doi.org/10.22498/pages.26.2.62

# LinkedEarth: supporting paleoclimate data standards and crowd curation

Julien Emile-Geay[1], D. Khider[1], N.P. McKay[2], Y. Gil[1,3], D. Garijo[3] and V. Ratnakar[3]

**Data science in the paleo sphere has been hindered by a lack of standards that limit interoperability and interdisciplinarity. Here we describe the LinkedEarth project, which lowered some of these barriers, and offers a blueprint for further erasing them.**

At present, scientists are asked to upload their data to various "silos" (loosely connected data centers like WDS-Paleo[1], PANGAEA[2], or Neotoma[3]), which use different formats and conventions, hampering interoperability. Further, there is sometimes little guidance on what information needs to be archived to provide long-lasting scientific value. While paleosciences offer a long-term perspective on environmental change, this cannot happen without a long-term perspective on environmental data stewardship. LinkedEarth[4] (Fig. 1) was funded by the EarthCube program[5] as a two-year "integrated activity", with the twin aims of putting paleoclimate data stewardship in the hands of data generators, and developing standards that promote effective reuse. Here we review LinkedEarth's successes and outstanding challenges, and take stock of its broader lessons for the PAGES community.

LinkedEarth has acted as a laboratory to advance the notion of decentralized paleodata curation, allowing data generators to curate their own and others' data, via standards and technologies. The basic premise of LinkedEarth is that no-one understands data better than the people who generated them. Therefore, data generators should be the ones describing their data, but in a consistent way to make them interoperable. Having participated in several PAGES' syntheses (e.g. PAGES 2k Consortium 2017), we also appreciate that publicly-archived datasets are nearly always incomplete, and may harbor errors - requiring collective curation and correction (that is, the ability for multiple actors to edit and annotate the same datasets). We thus set out to develop a platform that would enable paleoclimatologists to interact with data in an intuitive way, resulting in standardized datasets that are (by construction) extensible, interoperable, and discoverable.

## Crowd-curation through standards

A data standard consists of three parts: (1) a standard terminology, to prevent ambiguity; (2) standard practices, which codify the information that is essential to long-term reuse and (3) a standard format for archival and exchange. The latter is emerging, in the form of Linked Paleo Data (LiPD[6]; McKay and Emile-Geay, this issue, 2016), so LinkedEarth only had to contend with the first two parts.

Standardizing terminology was accomplished by means of the LinkedEarth ontology[7]. An ontology is a formal representation of the knowledge common to a scholarly field. It allows unambiguous definitions of common terms describing a paleoclimate dataset, as well as the relationships among these terms (e.g. a proxy observation is measured on a proxy archive at a particular depth). Ontologies are necessary to organize information so machines can take advantage of digitally-archived data. Ontologies are inherently flexible, allowing to specify ecological properties such as habitat depth and seasonality to previously-archived foraminiferal-based records. Ontologies have had an enormous impact in biomedical research, ranging from genomics to drug discovery, and are beginning to permeate the geosciences[8].

Ontologies need to be sufficiently rigid so that dependent applications can rely on their structure being stable over time, yet sufficiently flexible to accommodate growth and evolution. The ontology maps closely to the LiPD structure, which serves as its stable skeleton. Extensibility was achieved via a new technology, the LinkedEarth platform[9]. At its core, it is a semantic wiki, similar to other wikis like Wikipedia, but based on the LinkedEarth ontology. The LinkedEarth wiki tracks changes and attributes them to authenticated contributors (an ORCID is all that is required to join LinkedEarth). The wiki facilitates extensions by allowing users to edit the non-core aspects of the ontology: they can define new classes or properties, create or change definitions, start discussions with other users, or request modifications to the core ontology when sufficient consensus emerges. These user roles and interactions were defined in a formal charter[10]. The flexible structure will accommodate advances in techniques and interpretations, and allow users to deprecate outdated terms.

Because LinkedEarth datasets are based on LiPD, they can be uploaded or downloaded in a few clicks, and benefit from

