Clim. Past Discuss., 11, C2946–C2950, 2016
www.clim-past-discuss.net/11/C2946/2016/


Climate of the Past Discussions

# *Interactive comment on* "Technical Note: The Linked Paleo Data framework – a common tongue for paleoclimatology" *by* N. P. McKay and J. Emile-Geay

**N. P. McKay and J. Emile-Geay**

nicholas.mckay@nau.edu

Received and published: 19 January 2016

**We thank Ines for her thoughtful comments and suggestions. We respond point by point below in bold.**

*Overall I think the article presented by McKay and Emile-Geay reflects a great initiative that intends to facilitate research data sharing, discovery and reuse within the paleoclimate community. The authors are correct in stating that there is currently no universal way to describe, store and share paleoclimate data, which, unfortunately, also applies to most other research disciplines. However, the question is if this is due to the lack of suitable data formats, metadata standards and/or available infrastructure or due to*



*other aspects such as competition, giving other tasks precedence or simply unawareness of how and where to share research data. As I am not a reviewer of the article by McKay and Emile-Geay, I simply would like to leave some food for thoughts.*

*1. In the introduction chapter the terms data standard, metadata standard, data container and format are used fairly inconsistently. At times it is hard to follow if the authors talk about data, metadata or data formats. A brief definition of the individual terms may be appropriate to ensure a consistent understanding of their meaning.*

**Agreed. We will be explicit about how we are defining these terms, and check for consistency in our use of them.**

*2. While talking about universally readable data formats the authors mention netCDF format, which is a widely applied and accepted 'self-describing, machine-independent data format that supports the creation, access, and sharing of array-oriented scientific data'. Therefore, the authors choice of the JSON-LD data format strikes me as unusual as this format, at least to my understanding, is not widely known and applied. Introducing a largely unknown/unused format to the community might result in it not being readily adopted. 3. Regarding the proposed metadata standard I am wondering about the necessity to introduce yet another standard. On the one hand netCDF offers a metadata convention for e.g. Climate and Forecast (CF) data that is easily incorporated into the netCDF files (netCDF CF 1.6) itself. On the other hand there are several metadata standards specifically designed for geo information (e.g., ISO19115) that have several multi purpose fields that can house otherwise non assignable information.*

**Regarding the suitability and broad use of JSON-LD; we recognize that JSON-LD may be unfamiliar to most paleogeoscientists, however over the past decade JSON has become a leading format to describe and exchange all types data over the web due to it's simplicity, flexibility and functionality; features that are desirable for paleoclimatic data as well. The newer, linked data flavor of json allows**

**us to leverage the semantic web, which is a particularly exciting way to integrate paleoclimate data with concepts and datasets from other fields of science, that holds the potential to lead to more efficient interdiscplinary discovery. It's unfamiliarity shouldn't be a problem for users because 1) most users won't interact with the JSON-LD directly, as we will better explain in our reply, and 2) because those interested in digging into JSON-LD component with LiPD will find it both human-readable and simple, and will be able use the incredibly rich set of utilities available for working with JSON data.**

**Regarding the possibility of NetCDF as a format for paleoclimate data, we recognize that it is possible to use it, but there several reasons why we believe it is not an ideal solution:**

**1.NetCDF is used extensively by climate modelers, atmospheric and ocean scientists, but it is completely opaque to most paleoclimatologists. (We once described LiPD as "NetCDF for paleo" to a renown paleoceagranographer. Her response: "What is NetCDF? I only work with Excel"). For 90% of paleoclimatologists, NetCDF is equally or even less approachable than JSON-LD.**

**2. NetCDF is designed for large, spatially-gridded ("array-oriented") datasets. Although it could potentially accommodate paleoclimate data, it was not designed for the highly heterogeneous datasets common to paleogeoscience, and its implementation to such datasets would be very clunky. It would make sense to force a square peg into a round hole if most holes were round and very few pegs were square. However, NetCDF is non-existent outside of the climate modeling and atmosphere/ocean sciences communities. JSON, on the other hand, is used to exchange all manner of information on the Web, making it inherently much more interoperable. Indeed, JSON modules already exist in the vast majority of modern programming languages (see http://www.json.org/), unlike NetCDF.**

**3. Other paleogeoscientists (e.g. paleoecologists and paleobiologists), share**



**many of the same challenges that paleoclimatologists do. Organizations like Neotoma, Ocean Drilling Program and the Paleobiology Database have expressed interest in the LiPD framework, which they are considering adopting for their own purposes. This would make all paleogeoscientific data much more interoperable. NetCDF never had that pull with these communities.**

**4. While some protocols do exist to extract part of a NetCDF file via a network (.e.g OpenDAP), this information generally cannot be found unless one knows what to look for. In this sense, NetCDF files essentially create news silos (clean, organized and self-describing silos, but silos nonetheless). In contrast, JSON-LD is designed around the concept of Linked Open Data, which makes datasets instantly discoverable on the Web by broadcasting the metadata via a universal protocol (the Resource Description Format). Many organizations already archive content as Linked Open Data (e.g. the BBC, the NY Times, DBPedia), and doing so with paleoclimate data would enables it to be more easily cross-referenced with a continuously growing body of knowledge across the Web.**

**5. An important component of the paper is that it promotes a more structured way of storing paleoclimate data. Once this structure is achieved it is easy to translate it to other formats. If a NetCDF implementation is deemed critical, one can always write code to export from LiPD to NetCDF.**

**Ultimately, NetCDF is the outcome of a user group of scientists developing a data format that richly fits their need. Rather than squeezing paleoclimate data into NetCDF; the community should take the opportunity to develop the format that best fits the need of paleogeoscientists.**

*4. Regarding the unique identifier I would recommend looking into Digital Object identifier (DOI) that could either be associated to the individual data set or to the data collection. Most countries have central agencies, universities or research institutions that provide DOI minting services.*

**We agree that data DOIs would be ideal solutions, and LiPD can readily encode data DOIs as a key value pair. We however, believe that minting DOIs for datasets, rather just encoding those identifiers that were minted at institutions like the ones you described is beyond the scope of the what we're aiming to describe here. That said, we will join you encouraging scientists and repositories to adopt data DOIs for their datasets.**

*5. To link information from data files, metadata, authors, publications and grants I would recommend looking into the infrastructures that are already in place in universities and/or libraries. These often fairly sophisticated systems have been set up for exactly such purposes. In general it might be a good idea to involve a (local) liaison librarian, a member from the universities eResearch group (if existent) or research office, as they are often familiar with issues related to data sharing, storing, discoverability, linkage and reusability, and are potentially a good source of information. I also would recommend looking into Research Data Alliance as they intent to 'build the social and technical bridges that enable open sharing of data' across various disciplines*

**Thank you for this suggestion. LiPD was designed to link into existing data stores, and we will investigate the possibility of extending these data connections into LiPD.**

---

Interactive comment on Clim. Past Discuss., 11, 4309, 2015.