to citing data sources and the editorial team did not prescribe a specific style. Some authors included the DOI (or URL issued by NOAA-WDS Paleoclimatology) identifiers within a table that listed the datasets used in the study. This structured format enables future users to conveniently access datasets directly from a list rather than having to locate the DOI (or URL) identifier within the references-cited section. On the other hand, including the data citations in the reference list makes it easier for publishers to capture and processes the citations in the same way as other references (Cousijin et al., 2017). In the future, it may be possible for automated data harvesters to recognize and resolve DOI (or URL) identifiers from within the text outside of the reference list. Other authors used the familiar in-text call-out (author, year) and integrated the data citation into the references-cited section. Some authors included both an in-text

[14] https://www.icsu-wds.org

call-out to the data citation and the DOI (or URL) within a table. Although somewhat redundant, this dual approach has merit.

The primary results of studies in PAGES 2k special issue were lodged at public paleo-data repositories, and data citations, or just the DOI (or URL) identifiers, that point to these new datasets were incorporated in the data availability statement. The data citations for most studies linked to a landing page at either NOAA-WDS Paleoclimatology[15] or PANGAEA[16] where multiple input and output datasets are listed. For some of the data-synthesis and review articles, authors added metadata to the input datasets to enhance the re-usability and formatted the data to improve machine readability. Rather than revising the original dataset, the modified datasets were included as part of the original contribution of the synthesis study and were listed on the landing page for that study, along with the major outcomes of the synthesis, sometimes in more than one digital format.

4. Outlook

Open-data sharing will accelerate as the advantages become more obvious and the practice becomes more widely expected. The advent of metrics that track when and where data have been cited and reused will enable data generators to quantify the impact of their scholarly product, further encouraging open-data sharing. Data stewardship is a natural element of the publication when editors and reviewers put established open-data policies into practice. Additional work is often required to prepare data with sufficient metadata for archival, but the effort is relatively modest compared to the work involved in the publication itself, and in consideration of the long-term benefits data stewardship. Keeping the focus of data sharing on the essential data that are most likely to be useful for future studies helps to avoid unnecessary effort for the sake of following a blanket data-sharing policy. Considering the wide variety of paleo-data types and the recency of open-data policies, more examples like the PAGES 2k special issue will help editors and authors to navigate the ambiguities and challenges associated with implementing data best practices.

Team members – PAGES 2k special-issue editorial team

Nerilie Abram (Research School of Earth Sciences, Australian National University, Canberra, Australia), Michael N. Evans (Department of Geology & ES-SIC, University of Maryland, USA), Pierre Francus (Institut National de la Recherche Scientifique, Centre Eau Terre Environnement, Québec, Canada and GEOTOP Research Center, Montréal, Canada), Hugues Goosse (Earth and Life Institute (ELI)/Georges Lemaître Centre for Earth and Climate Research (TECLIM), Université Catholique de Louvain, Belgium), Hans Linderholm (Regional Climate Group, Department of Earth Sciences, University of Gothenburg, Sweden), Marie-France Loutre (PAGES International Science Office, Bern, Switzerland), Belen Martrat

[15] https://www.ncdc.noaa.gov/data-access/paleoclimatology-data
[16] https://www.pangaea.de

(Department of Environmental Chemistry, Institute of Environmental Assessment and Water Research (IDÆA), Spanish Council for Scientific Research (CSIC), Barcelona, Spain), Helen V. McGregor (School of Earth & Environmental Sciences, University of Wollongong, Australia), Raphael Neukom (Oeschger Centre for Climate Change Research & Institute of Geography, University of Bern, Switzerland), Scott St. George (Department of Geography, Environment and Society University of Minnesota, Minneapolis, USA), Chris Turney (School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia), and Lucien von Gunten (PAGES International Science Office, Bern, Switzerland)

Acknowledgements. We thank the authors of the PAGES 2k special issue, Copernicus Publications, and the paleo-data repositories who worked with us to implement open-data policies. This is a contribution to the Past Global Changes (PAGES) Data Stewardship Integrative Activity. PAGES is supported by the US and Swiss national science foundations, and the Swiss Academy of Sciences. DSK is funded by NSF-AGS-1602105.

References

CODATA-ICSTI Task Group on Data Citation Standards and Practices: Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Science Journal 12, CIDCR1–CIDCR7. doi: 10.2481/dsj.OSOM13-043, 2013.

