doi.org/10.22498/pages.29.1.59

Towards increased interoperability of paleoenvironmental observation data

Oliver Bothe[1], K. Rehfeld[2], B. Konecky[3] and L. Jonkers[4]

Data is an important foundation of scientific progress. It allows us to contrast hypotheses with observational evidence. Sharing and providing data openly have a long tradition in paleoenvironmental research, supported by repositories such as WDS-Paleo[1], PANGAEA[2],and Neotoma.[3]

The 2018 *Past Global Changes Magazine* issue (Williams et al. 2018) "Building and Harnessing Open Paleodata" touches on all the questions from the production of individual records to the reuse of compilations. Common themes were conventions for reporting, for metadata, and for data structures; crediting mechanisms, community as well as external support in data curation and infrastructure; automating processes; and making data more widely usable.

Today, with many new published data compilations (e.g. Iso2k[4], Konecky et al. 2020; SISAL[5], Comas-Bru et al. 2020; PalMod, Jonkers et al. 2020[6], Cao et al. 2020[7]; PlioVAR[8], McClymont et al. 2020), the need for improving reusability and interoperability of data is becoming more pressing. Each of those compilations adheres, to some extent, to the principles of Findability, Accessibility, Interoperability, and Reusability (FAIR; Wilkinson et al. 2016). The creation of such compilations, which includes quality controlling large numbers of original data records, improves the interoperability of available data records and increases the amount of usable data for understanding past environments and assessing uncertainty. But are the syntheses themselves interoperable enough?

Interoperability benefits from common standards about what is reported, using which vocabularies, and in which storage structures (see e.g. Khider et al. 2019). The highlighted compilations still use a variety of vocabularies and metadata elements. They are provided in a number of different formats including LiPD files, a SQL database, and tab-limited text files. Working with multiple compilations requires becoming fluent enough in each of them to write code to harmonize data formats, interact with files, or produce new files.

A harmonized workflow would allow data from different compilations to be used together more efficiently. This in turn would mean that findings could rely on a larger amount of data and better account for uncertainties. In short, we could more reliably establish agreement and disagreement between data sources (including simulation output), and we could obtain more complete pictures of past environments. Standardization of data synthesis products would therefore be a valuable step towards standardization of all paleoenvironmental observation data and towards using all paleoenvironmental data to their fullest, which certainly motivates many PAGES working group activities.

A number of recent initiatives provide key elements of such a toolchain. Curated repositories assist in harmonizing reporting standards, vocabularies, as well as data formats. These repositories cater to a number of research fields with different conventions. Requirements may also differ between the data producers and the data users. Of particular interest for interoperability are the storage conventions and the vocabularies.

In contrast to paleo-observational data, established sharing and access channels as well as utilities provide standardized workflows and a high degree of FAIRness for simulation output. Paleo-observational data standardization efforts can benefit from the experiences of the wider Earth system modeling community. However, harmonizing climate simulation output with tools like the Climate Model Output Rewriter (CMOR[9]) may be more straightforward than harmonizing paleoenvironmental observations. For the latter, we have yet to finish coordinating vocabularies among research communities and may still have to optimize multiple ways of organizing and storing research data before a standard emerges. Finally, we might find that we cannot use one common format but rather that we need a well-designed, automatable, and well-documented set of tools for interacting with multiple community specific standards to create, modify, and update (parts of) files, as well as read files from different formats.

Community engagement is necessary for tools to be adopted for community specific-use cases. Development and maintenance of tools must not depend on individuals and short funding cycles. Community governance as well as technical solutions can ensure sustainable long-term support for standards for reusable and interoperable paleoenvironmental data that maximally serve our understanding of past and future environmental changes. The paleoenvironmental community, as a community of many research communities, has to provide guidance. For this to be established and adhered to, communities as represented, for example, by the PAGES working groups, have to talk to each other, the repositories for paleoenvironmental data, and providers of technological infrastructure. Then, we can tailor standards, formats, and tools to community needs.

PAGES has taken up data stewardship as an integrative activity with relevant structures and cooperations. Thus, PAGES and comparable efforts are in an ideal position to assist sustainable solutions with a long-term commitment, for which the new Data Stewardship Scholarship[10] offered to PAGES working groups may be a valuable stepping stone. Another step can be for PAGES working groups and PAGES governance to instigate and moderate the necessary conversations, e.g. in the form of a virtual data roundtable bringing together all interested parties.

Figure 1: The diverse formats of paleoenvironmental datasets resemble an assortment of gear wheels that do not necessarily work together (Image credit: Laura Ockel, Unsplash[11]).

AFFILIATIONS

[1]Helmholtz-Zentrum Hereon, Geesthacht, Germany
[2]Institute of Environmental Physics, Ruprecht-Karls-Universität Heidelberg, Germany
[3]Department of Earth and Planetary Sciences, Washington University, St. Louis, MO, USA
[4]MARUM – Center for Marine Environmental Sciences, University of Bremen, Germany

CONTACT

Oliver Bothe: oliver.bothe@hereon.de

REFERENCES

Cao X et al. (2020) Earth Syst Sci Data 12: 119-135

Comas-Bru L et al. (2020) Earth Syst Sci Data 12: 2579-2606

Jonkers L et al. (2020) Earth Syst Sci Data 12: 1053-1081

Khider D et al. (2019) Paleoceanogr Paleoclimatol 34: 1570-1596

Konecky BL et al. (2020) Earth Syst Sci Data 12: 2261-2288

McClymont EL et al. (2020) Clim Past 16: 1599-1615

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

Williams JW (Eds) (2018) PAGES Mag 2(26), 52 pp

LINKS

[1]https://www.ncdc.noaa.gov/data-access/paleoclimatology-data
[2]https://www.pangaea.de/
[3]https://www.neotomadb.org/
[4]https://doi.org/10.25921/57j8-vs18
[5]https://doi.org/10.17864/1947.256
[6]https://doi.org/10.1594/PANGAEA.908831
[7]https://doi.org/10.1594/PANGAEA.898616
[8]https://doi.org/10.1594/PANGAEA.911847
[9]https://cmor.llnl.gov/
[10]http://pastglobalchanges.org/science/wg/data-stewardship-scholarship
[11]https://unsplash.com/photos/e_hQZ2EM-Qg

CC-BY