The Journal of Open Source Software

PyLiPD: A python package for the manipulation of paleoclimate datasets

Varun Ratnakar [1] **and Deborah Khider** [1]

1 Information Sciences Institute, University of Southern California

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: William Gearty

Reviewers:

- @spencerkclark
- @PennyHow

Submitted: 25 April 2025
Published: unpublished

Summary

PyLiPD is a Python package designed to support the reading, querying, editing, and writing of paleoclimate datasets formatted in the Linked Paleo Data Format (LiPD) (McKay & Emile-Geay, 2016). Built on the rdflib library, it transforms LiPD data into Resource Description Framework (RDF) graphs aligned with the LinkedEarth ontology, enabling semantic querying with the SPARQL Protocol and RDF Query (SPARQL). The API is structured into four modules that handle data access, variable-level filtering, and dataset creation and editing, with support for both graph-based and tabular workflows. Comprehensive documentation and tutorials—including ontology concepts and scientific use cases—are provided to guide users.

Statement of Need

Paleoclimate data (obtained from biogeophysical measurements made on ice cores, sediments, corals, trees, etc.) often come with diverse structures and metadata conventions, making synthesis efforts challenging and highlighting the necessity of standardized data formats. Achieving standardization involves three key components: (1) a uniform data format, (2) a consistent terminology for describing metadata, and (3) clear guidelines for how data should be reported. The LiPD framework emerged to address the first need, providing a universally readable data container for paleoclimate data and metadata (McKay & Emile-Geay, 2016). The metadata is stored in a JSON-LD file while the data is organized in multiple tables saved as csv files. LiPD has six distinct components: root metadata (e.g., dataset name, and version); geographic metadata (e.g., coordinates); publication metadata (including unique identifiers); funding metadata (e.g., grant number); PaleoData, which includes all the measured (e.g., the width of tree rings) and inferred (e.g., temperature) paleoenvironmental data; and ChronData, which mirrors PaleoData for information pertaining to time. These components provide the rigidity necessary to write robust codes around the format while remaining extensible enough to capture (meta)data as rich as the users want to provide for them. Paired with the Paleoenvironmental Standard Terms (PaST) Thesaurus developed by the National Oceanic and Atmospheric Administration (NOAA) (Morrill et al., 2021), LiPD standardization facilitates efficient querying and analysis of multiple records, enabling large-scale paleoclimate syntheses. Community-driven initiatives to reconstruct temperature and hydroclimate over Earth's history (Emile-Geay et al., 2017; Jonkers et al., 2020; Kaufman et al., 2020; Konecky et al., 2020; Routson et al., 2021) have compiled paleoclimate records into this standardized format and have made them available to the paleoclimate community in an online database. As the database continued to grow, there was an increasing demand for tools capable of accessing, reading, modifying, and writing files in the LiPD format. Although there is a package in R (Heiser et al., 2018), a Python tool built for scientists and interoperable with libraries such as Pyleoclim (Khider et al., 2022) for time series analysis and cfr (Zhu et al., 2024) for climate field reconstruction is lacking. To address this need, we introduce PyLiPD, a Python-based tool designed specifically for these tasks.

The Journal of Open Source Software

Implementation

PyLiPD is built on top of the Python `rdflib` library (Krech et al., 2025), which supports working with RDF data and provides robust capabilities for parsing, manipulating, and querying semantic graphs using SPARQL. To leverage this functionality, LiPD-formatted datasets are converted into RDF graphs upon loading, using the LinkedEarth ontology (Emile-Geay et al., 2019). The LinkedEarth Ontology captures the LiPD components and the relationship among them and integrates the standardized terms from the PaST Thesaurus.

`PyLiPD`'s user-facing APIs are organized around four main modules. The `LiPD` module allows users to load, manipulate, query, and filter LiPD-formatted datasets stored locally, retrieved via URL, or accessed from an online knowledge base. The API supports typical paleoclimate research queries, including filters by geographic extent, archive type, and temporal coverage. For users more comfortable with tabular formats, graph content can be flattened into a `pandas.DataFrame`. This feature facilitates integration with familiar data analysis tools in the Python ecosystem. Because the LiPD object is a subclass of `rdflib.Graph`, it also supports direct SPARQL querying. Given that PyLiPD's querying capabilities are largely built on `rdflib`'s SPARQL integration, the `pylipd.globals.queries` module includes a range of examples to illustrate common use cases. The `LiPDSeries` module provides functionality for querying and filtering data at the variable level.

All other modules – `pylipd.classes` (referred to as LiPD Classes in the documentation) – and their associated submodules are primarily used for editing and creating LiPD-formatted datasets. Some of the modules (e.g., `pylipd.classes.dataset`, `pylipd.classes.variable`) include Python classes auto-generated from ontology definitions, each equipped with methods to `get`, `set`, or `append` (`add`) property values. Others (e.g., `pylipd.classes.archivetype`, `pylipd.classes.paleovariable`) are used to align with the PaST Thesaurus, offering standardized vocabularies for properties such as variable names, archive types, and units.

The APIs are fully documented and include minimal working examples. Documentation is available on Read the Docs, where users can also find installation instructions and guidelines for contributing to the codebase.

Research Applications

In addition to the minimal working examples in the documentation, more comprehensive tutorials (Khider et al., 2025) are available as a Jupyter Book. These tutorials introduce key ontology concepts—particularly the LinkedEarth Ontology—and offer scientific use cases that demonstrate how to apply the PyLiPD software.

