Clim. Past Discuss., 11, C1763–C1773, 2015
www.clim-past-discuss.net/11/C1763/2015/

Climate of the Past Discussions
Open Access

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: 23 September 2015

We thank the reviewer for bringing up several concerns with the data format and community use and support for Linked PaleoData. All of these concerns are easily addressed - indeed most have come up before in previous iterations of the discussion and feedback with the community. In this technical note we focused on the technical details of the format and the standard, without addressing the larger context of how LiPD would fit into the workflow of individual users, or the extent to which the community has been involved in its development. The review makes the valid point that we did not adequately make these points clear in the

manuscript, and will attempt to do so here and as well as in a revised version.

For ease of reading, we will include the reviewer's comments in italics, and our replies in bold.

1 The manuscript presented by McKay and Emile-Geay is putting forward a proposal for definition of a palaeoclimate file format for data exchange. This is done on the basis that a) it is arguably difficult to exchange large amounts of palaeoclimate data in a consistent way b) this prevent the automatic construction and parsing of large (proxy) databases c) the underlying format for the data and meta-data is researcher dependent, creating a large number of possibilities. The proposed format to be used is the JSON-LD format (a JavaScript originated avatar) for the metadata overlying a CSV (Comma Separated Value format) data file for the actual proxy data.

2 In the following I will refer to "data" without further mention as a shortcut for "Paleoclimatological Proxy Data", covering potentially the proxy data itself, the age models etc.

3 Several questions need to be stated when thinking about data exchange. The first one is obviously to define the communities (data gatherers, data users, data compilers, data providers) and their work methods to check the adequation of the proposed format with the present routine work with minimal adaptation. Second is the adequation of the format proposed to the work it should be used for. Third is the community support that is gathered around that proposal.

4 Taking up the first community type: data gatherers. The generation of proxy data is generally ending up in the production of non standard excel sheets (or other equivalent non-proprietary format, but the latter is very rare), that are produced on different operating systems. The proxy data gatherers (or producers) are themselves not using any other format to my knowledge for their daily work. Hence, creating all their work in a format that is not natively supported by an office production suite (without reference to a proprietary system) is not likely to be largely adopted. The underlying format is

CPD

11, C1763–C1773, 2015

hence necessary to be compatible with this.

We recognize that data producers predominantly rely on Excel or spreadsheet software to report and manage data. The expectation is not that data producers will be entering the data directly into LiPD; indeed, we have developed utilities to import Excel records into LiPD, and this was the primary way that the PAGES2k were translated to LiPD. However, we also emphasize that innovation may require the use of new technology. This should clearly be a gradual process, and we suggest that the local formats that data producers use to create datasets are excellent for record development, but should not be used as the format to archive data or to pass published data between scientists. Ultimately, we envision web-based and open-source desktop tools that will also users to translate their Excel spreadsheets into LiPD. But first we need to publish and iterate on this format.

5 Furthermore, the format for actual data (measurement) exchange requires an excellent portability of the underlying format to any system or language (by the latter I infer the common computer "localization" problem). It thus requires a format that do not include any structure that could be misinterpreted in that regard.

6 Data users can be other data gatherers, data compilers or other communities (e.g. modellers). For other data gatherers, the exchange format is non-critical since the common use of the excel-type format is very comparable in all communities and therefore do not hamper the data exchange. For data compilers, the issue is more complex: it requires to use a common format indeed (as pointed by McKay and Emile-Geay) for the actual measurements and for the metadata. As an example, the largely reknown PANGAEA database is using HTML metadata and TSV (Tabulated Separated Values) text files for the actual data. For the other communities (in particular the modellers or the model and data intercomparison group), the expectation is to have a format that is robust (as opposed to error-prone) and easily transferrable.

7 The JSON-LD format proposed for the metadata is a very peculiar choice. Though

CPD
11, C1763–C1773, 2015

is known for being relatively lightweight, it is very seldom used to gather databases. In the instance of proposing a common language in the view of creating large databases, it is a very contradicting approach. The only long-lasting advantage I can see in this choice is the capability of being easily read when opened in a simple text editor (e.g. Notepad or emacs).

This is an odd objection. JSON is a leading format for data exchange on the web, and extensible databases are becoming increasingly common for datasets that are not well suited for relational databases. Paleoclimate data clearly fall into this category. For extendable databases JSON and XML are essentially the two choices, and the simplicity and human-readability of JSON, as well as the rapidly evolving community support for the format makes it preferable to the far more verbose XML choice. Besides, the volume of paleoclimate data is not ever expected to be something so large that JSON would not be able to handle it. Certainly, it is equally as capable as XML on this front.

8 On the over hand, thinking about the daily work of a data gatherer, this format is largely at odd with the tool used. There is no simple way to create a template or better input-ready form that can be used in an office suite. I have made the simple test to open the provided jsond file in an office suite software and the result is not something you will want to work with on a daily basis (I encourage the editor and other reviewers to perform that simple and convincing test).

Again, our intention, which we didn't make sufficiently clear in the manuscript, is that data producers will not be editing the json-ld file directly. The standard is meant to be explicit and self-describing, so the meaning is built into the file in a uniform way. This necessarily makes it not ideal for manual development, which is why we are continuing to develop tools for input and output from various platforms. We did not expand on these tools in the paper as we wanted to focus on the format itself, but if the reviewer think it is appropriate, we would be happy to discuss these tools in the paper.

CPD
11, C1763–C1773, 2015

9 The choice of CSV for the underlying data is simply disastrous. There is no formal description of what the CSV format is internationally and any software can decide the method, separators, text identifiers etc. that is to be interpreted. The method to be used by a software to open a CSV file depends on the localization characteristics of the medium of support!!. To put that in plain language, depending on the support of origin of your file (usb key, data hardrive, http etc.) the software will interpret its localization differently. If I open a CSV text on a french hardrive that contains http UK transfered data files, themselves generated in a lab in Danemark (eurocentric arbitrary choice of countries) the definition of the localization for the file is likely to be undefined, or defined by the last medium! This is not acceptable as an international exchange data format.

