The Linked Paleo Data framework: a common tongue for paleoclimatology

Nick McKay

Affiliation not available

Julien Emile-Geay

Affiliation not available

May 23, 2016

Abstract

Paleoclimatology is a highly collaborative scientific endeavor, increasingly reliant on online databases for data sharing. Yet, there is currently no universal way to describe, store and share paleoclimate data: in other words, no standard. Data standards are often regarded by scientists as mere technicalities, though they underlie much scientific and technological innovation, as well as facilitating collaborations between research groups. In this article, we propose a preliminary data standard for paleoclimate data, general enough to accommodate all the proxy and measurement types encountered in a large international collaboration (PAGES2K). We also introduce a vehicle for such structured data (Linked Paleo Data, or LiPD), leveraging recent advances in knowledge representations (Linked Open Data).

The LiPD framework enables quick querying and extraction, and we expect that it will facilitate the writing of open-source, community codes to access, analyze, model and visualize paleoclimate observations. We welcome community feedback on this standard, and encourage paleoclimatologists to experiment with the format for their own purposes.

1 Introduction

Science is entering a data-intensive era, where insight is increasingly gained by extracting information from large volumes of data [5]. This is particularly critical in paleoclimatology, as understanding past changes in climate system requires observations across large spatial and temporal scales. Paleoclimatic observations are typically limited to small geographic domains, so investigating large scales requires integrating many disparate studies and datasets. Observational work in paleoclimatology exemplifies the “long-tail” approach to data collection [4]: the majority of observations are gathered by independent scientists with no formal language for describing their data and meta-data to each other – or to machines – in a standardized fashion. This results in a “Digital Tower of Babel”, making the curation, access, re-use and valorization of paleoclimate data far more difficult than it should be, hindering scientific progress.

Recognizing the need for data sharing, paleoclimate investigators have made a major effort over the past decade to make their data available to the broader community, largely through online archiving systems like the World Data Center for Paleoclimatology and Pangaea. Nonetheless, the lack of consistent formatting and metadata standards (i.e. a common tongue) has made the re-use of such data needlessly labor-intensive by preventing computers from participating in the task of making connections across datasets. As the number of records in these archives has grown, making connections manually has become more and more challenging, hampering integrative efforts at the very time they should be flourishing. Paleoclimatologists thus need a common tongue to describe their datasets to each other and to machines. Achieving this goal requires addressing two major hurdles: (1) the lack of an accepted data container for paleoclimate data; (2) the lack of a community standard for such data.

These two issues are clearly related, but somewhat distinct in practice. The data container must be universally readable, a condition satisfied by, for instance, netCDF files, which have been used for paleoclimate syntheses [9]. However, such files only allow for fixed schemas and require identical fields for all proxies. In reality, each proxy dataset may have a unique set of data and metadata properties. For broader applicability, we thus require a more flexible format. Further, to enhance the relevance of paleoclimate data to other fields, one would like this data container to be compatible with the Linked Data paradigm [2], which allows for data-driven discovery between datasets that would otherwise be unlikely or impossible.

In this technical note, we present LiPD (Linked Paleo Data) a new, flexible linked-data container designed for paleoclimate data. Such a data container is a necessary first step towards a “semantic web of paleoclimatology” [3], and provides a straightforward framework in which communities and researchers can explicitly describe their data and metadata in common terms that the community, and computers, can understand. In the process, we introduce a preliminary data standard for paleoclimatology. Indeed, such a standard is essential to structuring the metadata, though the container is flexible enough to accommodate many revisions and updates. Ideally, such a standard would proceed from a community-wide discussion, and the establishment of a consensus, which has yet to take place in our field. One goal of the present work is to spark such a discussion by giving the worldwide paleoclimate community a strawman to improve upon.

This article is structured as follows: In section 2 we describe the new container, LiPD. In section 3 we describe the proposed metadata standard. We close with a discussion section.

