Paleoclimate data standards

Julien Emile-Geay[1] and Nicholas P. McKay[2]

Boulder, USA, 22-23 June 2016

The progress of science can be directly tied to increased standardization of measurements, instruments, and information. Hosted by the US National Centers for Environmental Information (NOAA-NCEI) World Data Service for Paleoclimatology (WDS-Paleo)[1] in Boulder, USA, 40 participants gathered to establish preliminary standards for paleoclimate data. The workshop was primarily supported by the EarthCube-funded LinkedEarth project[2], with auxiliary support from PAGES to ensure international participation.

A common tongue

Building upon the machine-readable Linked Paleo Data framework (LiPD; McKay and Emile-Geay 2016; Fig. 1), the group recognized the need for (i) a structure to organize data and metadata and (ii) a consensual terminology. Group activities highlighted the difficulty to agree upon such terminology, though the conceptualization of Evans et al (2013) emerged as a sound foundation for the structure. WDS-Paleo presented efforts to normalize the terminology used to report observations and their units, for each archive type[3].

Shades of metadata

While more metadata are always desirable, a key discussion focused on distinguishing a set of essential, recommended and optional properties for each dataset. A consensus emerged that these levels are archive-specific (e.g. what is essential to intelligently re-use a dataset is different for a speleothem and a marine sedimentary record). Therefore, the community should coalesce around archive-specific working groups to propose such recommendations. The group also recognized the asymmetry between modern and legacy records (metadata for the latter being inherently more difficult to gather), so what counts as "essential" will be different depending on the age of the study.

A cornerstone of group discussions was to identify what metadata allow to reproduce, cite, and reuse, paleoclimate data. Rewarding data producers emerged as a central theme, thanks in part by presentations on connected efforts on data citations[4], links to scientific expeditions[5], and physical samples[6]. The difficulty to mint data digital object identifiers with the appropriate level of granularity, as well as the current cap on the number of citations in mainstream science journals, were recognized as a major hurdle to crediting data producers; a dialogue with the publishing community should be initiated to remove these hurdles.

Characterizing data records

Part of the drive to enhance the description of paleoclimate data is the need to better communicate their inherent uncertainties. The indirect nature of paleoclimate observations also requires a precise language for their interpretation. This motivated a discussion of a formal paleoclimate ontology, a preliminary version of which was unveiled[7]. Ontologies are formal representations of concepts and their logical connections, and their dividends in other fields were presented. Despite their scientific importance, the concepts of uncertainty and interpretation were found difficult to define unambiguously. However, the group underscored the importance of quantifying how completely these notions are described in the metadata. A multivariate completeness score was proposed[8], allowing users to rapidly evaluate metadata completeness along several dimensions, thus incentivizing best curation practices.

Building consensus

Standards arise by consensus. The group acknowledged the necessity to engage a broad community of paleoscientists, and establish a mechanism to gather their input. The LinkedEarth wiki[9] was designed to facilitate initial discussions by communities of interest, and elaborate recommendations addressing the various issues raised at the workshop. The community engagement process is described at http://linked.earth/community-standards-development/. PAGES will play a key role in this feedback elicitation, and anyone interested in the process should reach out to linkedearth@gmail.com. An official communication from the PAGES IPO about the role of this process in the PAGES Data Stewardship[10] Initiative is forthcoming.

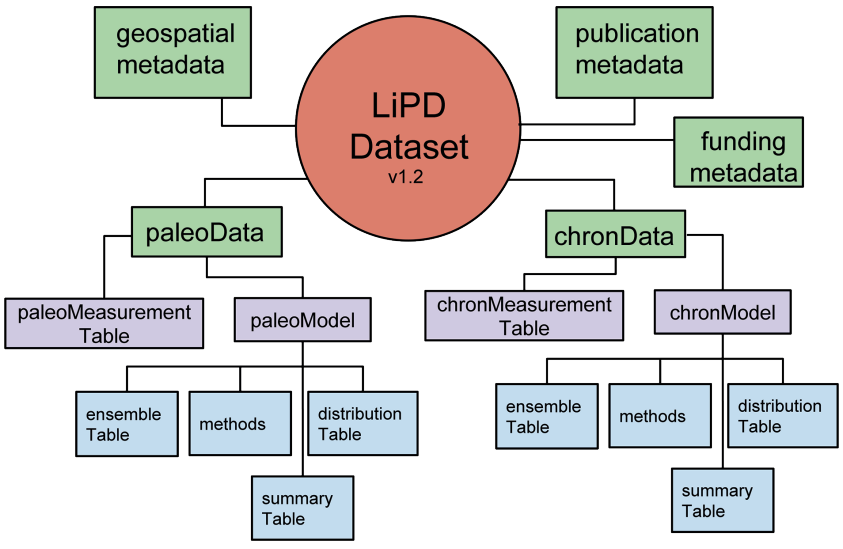

Figure 1: Data and metadata compartments in the current version of LiPD. The format is designed to store, along with several kinds of metadata (location, publication, funding), data tables pertaining to the paleosystem of interest (most often, climate), as well as chronology data, including raw chronological tie points and models based on them.

AFFILIATIONS

[1]Climate Dynamics Lab, University of Southern California, Los Angeles, USA

[2]School of Earth Sciences and Environmental Sustainability, Northern Arizona University, Flagstaff, USA

CONTACT

Julien Emile-Geay: linkedearth@gmail.com

LINKS

[1]www.ncdc.noaa.gov/data-access/paleoclimatology-data; [2]http://linked.earth; [3]https://agu.confex.com/agu/fm15/webprogram/Paper65476.html; [4]www.datacite.org; [5]www.geolink.org; [6]www.geosamples.org; [7]http://vowl.visualdataweb.org/webvowl/#iri=http://linked.earth/ontology; [8]http://wiki.linked.earth/WG_completeness; [9]http://wiki.linked.earth; [10]www.pastglobalchanges.org/ini/int-act/data-stewardship

REFERENCES

Evans MN et al. (2013) Quat Sci Rev 76: 16-28

McKay NP, Emile-Geay J (2016) Clim Past 12: 1093–1100