Availability

PyLiPD is an open-source software released under the Apache 2.0 license and is actively maintained as part of the LinkedEarth project. It is available through PyPi and GitHub. Documentation is available through readthedocs.

Acknowledgements

This work is supported by the US National Science Foundation grant RISE 2126510 to Khider.

References

Emile-Geay, J., Khider, D., Garijo, D., McKay, N., Gil, Y., Ratnakar, V., & Bradley, E. (2019). *The linked earth ontology: A modular, extensible representation of open paleoclimate data*. Zenodo. https://doi.org/10.5281/ZENODO.2577604

Emile-Geay, J., McKay, N. P., Kaufman, D. S., Gunten, L. von, Wang, J., Anchukaitis, K. J., Abram, N. J., Addison, J. A., Curran, M. A. J., Evans, M. N., Henley, B. J., Hao, Z., Martrat, B., McGregor, H. V., Neukom, R., Pederson, G. T., Stenni, B., Thirumalai, K., Werner, J. P., … Zinke, J. (2017). A global multiproxy database for temperature reconstructions of the common era. *Scientific Data*, *4*(1). https://doi.org/10.1038/sdata.2017.88

Heiser, C., McKay, N., Simpson, G., & Routson. (2018). *Nickmckay/LiPD-utilities: v0.2.5.5*. Zenodo. https://doi.org/10.5281/ZENODO.1256889

Jonkers, L., Cartapanis, O., Langner, M., McKay, N., Mulitza, S., Strack, A., & Kucera, M. (2020). Integrating palaeoclimate time series with rich metadata for uncertainty modelling: Strategy and documentation of the PalMod 130k marine palaeoclimate data synthesis. *Earth System Science Data*, *12*(2), 1053–1081. https://doi.org/10.5194/essd-12-1053-2020

Kaufman, D., McKay, N., Routson, C., Erb, M., Dätwyler, C., Sommer, P. S., Heiri, O., & Davis, B. (2020). Holocene global mean surface temperature, a multi-method reconstruction approach. *Scientific Data*, *7*(1). https://doi.org/10.1038/s41597-020-0530-7

Khider, D., Emile-Geay, J., James, A., Landers, J., Zhu, F., & Lee, P.-T. (2025). *PyleoTutorials: A gentle introduction to the pyleoclim package.* Zenodo. https://doi.org/10.5281/ZENODO.14782883

Khider, D., Emile-Geay, J., Zhu, F., James, A., Landers, J., Ratnakar, V., & Gil, Y. (2022). Pyleoclim: Paleoclimate timeseries analysis and visualization with python. *Paleoceanography and Paleoclimatology*, *37*(10). https://doi.org/10.1029/2022pa004509

Konecky, B. L., McKay, N. P., Churakova (Sidorova), O. V., Comas-Bru, L., Dassié, E. P., DeLong, K. L., Falster, G. M., Fischer, M. J., Jones, M. D., Jonkers, L., Kaufman, D. S., Leduc, G., Managave, S. R., Martrat, B., Opel, T., Orsi, A. J., Partin, J. W., Sayani, H. R., Thomas, E. K., … Gunten, L. von. (2020). The Iso2k database: A global compilation of paleo-δ18O and δ2H records to aid understanding of common era climate. *Earth System Science Data*, *12*(3), 2261–2288. https://doi.org/10.5194/essd-12-2261-2020

Krech, D., Grimnes, G. A., Higgins, G., Car, N., Hees, J., Aucamp, I., Lindström, N., Arndt, N., Sommer, A., Chuc, E., Herman, I., Nelson, A., McCusker, J., Gillespie, T., Kluyver, T., Ludwig, F., Champin, P.-A., Watts, M., Holzer, U., … Stuart, V. (2025). *RDFLib* (Version 7.1.2). https://doi.org/10.5281/zenodo.6845245

McKay, N. P., & Emile-Geay, J. (2016). Technical note: The linked paleo data framework – a common tongue for paleoclimatology. *Climate of the Past*, *12*(4), 1093–1100. https://doi.org/10.5194/cp-12-1093-2016

Morrill, C., Thrasher, B., Lockshin, S. N., Gille, E. P., McNeill, S., Shepherd, E., Gross, W. S., & Bauer, B. A. (2021). The paleoenvironmental standard terms (PaST) thesaurus: Standardizing heterogeneous variables in paleoscience. *Paleoceanography and Paleoclimatology*, *36*(6). https://doi.org/10.1029/2020pa004193

Routson, C. C., Kaufman, D. S., McKay, N. P., Erb, M. P., Arcusa, S. H., Brown, K. J., Kirby, M. E., Marsicek, J. P., Anderson, R. S., Jiménez-Moreno, G., Rodysill, J. R., Lachniet, M. S., Fritz, S. C., Bennett, J. R., Goman, M. F., Metcalfe, S. E., Galloway, J. M., Schoups, G., Wahl, D. B., … Cumming, B. F. (2021). A multiproxy database of western north american holocene paleoclimate records. *Earth System Science Data*, *13*(4), 1613–1632. https://doi.org/10.5194/essd-13-1613-2021

Zhu, F., Emile-Geay, J., Hakim, G. J., Guillot, D., Khider, D., Tardif, R., & Perkins, W. A. (2024). Cfr (v2024.1.26): A python package for climate field reconstruction. *Geoscientific Model Development*, *17*(8), 3409–3431. https://doi.org/10.5194/gmd-17-3409-2024