2 A flexible container for paleoclimate data

Paleoclimate observations come in many varieties; standardizing the data and providing the framework to encode meaning to the parameters and metadata requires a flexible, and extensible format. The linked data variety of JavaScript Object Notation (JSON-LD), provides a lightweight, and human-readable solution to this problem. JSON-LD is almost infinitely customizable, here we present conventions for Linked Paleo Data (LiPD), which utilizes JSON-LD and provides a structure that is common to the overwhelming majority (if not all) of paleoclimate observational datasets. Despite their variety, all paleoclimate datasets, commonly comprised of both direct observations and inferred variability, share the same major features.

1. Some base metadata about the dataset (e.g.)
   (a) Identifiers (dataset name, version number, dataset DOI, investigators)
   (b) Archive[1] type
2. Geographic metadata (e.g.,)
   (a) latitude, longitude, elevation above or depth below sea level
   (b) site name
3. Publication metadata (e.g.,)
   (a) DOI (which resolves the following information)
   (b) authors, title, journal, publication date
4. Proxy data and metadata, including:
   (a) One or more tables of measurements, and their metadata
   (b) Parameter names, units, standards, and interpretations (including forward models)
5. Geochronological data and metadata, which can include
   (a) Table(s) of radiometric dating measurements and associated metadata
   (b) Age model ensembles
   (c) Author interpretation and methodological choices

LiPD encodes these data and metadata into a structured hierarchy that allows explicit description of any aspect of the dataset at any level of the data (Figure 1). LiPD serializes this hierarchy using JSON-LD, using nests of lists and key-value pairs. LiPD adopts the GeoJSON standard to describe the geospatial metadata of a given site like this:

[1] the archive is the medium in which the paleoclimatic signal is imprinted: e.g. coral aragonite, ice core, foraminiferal calcite in a sediment core, etc...

```
"geo": {
"type": "Feature",
"geometry": {
"type": "Point",
"coordinates": [-17.82, 62.08, -1938]
},
"properties": {
"siteName": "RAPiD-12-1K, South Iceland Rise, northeast North Atlantic"
}
},
```

Note that the GeoJSON standard defaults to the WGS84 ellipsoid, and units of decimal degrees for latitude and longitude and meters above sea level for elevation. This standard readily accommodates polygonal geographic features and additional location metadata.

LiPD adopts the Linked Data extension of BibJSON [6] to describe publication metadata, for example:

```
"pub": {
"author": [
{"name" : "Thornalley, D.J.R"},
{"name" : "Elderfield, H.},
{"name" : "McCave, N"}
]
"type" : "article"
"identifier" : [
{"type": "doi",
"id": "10.1038/nature07717",
"url": "http://dx.doi.org/10.1038/nature07717"}
],
"pubYear": 2009
},
```

For the paleoData and chronData components of LiPD, which include tabular data, LiPD does not store the actual tabular data in the JSON-LD format, as this becomes increasingly verbose and inefficient with large data tables. Rather, the tabular data are stored in headerless comma separated value (CSV) files that are referenced and described by the JSON-LD file, using the W3C's CSV on the Web working groups recommendations, like this:

```
"paleoData": [{
"paleoDataTableName": "data",
"filename": "Atlantic0220Thornalley2009.csv",
"columns": [{
"number": 1,
"parameter": "depth",
"parameterType": "measured",
"description": "depth below ocean floor",
"units": "cm",
"datatype": "csvw:NumericFormat",
"notes": "depth refers to top of sample"
},
{
"number": 2,
"parameter": "year",
"parameterType": "inferred",
"description": "calendar year AD",
"units": " AD",
"datatype": "csvw:NumericFormat",
"method": "linear interpolation"
},
{
"number": 3,
"parameter": "temperature",
"parameterType": "inferred",
"description": "sea-surface temperature inferred from Mg/Ca ratios",
"datatype": "csvw:NumericFormat",
"material": "foramifera carbonate",
"calibration": {
"equation": "BAR2005: Mg/Ca=0.794*exp(0.10*SST)",
"reference": "Barker et al., (2005), Thornalley et al., (2009)",
"uncertainty": 1.3
},
"units": "deg C",
"proxy": "Mg/Ca",
"climateInterpretation": {
"parameter": "T",
"parameterDetail": "seaSurface",
"seasonality": "MJJ",
"interpDirection": "positive",
"basis": "Mg/Ca calibration to SST"
}
}]
}]
```

