Building a controlled vocabulary for the international lipid biomarker database

Harleena Franklin[1], E.K. Thomas[1], J.W. Williams[2], J.M. Aguilar[3], I.S. Castañeda[4], K.H. Freeman[5], N. McKay[6] and C. Morrill[7]

AFFILIATIONS
[1]Department of Geology, University at Buffalo, USA
[2]Department of Geography, University of Wisconsin Madison, USA
[3]Department of Chemistry, University at Buffalo, USA
[4]Department of Earth, Geographic, and Climate Sciences, University of Massachusetts Amherst, USA
[5]Department of Geosciences, Penn State University, University Park, USA
[6]School of Earth and Sustainability, Northern Arizona University, Flagstaff, USA
[7]Climatic Science and Services Division, NOAA's National Centers for Environmental Information, Washington DC, USA

ORCID
H Franklin - 0000-0002-9960-2727
EK Thomas - 0000-0002-6489-7123
JW Williams - 0000-0001-6046-9634
IS Castañeda - 0000-0002-2524-9326
N McKay - 0000-0003-3598-5113

CONTACT
Harleena Franklin: harleena@buffalo.edu

Achieving global-scale insights into past climate variations requires the careful assembly and standardization of networks of proxies databases (Kaufman et al. 2020; Konecky et al. 2020; Walter et al. 2023). Moreover, it is expected that scientific data are openly and readily shared online. These expectations were formalized through the FAIR Guiding Principles (Wilkinson et al. 2015), which created a standard framework that open scientific data should be findable, accessible, interoperable, and reusable.

Controlled vocabularies are essential infrastructure to meet the FAIR principles, thereby enabling global-scale data syntheses and subsequent scientific research. Controlled vocabularies are sets of terms constrained by specific rules that allow for concise and unambiguous usage (Wojcik 2006). Several community-led controlled vocabularies are emerging in paleoclimatology and paleoecology, including the PaST Thesaurus employed by the NOAA World Data Service for Paleoclimatology (Morrill et al. 2021) and the steward-curated taxonomies used by the Neotoma Paleoecology Database (Williams et al. 2018). As the volume and variety of empirical data in paleoclimate research expands, controlled vocabularies developed by experts and consistently shared paleodatabases become ever more essential.

Lipid biomarkers are common in climate and environmental studies, especially in the near-recent times, and represent readily analyzed lipids that have homologous series distributions, ratios, and isotope abundances with high utility for the paleoclimate community. Despite this, these have no comprehensive controlled vocabulary for paleoclimate and environmental use, although the International Union of Pure and Applied Chemistry (IUPAC) dictionary exists for many compounds. Here, we present a draft controlled vocabulary that encompasses several major classes of lipid biomarkers commonly applied for paleoclimate research. To facilitate interoperability among data resources, the NOAA World Data Service, LiPDverse, and Neotoma have all agreed to adopt this

vocabulary. This vocabulary is being developed as an open process, and we welcome community input.

Because the task of cataloguing and establishing vocabulary rules for thousands of lipid biomarkers is non-trivial, we have begun with some of the most commonly used lipid biomarkers in paleoclimate research: branched and isoprenoidal glycerol dialkyl glycerol tetraethers, n-alkanoic acids, n-alkanes, alkenones, and long-chain diols. We have developed a list of lipid biomarker names as they are commonly used in the paleoclimate literature, and include the IUPAC term for each compound, to avoid ambiguity. This list is published as v 0.1.0 on Google Sheets (tinyurl.com/ILBD010)1 and is available for comment. We are seeking community input to check for completeness and accuracy within these classes until 31 January 2024. We would also welcome participation by individuals or teams interested in leading development of a list for other classes of lipid biomarkers.

When the community input period is complete, we will update and publish v 1.0.0 of the International Lipid Biomarker Controlled Vocabulary on Zenodo (https://doi.org/10.5281/zenodo.8284383), with updates and future versions possible afterwards. We will also incorporate v 1.0.0 and subsequent versions into the controlled vocabularies maintained by NOAA, LiPDverse, and Neotoma. If there are other databases interested in using this controlled vocabulary, please contact us. With a controlled vocabulary in place, the next steps will be to harmonize the vocabulary in lipid biomarker datasets currently on public paleoclimate databases, and gather and add datasets not yet on these public databases. Anyone interested in contributing vocabulary or datasets can contact Harleena Franklin at harleena@buffalo.edu. Documentation of the process being developed here may be useful to experts seeking to develop controlled vocabularies for other proxies.

Caption: Figure 1: (A) schematic representation of the process used to develop a controlled vocabulary for use in the International Lipid Biomarker Database. This vocabulary is based on previously established language frameworks (semantics) developed by global paleodata resources such as Neotoma, NOAA, and LiPD. Each new biomarker name is assigned a long name (e.g. branched diakyl glycerol tetraether Ia), short name (e.g. brGDGTIa), a higher-order name (brGDGT), and equivalent names from the PaCTs, PaST, and IUPAC controlled vocabularies. Usually, the PaST and PaCTs names are more generic than those proposed here. (B) Once the controlled vocabulary is established, we will add those new terms back into other databases for increased interoperability across paleodata resources. We view this as an iterative and on-going process, with new additions to the controlled vocabulary list to be provided by participating experts and new publications describing lipid biomarker compounds. We invite community input to these names and processes.

References

Kaufman D et al. (2020) Sci Data 7: 115 https://doi.org/10.1038/s41597-020-0445-3
Konecky BL et al. (2020) Earth System Science Data 12: 2261-2288 https://doi.org/10.5194/essd-12-2261-2020
Morrill C et al. (2021) Paleoceanogr Paleoclimatol 36https://doi.org/10.1029/2020PA004193
Walter RM et al. (2023) Earth Syst Sci Data 15: 2081-2116 https://doi.org/10.5194/essd-15-2081-2023
Wilkinson MD et al. (2016) Sci Data 3: 160018 https://doi.org/10.1038/sdata.2016.18
Williams JW et al. (2018) Quaternary Research 89: 156-177 https://doi.org/10.1017/qua.2017.105
Wojcik R (2006) Controlled Languages, 2nd ed. Elsevier Science & Technology, 3 pp https://www.sciencedirect.com/science/article/pii/B0080448542050811