Clim. Past Discuss.,
https://doi.org/10.5194/cp-2017-157-SC7, 2018

Climate
of the Past
Discussions
Open Access

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

J. Williams

jww@geography.wisc.edu

Received and published: 30 March 2018

I have read this article and discussion with great interest: it speaks to fundamental questions about how we conduct open, transparent, and global-scale earth system paleoscience, in synthesis projects that draw upon the collected data, knowledge, and labor of many (dozens to thousands) of individual scientists, distributed around the world. My perspective is that of a scientist who has worked on many continental- to global-scale data syntheses over the years, as someone not involved in the PAGES 2K initiative, and as one of the leaders of the Neotoma Paleoecology Database (www.neotomadb.org), which seeks to support open data, building commu-

nities of Data Stewards, and global-scale paleoscience. Here I first reflect on four emergent themes in this discussion, then briefly note Neotoma's role within the open-data ecosystems that are emerging and the solutions that we are seeking to build.

1) Where are the ethical limits to open data? Open data is clearly a good: it enables transparency, reproducibility, and accountability in science (see comments by Simpson, Bothe, response by Kaufman et al.). Open data enables our field to move from our local-scale records, collected at great effort and cost, to a global-scale understanding of the earth system and its past dynamics. Open data and open workflows accelerate the pace of science and knowledge transmission, by enabling new advances developed by one research team to be quickly adopted by other researchers.

However, there are other goods. Karoly and Cook in their comments raise the important good of protecting early career researchers and their intellectual output. One can imagine an extreme open-data ethos that did not fairly account for this competing good: e.g. requiring all measurements made by an early career researcher to be instantly made available on-line and usable by all. This hypothetical extreme, in which an ECR provided the labor and others instantly reaped the fruits, clearly would carry the open-data ethos too far.

So, we need balancing mechanisms that both encourage open data and create first-use protections for the intellectual work by data generators and early career researchers. Embargoes, as suggested by Simpson, are one important mechanism, and a critical priority for our field should be developing better systems for creating and managing data embargoes. (We are beginning embargo development in Neotoma, see below.) A second mechanism should be to establish different open data norms for primary data papers versus large data syntheses. Primary data papers, that present new data collected and generated by a research team, should be held to a high standard of data openness and publication, with the general expectation that all presented data be contributed to a community open data repository. Large data syntheses, such as the PAGES 2K synthesis, that draw upon both published and unpublished records, need more of a tiered system that balances open data missions with protections for data generators. Such a tiered system would establish full openness for workflows that use published data, and partial openness for workflows that use unpublished data. Or, efforts such as PAGES 2K may simply opt for simplicity, use published data only, and thereby achieve full openness.

2) Where are the practical limits to open data? This point, raised by Bothe, falls under the general topic of data reduction. In practice we must always curate data and knowledge, in which we make decisions about which data are important and which are not. Some limits relate to data volume, e.g. the question by Bothe about earth system models and how best to store and share their large-volume outputs. Similarly, Bayesian modeling approaches generate large volumes of Monte Carlo Marko Chain (MCMC) traces that are used to generate posterior probability estimates (Blaauw and Christen, 2011; Dawson et al., 2016; Parnell et al., 2016). Should the full MCMC traces be stored, or simply the summary statistics? Some data reductions or transformations are motivated by scientific convention. For example, radiocarbon dates are usually reported as estimated ages, rather than the primary measurements of count statistics for individual isotopes. Some data reductions occur because science data collection operates at the real/virtual interface, with some information easily captured and shared with others (e.g. primary data tables, instrumental outputs, photos) and others less so (e.g. field notebooks, lab notebooks, personal experience, judgment, and decision-making). In general, the advances in data science will make it possible to extend data openness to information that was previously impractical to share (e.g. raw data output from geochemical instrumental systems). But there will always be limits to what can be made readily open, and hence some need for expert judgment and community norms about data curation and sharing.

3) Paleodata are high-effort and therefore high-value. Our proxy records are collected at great cost: they usually require days to weeks of fieldwork in remote locations, months to years of labwork, and months to years of analysis and interpretation by

highly trained experts. Most of this work is supported by public taxpayers via scientific foundations. Often, these data cannot be collected again, for economic reasons (costs associated with field and labwork) and physical reasons (many archives are now lost or in danger of being lost). To me, one of the strongest arguments for open data is as our field's bulwark against data entropy (Michener et al., 1997) and knowledge loss (Jackson, 2012). Our sum total of scientific knowledge can be viewed as a dynamic balance between rates of knowledge and data generation versus rates of knowledge and data loss. The peer-reviewed literature is a long-established and good (but imperfect) system for transmitting and saving scientific knowledge. A grand challenge for our generation is to build equally strong open data systems for transmitting and saving the scientific data that support our knowledge.