Describing the columns in the datatable in the LiPD framework allows explicit encoding of key metadata that are commonly lost or misunderstood in current data structures. For example, the “climateInterpretation” section above allows the scientist to explicitly describe the details of how the parameter “senses” climate. When encoded as above and explicitly defined and linked, the knowledge that this record is interpreted to record May through July sea surface temperature, and that those temperature estimates were derived from the Mg/Ca calibration equations of **(author?)** [1] and **(author?)** [8] becomes built into the dataset, and readable to both people and computers. It's queryable, and linked to other datasets, and transparent when datasets are used in ways that are outside the published interpretations.

An advantage of using JSON as the default container for this information is that it is an extremely common vehicle for all manner of data, and can be parsed by nearly all modern programming languages. As each LiPD dataset is comprised of a JSON-LD file and one or more csv files; each dataset is packaged using BagIt [2], which provides a simple format for collecting and validating files for distribution, and that can be readily serialized into a compressed file for exchange between users. To facilitate input and output of LiPD datasets, we are developing code to easily export LiPD datasets into and out of Matlab (as structured arrays), R (as lists) and Python (as dictionaries).

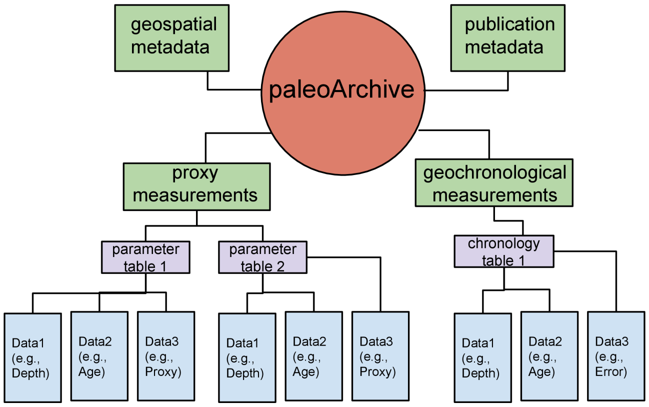

Figure 1: Figure 1. Schematic data model example of Linked PaleoData.

3 A preliminary data standard for paleoclimatology

The flexible container described in section 2 can serialize any set of paleoenvironmental data with rich metadata. However this framework only becomes useful when a common vocabulary with explicit meanings is applied to the data. Developing this vocabulary requires buy-in from experts across the disparate domains of the paleogeosciences, and will be a gradual process of evolving standards. To begin this conversation, here we outline a preliminary metadata standards for required metadata, based on phase 2 of the Past Global Changes (PAGES) past two thousand years (2k) project. The following are the minimal metadata for every paleoArchive in the network. Many records include additional desirable data and metadata; an ongoing extended metadata table is available here. To