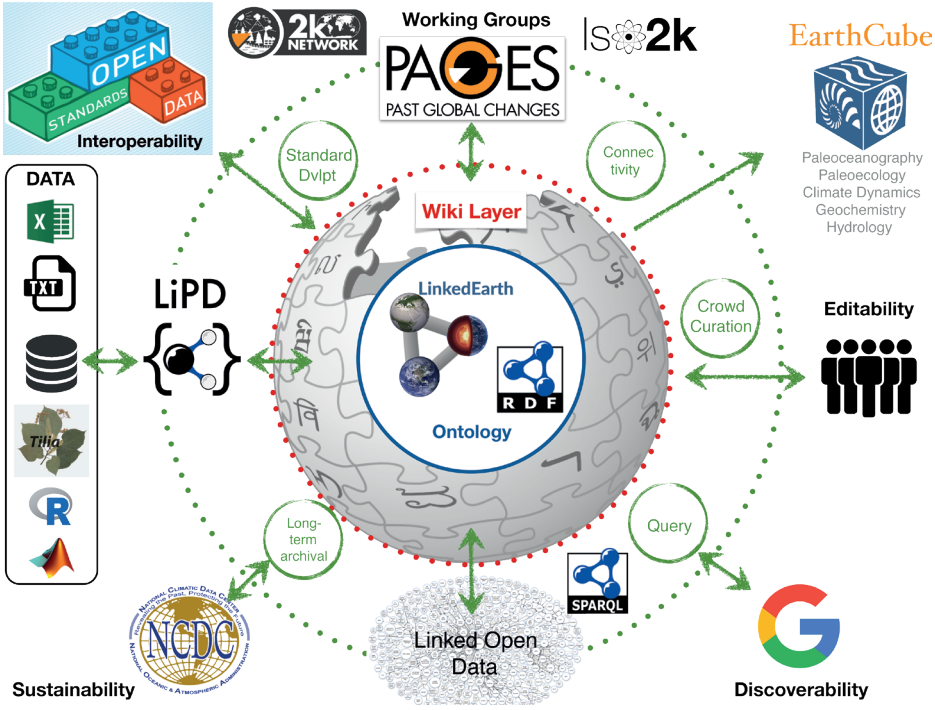


**Figure 1:** The LinkedEarth design (Gil et al. 2017) is structured around a semantic core (ontology), which can be easily interacted with thanks to a wiki. In addition, the framework supports import/export in LiPD, rich queries, the elaboration of community standards, the crowd-curation of datasets and a natural link to the Web of Data, ensuring discoverability.



CC-BY

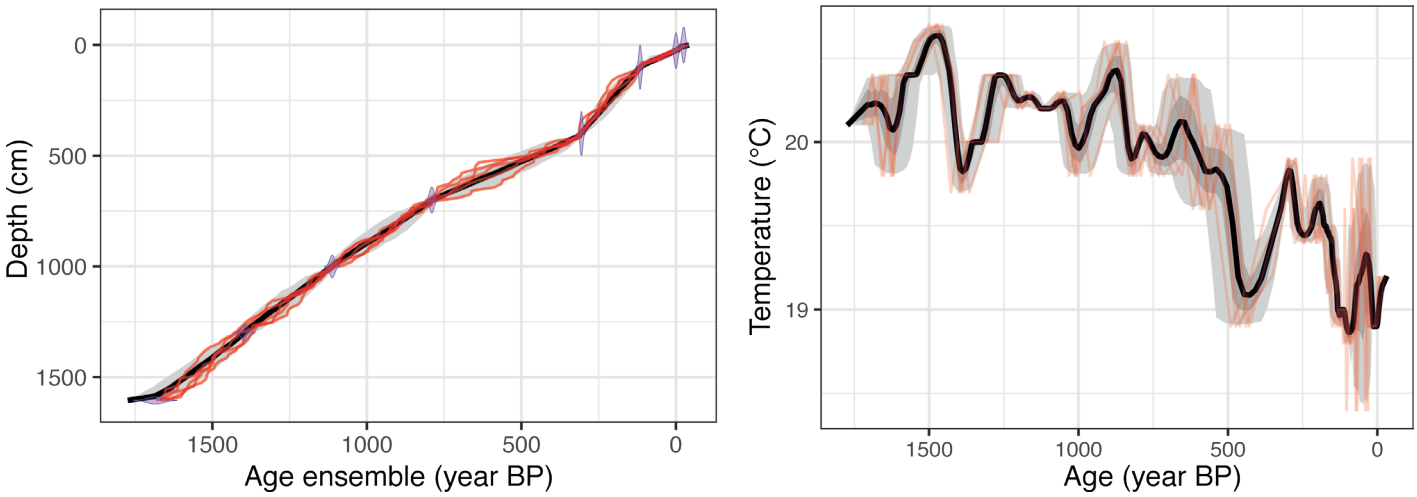


**Figure 2: (A)** Chronology for Basin Pond sediment core using the BACON software; **(B)** Basin Pond temperature reconstruction placed on an ensemble of age models (red line). The bold black line represents the median. The grey areas represent the 1σ (dark) and 2σ (light) uncertainty bands on calendar age. For details, see nickmckay.github.io/LinkedEarth-Neotoma-P418/LE-Neo_UseCase.html

the entire LiPD research ecosystem (McKay and Emile-Geay, this issue). This makes LinkedEarth-hosted data inherently ***interoperable***. In order to ensure the lasting utility of the data, LinkedEarth sparked the first international discussion on community-led data reporting standards[11], to build consensus on the most important information that should be reported in paleoclimate datasets. This consensus-building enterprise was facilitated by the LinkedEarth platform, including working groups, discussions, and polling (Gil et al. 2017).