4) Good data management begins at point of capture. I fully support Simpson's point that good data management begins at the point of data collection, not at the point of publication. Most of our data losses occur because data management effort is mostly invested at the end of a project cycle, when it is particularly laborious and when scientists are ready to move on to their next paper, grant, or project. Our field needs sustained research and investment in data systems that support and facilitate data management at all stages of the process, with as little burden as possible placed on individual scientists. For example, when coring and drilling lakes, we need integrated data management systems for easily capturing coring metadata, the data and metadata when splitting and imaging cores, the depth models and age-depth models generated from these cores, the proxy measurements made by multiple research groups on these cores, and the eventual analyses and papers that result from these cores.

At Neotoma, our mission is to support global-change research by providing a high-quality community-curated data resource for paleoecological and paleoenvironmental data (www.neotomadb.org) (Williams et al., 2018). We traditionally specialize in paleoecological proxy data from a variety of terrestrial archives (e.g. diatoms, ostracodes, pollen, testate amoebae, vertebrates) and are expanding our data models to store geochemical data such as stable isotopes and organic biomarkers. We view ourselves as one node in a larger open-data ecosystem, complementary to other primary data archives (e.g. NOAA/NCEI Paleoclimatology, Pangaea) and supportive of high-level synthesis efforts such as PAGES2K, PalEON, or SKOPE. A key element of Neotoma's approach is to stay closely engaged with data generators and data users and to build a network of expert Data Stewards, serving a role akin to editors in peer-reviewed journals. Our governance model is open and built around the concept of Constituent Databases, each representing a particular proxy type or region with associated communities of Data Stewards. Neotoma seeks to enable living data systems, with tools for data updates and amendments by Data Stewards. The most common amendment is addition of new age-depth models, as age-depth modeling approaches improve and data synthesis efforts rebuild age models. But, more generally, we seek to support ongoing improvements to Neotoma's data, mediated by trained Data Stewards, because many forms of data error and corruption are only uncovered by data use.

Neotoma's data use policy includes an embargo policy (https://www.neotomadb.org/data/category/use). For now, policy is ahead of technical implementation: data embargoes are currently implemented by receiving data submissions and preparing them for upload (using the Tilia software system for data cleaning and validation), but avoiding actual upload to the online Neotoma database until embargo is released. We are working on technical implementation of an embargo system for the main database so that data can be submitted to the database and DOIs assigned, but no actual data are exposed or released until the embargo is lifted. The larger goal here is to create systems that encourage good data management practices by data generators (encouraging early data submissions and incorporation of Neotoma into lab-scale workflows) while also protecting the first-use rights of data generators.

A larger and final point is that paleoclimatology and paleoecology have an excellent tradition of data synthesis and data sharing, thanks to our recognition that a global-

scale understanding of the climate system demands a pooling of our many site-level records and thanks to pioneering efforts such as CLIMAP and COHMAP. We have a good culture of data sharing, and an awareness of its complexities and tradeoffs, as this discussion by Kaufman and others shows well. Our field is well positioned to create new institutional and social solutions to these new opportunities and challenges of open data sharing, and to be an example to other disciplines wrestling with similar challenges.

References Blaauw, M., Christen, J.A., 2011. Flexible paleoclimate age-depth models using an autoregressive gamma process. Bayesian Analysis 6, 1-18.

Dawson, A., Paciorek, C.J., McLachlan, J.S., Goring, S., Williams, J.W., Jackson, S.T., 2016. Quantifying pollen-vegetation relationships to reconstruct forests using 19th-century forest composition and pollen data. Quaternary Science Reviews 137, 156-175.

Jackson, S.T., 2012. Representation of flora and vegetation in Quaternary fossil assemblages: known and unknown knowns and unknowns. Quaternary Science Reviews 49, 1-15.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G., 1997. Nongeospatial metadata for the ecological sciences. Ecol. Appl. 7, 330-342.

Parnell, A.C., Haslett, J., Sweeney, J., Doan, T.K., Allen, J.R.M., Huntley, B., 2016. Joint palaeoclimate reconstruction from pollen data via forward models and climate histories. Quaternary Science Reviews 151, 111-126.

Williams, J.W., Grimm, E.G., Blois, J., Charles, D.F., Davis, E., Goring, S.J., Graham, R., Smith, A.J., Anderson, M., Arroyo-Cabrales, J., Ashworth, A.C., Betancourt, J.L., Bills, B.W., Booth, R.K., Buckland, P., Curry, B., Giesecke, T., Hausmann, S., Jackson, S.T., Latorre, C., Nichols, J., Purdum, T., Roth, R.E., Stryker, M., Takahara, H., 2018. The Neotoma Paleoecology Database: A multi-proxy, international community-curated data resource. Quaternary Research 89, 156-177.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2017-157, 2017.