[2] https://en.wikipedia.org/wiki/BagIt

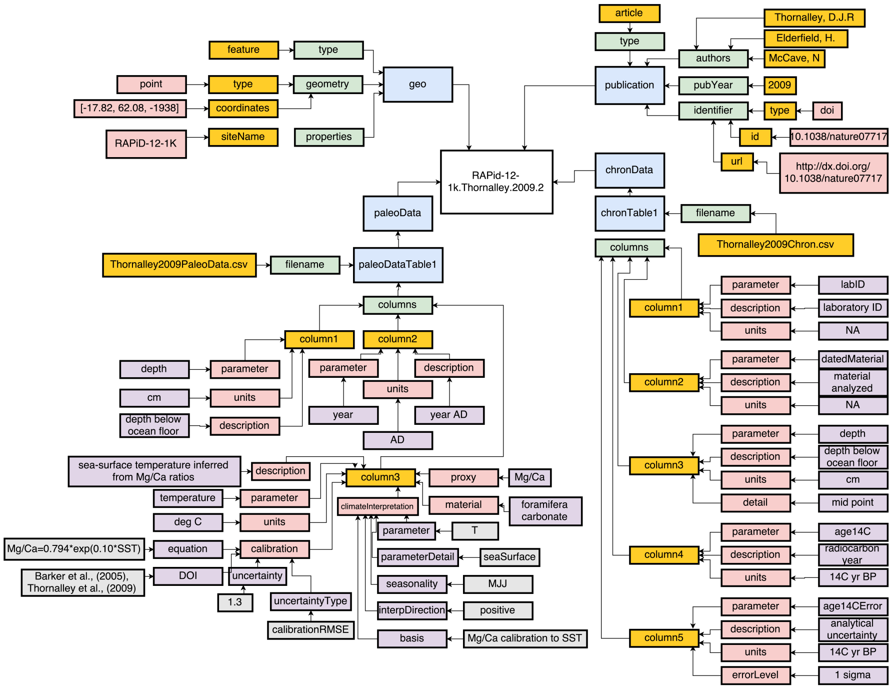

Figure 2: Figure 2. Data model for RAPiD-12-1k example used here.

illustrate this standard more concretely, we use the dataset of **(author?)** [8] as an example (Figure 2 and Table below).

Base metadata metadata that apply to the dataset as a whole:

dataSetName name of the dataset; that is, an alphanumeric string that uniquely characterizes this record in the database, often based on site, authors, year and ancillary information *example: RAPiD-12-1K.Thornalley.2009*

archiveType *example: marine sediments*

investigator *example: David Thornalley*

Geospatial metadata metadata that apply to the location of the study site, following the GeoJSON convention:

coordinates longitude, latitude, and elevation, in units of decimal degrees and meters above sea level; *example: "coordinates": [-17.82, 62.08, -1938]*

type geographic feature type (typically "point" or "polygon" *example: point*

siteName *RAPiD-12-1K*

Publication metadata metadata that apply to the publication(s) associated with the paleoArchive:

DOI publication Digitial Object Identifier; *example: 10.1038/nature07717*

citation long publication string if DOI is unavailable; *example: Thornalley, David JR, Harry Elderfield, and Nick McCave. ”Holocene Oscillations in Temperature and Salinity of the Surface Subpolar North Atlantic.” Nature 457, no. 7230 (2009): 711–14.*

pubString short text citation; *example: Thornalley et al., 2009*

paleoDataTable metadata metadata associated with paleoDataTable:

paleoDataTableName short name of the paleoData table; *example: data*

filename measurement table filename (.csv); *example: Atlantic0220Thornalley2009.csv*

chronology short name of chronology table used in this measurement table; *example: chronology*

paleoData table columns columns required in measurement table, will vary by archive and proxy type

depth depth of sample/measurement

age/year best estimate age/year of sample/measurements

climate-sensitive parameter measurement interpreted in terms of past climate or environmental change *example: Mg/Ca*

paleoData column metadata metadata characterizing each column of the paleoDataTable

column column number; *example: 3*

parameter short parameter name *example: SST*

description longer description of the parameter *example: sea-surface temperature inferred from Mg/Ca ratios*

units *example: deg C*

climateInterpretation five parameters that allow for a concise description of how the climate-sensitive parameter is related to climate. This is required for at least one column in the PAGES 2k database, but may not be appropriate for all paleoArchives.