Lastly, the semantic part of LinkedEarth means that datasets are broadcast to the web using standard schemas[12], which make them ***discoverable*** by various search engines, including Google. Because of this outward-facing design, LinkedEarth datasets were the first to be integrated into EarthCube's Project 418[13] (P418), an EarthCube initiative to demonstrate common publishing approaches for data holdings using such standard schemas.

**Towards Interoperable Paleo Data**

Discovering data pertinent to a scientific question is critical, but what to do once you find them? Imagine a user interested in the impact of time uncertainty on a pollen-based temperature reconstruction at Basin Pond, Massachusetts, USA. After a quick search through the P418 interface, our user realizes that the temperature reconstruction is hosted on LinkedEarth while the geochronological information is stored on Neotoma. Using the P418 service, they can find and download the datasets of interest. The GeoChronR software package (McKay et al. 2018) can then facilitate their analysis. GeoChronR was built around LiPD, which has been mapped to the Neotoma data model (that is, Neotoma datasets can be read by any LiPD-based code). This enables fast integration between the LinkedEarth-hosted temperature reconstruction and the Neotoma-hosted chronological data. Within the GeoChronR framework, our user has access to a variety of age-modeling tools, including Bacon (Blaauw and Christen 2011). They can then readily visualize the new age model (Fig. 2a) and assess the impact of age uncertainty on the temperature evolution (Fig 2b). Such is the promise of holistic data stewardship: more than putting data online, it's about drastically simplifying their reuse.

**Beyond LinkedEarth**

In a short two years, LinkedEarth has brought to life a functional platform for the crowd-curation of paleoclimate data and an emerging data standard. Along the way, it provided a nucleus for interoperability via synergistic software (GeoChronR, Pyleoclim[14]).

Despite these accomplishments, the vision still faces notable challenges. First, it has proven difficult to elicit broad participation: only 100 paleoclimatologists have answered our survey on paleoclimate data standards so far. We have found that overburdened scientists have little inclination to participate in such activities unless there are clear incentives. We argue that only publishers and funding agencies can provide these incentives, but have yet to do so. We do not envision meaningful progress until they do. Another issue concerns adoption: despite a non-trivial investment of resources (funding, personal time for participants), very few scientists are actively using LinkedEarth. PAGES is playing a leading role in incentivizing a new generation of paleoscientists to curate high-quality data compilations and take advantage of the LiPD-based research ecosystem, which was built for them. PAGES 2k[15] is a case in point, having motivated the birth of LiPD, the need for crowd-curation, and many of the ontologies' categories. One persistent obstacle to adoption is the perceived redundancy with data repositories. LinkedEarth is a framework, and works in tandem with repositories. It has strong links to WDS-Paleo, which now accepts LiPD as a submission format, and can ensure long-term archival. Because of LiPD's structured nature, LinkedEarth also integrates well with Neotoma; links to other repositories are in the works. The success of LinkedEarth will be measured over time by adoption and extension of its various tools and standards. We look forward to many more PAGES compilations being generated, discussed, and published on LinkedEarth. Every new PAGES working group brings with it new requirements; so far, LinkedEarth's intrinsic flexibility has enabled it to accommodate them all, and likely will for the foreseeable future.

AFFILIATIONS

[1]Department of Earth Sciences, University of Southern California, Los Angeles, USA
[2]School of Earth and Sustainability, Northern Arizona University, Flagstaff, USA
[3]Information Sciences Institute, University of Southern California, Marina Del Rey, USA

CONTACT

Julien Emile-Geay: julieneg@usc.edu

REFERENCES

Blaauw M, Christen JA (2011) Bayesian Anal 6: 457-474

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

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

McKay NP et al. (2018) GeoChronR (Version 1.0.0). Zenodo

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

LINKS

[1]ncdc.noaa.gov/data-access/paleoclimatology-data
[2]pangaea.de
[3]neotomadb.org
[4]linked.earth
[5]earthcube.org
[6]lipd.net
[7]climdyn.usc.edu/publication/leo
[8]geoscienceontology.org/about.html
[9]wiki.linked.earth
[10]linked.earth/aboutus/governance/charter
[11]wiki.linked.earth/Paleoclimate_Data_Standards
[12]schema.org
[13]earthcube.org/group/project-418
[14]doi.org/10.5281/zenodo.1205661
[15]wiki.linked.earth/PAGES2k