Cousijin, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., Clark, T.: A data citation roadmap for scientific publishers. bioRxiv. doi: 10.1101/100784, 2017.

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles: Martone M. (ed.), San Diego CA: FORCE11. doi: 10.25490/a97f-egyk, 2014.

Dellavalle, R.P., Hester, E.J., Heilig, L.F., Drake, A.L., Kuntzman, J.W., Garber, M., Schilling, L.M.: Going, going, gone: Lost Internet references. Science 302, 787-788. doi: 10.1126/science.1088234, 2003.

Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., Clark, T.: A data citation roadmap for scholarly data repositories. bioRxiv. doi: 10.1101/097196, 2016.

Kattge, J., Díaz, S., Wirth, C.: Of carrots and sticks. Nature Geoscience 7, 778-779, 2014.

Kaufman, D.S.: Bookkeeping or science: what's behind a paleo data compilation. http://blogs.nature.com/soapboxscience/2017/07/11/bookkeeping-or-science-whats-behind-a-paleo-data-compilation, 2017.

Khider, D., Emile-Geay, J., McKay, N., Garijo, D., Ratnakar, V., Gil, Y., Zhu, F.: LinedEarth and 21st century paleoclimatology: reducing data friction through standard development. American Geophysical Union fall meeting, New Orleans, 11-15 Dec., Abstract IN32A-03, 2017.

Lemprière, S.: Five things you can do today to make tomorrow's research open. http://blogs.nature.com/naturejobs/2017/11/22/five-things-you-can-do-today-to-make-tomorrows-research-open, 2017.

McKay, N.P., Emile-Geay, J.: Technical note: The Linked Paleo Data frameworks – a common tongue for paleoclimatology. Climate of the Past 12, 1093-1100, 2016

McNutt, M., Lehnert, K., Hanson, B., Nosek, B.A., Ellison, A.M., King, J.L.: Liberating field science samples and data. Science 351, 1024-1026, 2016.

Molloy, J.C.: The open knowledge foundation: Open data means better science. PLoS Biology 9(12), e1001195. doi: 10.1371/journal.pbio.1001195, 2011.

PAGES 2k Consortium: A global multiproxy database for temperature reconstruction of the Common Era. Nature Scientific Data 4, 170088. doi: 10.1038/sdata.2017.88, 2017a.

PAGES 2k Consortium: A global multiproxy database for temperature reconstructions of the Common Era. World Data Center for Paleoclimatology, https://www.ncdc.noaa.gov/paleo/study/21171, 2017b.

Piwowar, H.A., Vision, T.J.: Data reuse and the open data citation advantage. PeerJ 1, e175. doi: 107717/peerj.175, 2013.

Schmidt, B., Gemeinholzer, B., Treloar, A.: Open data in global environmental research: The Belmont Forum's open data survey. PLoS ONE 11(1), e0146695. doi: 10.1371/journal.pone.0146695, 2016.

Shuman, B.N., Routson, C., McKay, N., Fritz, S., Kaufman, D., Kirbey, M., Nolan, C., Pederson, G.T., St Jacques, J.M.: Millennial-to-centennial patterns and trends in the hydroclimate of North America over the past 2000 years. Climate of the Past Discussions. doi: 10.5194/cp-2017-35, 2017a.

Shuman, B.N., Routson, C., McKay, N., Fritz, S., Kaufman, D., Kirbey, M., Nolan, C., Pederson, G.T., St Jacques, J.M.: Millennial-to-centennial patterns and trends in the hydroclimate of North America over the past 2000 years. World Data Center for Paleoclimatology, https://www.ncdc.noaa.gov/paleo/study/22732, 2017b.

Starr et al.: Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1, e1. doi: 10.7717/peerj-cs.1, 2015.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M.: Data sharing by scientists: Practices and perceptions. PLoS ONE 6(6), e221101. doi: 10.1371/journal.pone.0021101, 2011.

Vines, T.H., Albert, A.Y.K., Andrew, R.L., Débarre, F., Bock, D.G., Franklin, M.T., Gilbert, K.J., Moore, J.-S., Renaut, S., Rennison, D.J.: The availability of research data declines rapidly with article age. Current Biology 24, 84-97. doi: 10.1016/j.cub.2013.11.014, 2014.

Wilkinson, M.D. et al.: the FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018. doi: 10.1038/sdata.2016.18 (2016).