- **parameter** what aspect(s) of climate are recorded in this archive; *example: temperature*
- **parameterDetail** provides detail on ”climateInterpretationParameter” *example: sea surface*
- **seasonality** *example: May, June , July*
- **interpretationDirection** Do the values have a positive or negative relation to the inferred parameter *example: positive*

basis quote from paper or other argument that justifies the interpretation *example: regional core top calibration equation (Barker et al., 2005*

chronData metadata metadata associated with chronDataTable:

chronDataTableName short name of the chronData table; *example: chronology*

filename chronology table filename (.csv); *example: Atlantic0220Thornalley2009Chronology.csv*

chronology short name of chronology table used in this measurement table; *example: chronology*

chronData table columns columns required in chronology table, will vary by archive and geochronological methodology type - *example for ^{14}C age.*

depth depth of sample/measurement

14CAge radiocarbon age

14CAgeUncertainty analytical radiocarbon age uncertainty

datedMaterial what was dated? (e.g., bulk sediment, terrestrial macrofossil, etc)

chronology column metadata metadata characterizing each column of the chronDataTable

column column number; *example: 2*

parameter short parameter name *example: 14CAge*

description longer description of the parameter *example: uncalibrated radiocarbon age*

units *example: 14C yr BP*

errorLevel error level for uncertainty columns *example: 14C yr BP*

4 Discussion

The data container and preliminary data standard described here are extremely flexible, and can accommodate any paleoclimatic or paleoenvironmental data that are based on any expansion of dependent/independent variable pairs. This encompasses all paleoclimate and paleoenvironmental datasets that we can imagine. The challenge for developing a sufficiently broad paleodata framework has long been 1) defining all of the relevant terms for such a diverse community and 2) managing the appropriate level of detail (lumping versus splitting) in the terminology. The framework presented here accommodates the first challenge by being accommodating the complexity and inevitable proliferation of terms, variables and interpretations inherent to the interdisciplinary field of paleoclimatology, and by assigning explicit meaning to the terms through the Linked Open Data framework. Implementation of these semantics will be an evolving, community-driven process. This is critical for two reasons: first, defining an ontology *a priori* has proven impossible to date; second, even if it were possible, such an ontology would be meaningless if it were not used. We will thus rely on usage and community discussion to reach agreement on terminology. The discussion may begin here, in the comments of this paper.

The LiPD framework accommodates the second challenge by adopting a hierarchical structure, that starts with more general terminology and allowing further detail to be assigned deeper in the structure. Consider the example of two $\delta^{18}Odatasets, onemeasuredonacoralarchive, andtheotherderivedfromforaminiferaextractedfromamarines$
$skeletalaragonite$"$andtheother$"$\delta^{18}O - foraminfera > 120\mu$m size class". By taking advantage of JSON's capacity to build hierarchical metadata structures, we can encode an entire set of metadata at the appropriate level in the dataset as:

```
{
    "parameter": "d18O",
    "description": "d18O measured on skeletal aragonite",
    "units": "permil",
    "standard": "VSMOW",
    "material": "skeletal aragonite",
    "instrument": "Micromass Optima gas source triple-collector mass spectrometer"
},
```

and:

```
{
    "parameter": "d18O",
    "description": "d18O measured on G. bulloides > 120 microns",
    "units": "permil",
    "standard": "VSMOW",
    "material": "foraminifera calcite",
    "instrument": "Micromass Optima gas source triple-collector mass spectrometer"
    "species":"Globigerina bulloides"
},
```

This makes the commonalities and differences between the datasets explicit. Moreover, additional levels of metadata may be introduced into the descriptor to accommodate climate interpretation, calibration procedures or forward modeling as described above. The power of the hierarchical structure is that it allows the metadata to be placed at the appropriate level, avoiding logical contradictions in lumping and splitting that become necessary when trying to incorporate information from several levels into a single term – or when several users describe the same dataset in slightly different ways.

