Earth System Science Data Discussions

Earth Syst. Sci. Data Discuss., community comment CC1
https://doi.org/10.5194/essd-2021-272-CC1, 2021

Reply on RC1

John Williams

Community comment on "The Reading Palaeofire database: an expanded global resource to document changes in fire regimes from sedimentary charcoal records" by Sandy P. Harrison et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-272-CC1, 2021

I'm adding a comment to Dan's review, in my role as one of the leaders of the Neotoma Paleoecology Database and its Leadership Council, to add follow-on information about Neotoma and how we might be able to help:

*Neotoma is a community-curated data resource, in which coalitions of expert Data Stewards upload data to Neotoma and curate data in Neotoma. These coalitions are organized as Constituent Databases, which usually represent a particular type of data, perhaps bounded by region, e.g. the International Ostracode Database, the European Pollen Database, the North American Pollen Database. In this framework, it would be quite possible for the Reading Paleofire Database to join Neotoma as a Constituent Database and become the body charged with uploading and curating (and, effectively, governing) charcoal data uploaded to Neotoma.

*As Dan noted, Neotoma's primary mission is to store original measured variables, such as charcoal counts. Neotoma rarely stores derived variables, such as e.g. influx. This focus on primary data allows Stewards to focus on catching and removing the kind of errors that Dan flags. In this approach, derived variables such as influx are calculated outside of Neotoma, by using Neotoma software services to extract data from Neotoma and then calculate the derived variables in, e.g. R. This is the same approach followed by paleoclimatological users of Neotoma data, who pull the raw data from Neotoma and then apply paleoclimatic transfer functions externally. This approach supports reproducible workflows in which everyone starts with the same raw data and then can build their own analytical pipelines to conduct their particular analyses of interest.

*All contributions of data to Neotoma are voluntary. We try to encourage data contributions by providing a series of services to contributors and users. These include: 1) All datasets in Neotoma are assigned persistent unique digital identifiers (DOIs) with associated landing pages (e.g. https://data.neotomadb.org/14194). 2) Neotoma data can be searched and retrieved through third-party portals such as NOAA's National Center for Environmental Informatics – Paleoclimatology (https://www.ncdc.noaa.gov/paleo-search/) and the Earth Life Consortium (https://earthlifeconsortium.org/). 3) The Neotoma APIs (https://api.neotomadb.org/api-docs/) and R package (https://cran.r-project.org/web/packages/neotoma/index.html) provide several options for large-scale search and retrieval of Neotoma data. The map-based graphical user interface Neotoma

Explorer (https://apps.neotomadb.org/explorer/) supports quick browsing and viewing of data. 4) Snapshots of the full Neotoma database are periodically posted to Neotoma's home page in PostgreSQL (https://www.neotomadb.org/snapshots) and to FigShare (https://figshare.com/search?q=neotoma+database). 5) Support of multi-proxy paleoecological datasets, so that e.g. charcoal datasets can be stored and analyzed in conjunction with other proxies such as fossil pollen.

*Neotoma is endorsed as a repository by the US National Science Foundation, the US Geological Survey, Past Global Changes (PAGES), and the American Quaternary Association. Neotoma is registered as an American Geophysical Union COPDESS data resource and is accredited by the ICSU World Data System, with a CoreTrustSeal accreditation pending. Data uploaded to the Neotoma relational database are housed at the Center for Environmental Informatics at Penn State and are protected by data backup and archiving policy that facilitates swift backup and long-term archiving. Data are protected by multiple measures, including redundant disk storage, off-site mirroring, file-system snapshotting, regular tape backup, and duplication of the backup set.

*There are three potential barriers to having the Reading Paleofire Database join Neotoma as a Constituent Database, all of which are real but have potential solutions:

1) The time involved in preparing data for upload via Tilia and doing the associated quality checks. Some of this time needs to be spent no matter what, given the issues flagged in Dan's review. Neotoma has some resources to help: A) We often run Tilia training workshops for interested Data Stewards; B) we have a network of Stewards that can offer advice and help troubleshoot, and probably could prepare a few demo charcoal datasets as examples; and C) we have an active Slack channel (https://join.slack.com/t/neotomadb/shared_invite/zt-cvsv53ep-wjGeCTkq7IhP6eUNA9NxYQ) where people can post questions and solutions.

2) Data governance – who makes decisions about the data? This issue is often a concern to first-time data contributors. Neotoma addresses this through its Constituent Databases, in which sets of Data Stewards have authority to upload and modify data of a particular type, but are not allowed to modify other data types. For example, a vertebrate paleontologist Data Steward cannot modify a fossil pollen record, nor can a palynologist Steward modify a vertebrate record. Neotoma also keeps track of which Stewards have worked on which records, so that curatorial decisions (e.g. adding a new age-depth model) are linked to the people making the decisions.

3) Data completeness in the Paleofire Database. Neotoma does require, for example, that original age controls such as radiocarbon dates be uploaded with the dataset. Counts are strongly preferred over derived variables. This may mean that not all records in the Paleofire database can be uploaded to Neotoma.

All of the above is informational. I personally see a strong logic for adding the charcoal database to Neotoma, for the reasons that Dan outlines, and particularly to better support multi-proxy paleoecological research. I'm happy to help if there's interest in connecting the resources. But, it's up to the authors, of course, to decide how best to proceed.