doi.org/10.22498/pages.26.2.71

Linked Paleo Data: A resource for open, reproducible, and efficient paleoclimatology

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

Paleoclimatology is a remarkably diverse field of research, revolving around hard-won and complex datasets that typically represent hundreds of hours of field work, laboratory analyses and nuanced interpretation. Integrating those diverse datasets to piece together a spatio-temporal understanding of how, when and why climate has changed in the past is a grand challenge of paleoclimatology; one that requires careful handling of these data and their interpretations. Researchers often spend up to 80% of their time collecting, organizing and formatting data, before they can even begin addressing the questions they set out to tackle (Dasu and Johnson 2003). This was certainly our experience, and is why we developed the Linked Paleo Data (LiPD) framework. As the number of records relevant to paleoclimate research continues to grow, and the methodologies for investigating datasets and data networks become more complex, our community cannot afford to continue wasting time on data wrangling when there's so much science to be done!

The linked paleodata solution

The technical details of LiPD are presented in McKay and Emile Geay (2016), but the concept is simple: LiPD provides a flexible structure that contains and describes any paleoclimatic or paleoenvironmental dataset, the metadata that describe the details and complexity of the data (at any level from observations to collections), as well as models that accompany the data and their output, such as age models and their ensemble output. This powers efficient, 21st century scientific workflows, and enables open science and reproducible research.

This is why LiPD has been used by multiple data-intensive PAGES working groups, including the 2k Network Temperature Database (PAGES 2k Consortium 2017), and Iso2k[1]. Being able to rely on consistently structured data with rich metadata has greatly reduced the "time to science" for projects relying on the PAGES 2k database, such as the forthcoming global temperature reconstruction intercomparison[2], and the Last Millennium Reanalysis project.

Having structured and standardized data also enables efficient access to state-of-the-art analysis tools. One example is age-uncertain data analysis using the GeoChronR package[3]. GeoChronR relies on LiPD's capacity to contain and describe age-model ensembles to simplify quantifying the effects of age uncertainty on paleoclimate analysis. For example, quantifying and visualizing the impact of age uncertainty on a calibration-in-time with temperature, both on the regression model and the reconstruction back in time, is greatly simplified with LiPD and GeoChronR[4].

A growing LiPD "ecosystem"

Data standards and formats are only as useful as the breadth of their adoption. Thankfully, a LiPD "ecosystem" of datasets, standards, and tools is emerging (Fig. 1).

Datasets: More than 3000 datasets have now been formatted as LiPD files, largely as part of PAGES working group efforts. These data are archived at WDS-Paleo and LinkedEarth (Gil et al. 2017). LiPD is also well suited to serve as an "interchange format", facilitating the transfer of datasets from researchers to repositories and tools. As LiPD is not tied to any particular repository, initial connectivity with WDS-Paleo and Neotoma has been developed, and two-way interoperability with other repositories, including LacCore, and Open Core Data is forthcoming as part of the Throughput project[5].

Standards: From the outset, LiPD was designed to support "Linked Open Data", an international effort to connect data and concepts and make them broadly accessible through the semantic web[6]. As part of the LinkedEarth project, we created the "LiPD Ontology", the first ontology for paleoclimatology[7]. LiPD also enables community-developed data standards (Emile-Geay and McKay 2016; Emile-Geay et al., this issue), including WDS-Paleo's controlled vocabulary[8].

Tools: A wide range of tools that "speak" LiPD have been developed. This includes the LiPD Utilities, which provide basic functionality for reading, writing and querying LiPD data in R, Matlab and Python, and provides the base-level functionality for more sophisticated packages, including GeoChronR[9] and Pyleoclim[10]. A rich set of interactive, graphical, web-based tools for creating and modifying LiPD files has been created at lipd.net.

CScience, an AI-powered tool for age modeling uses LiPD as an input and output format (Bradley et al., this issue).

LiPD has always been collaborative and open-source, and we look forward to the continued expansion and evolution of these data, standards and tools by the community. To learn more about LiPD, how to use if for your research, and upcoming training opportunities, please visit lipd.net

AFFILIATIONS

[1]School of Earth & Sustainability, Northern Arizona University, Flagstaff, USA

[2]Department of Earth Sciences, University of Southern California, Los Angeles, USA

CONTACT

Nicholas P. McKay: Nicholas.McKay@nau.edu

REFERENCES

Dasu T, Johnson T (2003) Exploratory Data Mining and Data Cleaning. Wiley, 212 pp

Emile-Geay J, McKay NP (2016) PAGES Mag 24: 47

Gil Y et al. (2017) The Semantic Web – ISWC 2017: 231-246

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

PAGES 2k Consortium (2017) Sci Data 4: 170088

LINKS

[1] pastglobalchanges.org/ini/wg/2k-network/projects/iso2k

[2] pastglobalchanges.org/ini/wg/2k-network/projects/gmst-recon-2k

[3] nickmckay.github.io/LinkedEarth-Neotoma-P418/LE-Neo_UseCase.html

[4] lipdverse.org/regression.htm

[5] throughput-ec.github.io

[6] w3.org/DesignIssues/LinkedData.html

[7] linked.earth/ontology

[8] ncdc.noaa.gov/data-access/paleoclimatology-data/contributing

[9] nickmckay.github.io/GeoChronR

[10] linkedearth.github.io/Pyleoclim_util

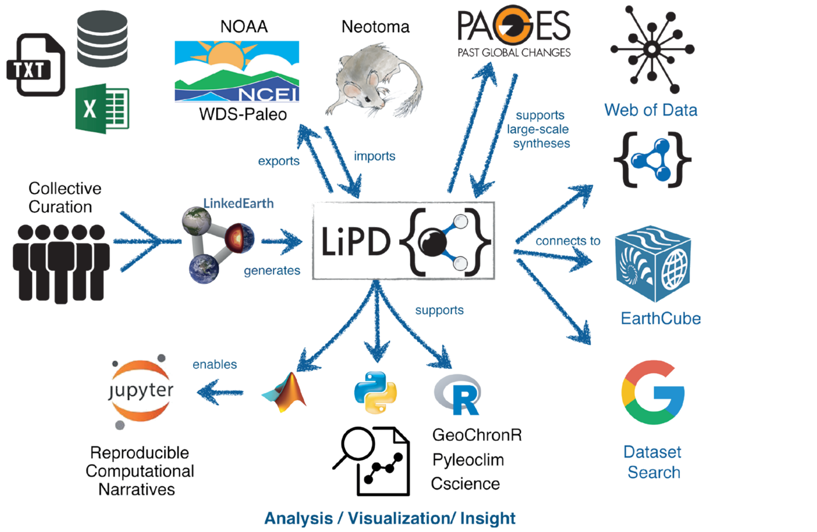

Figure 1: The LiPD ecosystem: a growing network of scientific communities, data repositories, and analysis tools connected and enabled by LiPD.