An important consideration for re-use and provenance tracking is **versioning**: each version of a LiPD record, or collection of LiPD records, should be associated with a unique identifier, which is crucial to reproducibility. We propose the following preliminary versioning scheme:

Individual records A number of the form $I_1.I_2$, where I_1 is an integer associated with a publication (e.g.[8]) and I_2 is a counter updated every time a modification is made to the data or metadata.

Data compilations A number of the form $C_1.C_2.C_3$, where C_1 is an integer associated with a publication [e.g. 7] and C_2 is a counter updated every time a record is added or removed, and C_3 is a counter updated every time a modification is made to the data or metadata in an individual record.

We are presently implementing a large-scale test of the LiPD framework by using it as the primary data archive for Phase 2 of the PAGES 2k global temperature database. Consequently, the described framework for describing the proxy data is fairly mature and field-tested. LiPD's relatively complex structure is in fact the simplest container we could devise that would accommodate the wide diversity of constraints provided by the ¿5,000 records in the database. However, it may not be universal, and we welcome suggestions for increased generality.

The standards for reporting and storing geochronological data are much less tested and will require far more community input. For instance, there seems to be no universal way of reporting radiocarbon, U/Th, or ^{210}Pb dates. Ideally, coordination between geochronologists would yield a universal standard for all radiometric age models; however, if there is to be any standard, it is more likely to first emerge within each sub-community. We are confident that JSON-LD is flexible enough to encompass any possibility, but doing so in a way that allows research codes to easily read those chronologies and generate age models from them will likely require more work.

More generally, it is important to realize that the LiPD framework is not a rigid container that one must force paleoclimate data into, but rather a flexible system designed to wrap around a data set. We are keen to continue developing and expanding LiPD and the preliminary data standard with input from the community. We welcome any and all feedback on this framework, in particular potential problems that we have not foreseen. Please take advantage of the commenting system in Authorea to let us know what you think.

The authors acknowledge NSF grants EAR 1347213 and AGS 1003818 for supporting this work and PAGES for support coordination of community-based initiatives. This work received useful input from many collaborators, especially Bruce Bauer, Peter Brewer, Chris Brierley, Mattero Cantiello, Olivier Cartapanis, Mark Chandler, Ken Collier, Mike Evans, Doug Fils, Simon Goring, Wendy Gross, Darrell Kaufman, Guillaume Leduc, Graham Mortyn, Jess Tierney, and Gene Wahl.

References

[1] Stephen Barker, Isabel Cacho, Heather Benway, and Kazuyo Tachikawa. Planktonic foraminiferal Mg/Ca as a proxy for past oceanic temperatures:

a methodological overview and data compilation for the Last Glacial Maximum. *Quaternary Science Reviews*, 24(7-9):821–834, apr 2005.

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[3] Julien Emile-Geay and Jason A. Eshleman. Toward a semantic web of paleoclimatology. *Geochemistry Geophysics, Geosystems*, 14(2):457–469, feb 2013.

[4] P. Bryan Heidorn. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, 57(2):280–299, 2008.

[5] Tony Hey. The Fourth Paradigm – Data-Intensive Scientific Discovery. In *E-Science and Information Management*, pages 1–1. Springer Science Business Media, 2012.

[6] Thomas Johnson. Indexing linked bibliographic data with JSON-LD, BibJSON and Elasticsearch. *Code4lib Journal*, 19:1–11, 2013.

[7] PAGES2k Consortium. Continental-scale temperature variability during the past two millennia. *Nature Geosci*, 6(5):339–346, apr 2013.

[8] David J. R. Thornalley, Harry Elderfield, and I. Nick McCave. Holocene oscillations in temperature and salinity of the surface subpolar North Atlantic. *Nature*, 457(7230):711–714, feb 2009.

[9] E. R. Wahl, D. M. Anderson, B. A. Bauer, R. Buckner, E. P. Gille, W. S. Gross, M. Hartman, and A. Shah. An archive of high-resolution temperature reconstructions over the past 2 millennia. *Geochemistry Geophysics, Geosystems*, 11(1):n/a–n/a, jan 2010.