We have three objections to this comment:

1. **Many of the localization problems are due to how software encodes their CSV files, not the CSV files themselves.**
2. **Whereas CSV doesn't have explicit standards, or we are following the conventions of the W3C's CSV on the Web working group.**
3. **Lastly, the localization options of the CSV are readily described in the json file that provides "instructions" on how to parse the data. The choice of delimiter is readily handled there. Again, it is important to recognize that we don't expect most users to interact directly with the data; rather, they will access the actual data via their platform of choice (e.g., Excel, R, Matlab, Python), from files created by input/output utilities. Matlab and R are presently operational, the others are work in progress.**

10 A common example for the localization problems are the commas. On an english system, the comma might define the field (in CSV, Comma Separated Values) or the

CPD
11, C1763–C1773, 2015

marker for thousands. On a french system, the comma defines the decimal. It naturally follows that a CSV file generated on a french system is not CSV purely since it is filed separated by a semicolon. This is most impractical and will lead to many, many errors.

See comment above. The semantic value of the Linked Data framework means that the data are shipped with 'instructions' on how to be properly read. However, we should have been more explicit about our intent to make other user-friendly files available.

11 Community support. I am extremely surprised to see that the proposal is presented by only two individuals and acknowledged, apparently, by the PAGES2K group. If that manuscript is the result of a group effort, why is this not a group co-authorship? If it is a individual proposal, how shall we expect to gather the group support and momentum that is required for the definition of something that could become a standard in the long-run? This (apparent) absence of community support and the statements made by the authors that " One goal of the present work is to spark such a discussion by giving the worldwide paleoclimate community a strawman to improve upon." makes this manuscript very inconsistent. A strawman is defined as being an "informal fallacy". Shall we, as a community accept that fallacy? Sparking a discussion is a good idea, but ideally this should not come as a formal published definition. Publishing formats first with the fallacious idea of suggesting a discussion will only lead to: a) a few groups will adopt it (like PAGES2k) b) some will not and either keep the usual format or adopt another one c) this will add layers of confusion, noise in the system and even more components to the "Digital Tower of Babel [sic]".

This is a major misunderstanding, that should have been made explicit in the manuscript. The described data format, as well as the manuscript itself, is the result of multiple iterations of community development. Certainly, some description of this should be added to the manuscript, and we're happy to do so; but let us be explicit about the community development that has supported the development of LiPD to its current point.

CPD
11, C1763–C1773, 2015

The reviewer acknowledges the recognized need for a standardized format for paleoclimate data; and the earliest development of these concepts arose from the clear recognition of such a need through two large community projects organized through Past Global Changes (PAGES), the PAGES 2k project, and the PAGES Arctic Holocene Transitions project; the authors have been involved with these projects since 2012. The call for standardization from the community working on these projects was clear, and PAGES has made the development of formats and standards a priority as part of its "Data Stewardship" integrated activities effort. To this end, the authors have worked with the PAGES International Program Office to reach out the large community of paleoscientists involved with PAGES, to solicit input and feedback on these ideas. Explicitly, through PAGES we reached in multiple ways:

- **2k Mailing lists mail to > 600 subscribers, 23 April**
- **e-news went to 5400 PAGES subscribers. It was e-news 3-2015**
- **Tweeted on May 8, https://twitter.com/PAGES_IPO/status/596604952363012096**
- **FB entry on May 8, https://www.facebook.com/PastGlobalChanges/timeline/**
- **"Latest news" entry on PAGES website, May 8: http://www.pages-igbp.org/news/all-news-items/9-latest-news/1162-give-your-feedback**

For the most part, we relied on the online platform Authorea, which allows online publishing, editing and feedback on manuscripts, https://www.authorea.com/users/17200/articles/19163/_show_article to share the information on this format and receive feedback. Through this process we received excellent feedback from

CPD
11, C1763–C1773, 2015

the community that great contributed to the framework. We view this as community product that has been evolving, and that we expect will continue to evolve through discussion and feedback here on Climate of the Past Discussions.

A couple more notes:

1. The term strawman was used in the sense of strawman plan (https://en.wikipedia.org/wiki/Straw_man_proposal), as in a first cut to be refined (or rebuilt altogether) based on subsequent input. Community feedback has been overwhelmingly positive, so we do not believe that the reviewer's perception is universal.

2. As for authorship, many community products evolve much more efficiently when a full fledged proposal is available to evaluate, rather than by trying to build consensus from the earliest stages. The authors developed this framework, and invested many research hours into it, and it is reasonable to publish it as their idea. It is a community product, that was greatly helped by extensive helpful advice from many community members, who acknowledged in the manuscript, as well as PAGES who engaged the community on our behalf. There are no clear guidelines about what constitutes authorship on such a product; nor would a long list of authors change the extent to which this is a community-developed product.

12 Overall, I find that the process and the definition outlined in the present manuscript are very much ill-fitting the purpose and should not be accepted. I urge the authors to seek some real community discussion and community support before engaging into formalizing it. A common group approach is to open a WiKi space, discuss over a few, tryout the definition, refine the norm definition and finally propose it for publication as a large group effort.

Again, extensive community support was involved in the development of this data structure. Climate of the Past Discussions is an excellent format for contin-

CPD
11, C1763–C1773, 2015

uing this refinement, and we appreciate additional comments from the broader community here. The beauty of this framework is that it can evolve organically based on community input and we look forward to taking even broader feedback to continue to evolve it to fit the needs of the community.

13 A better text delimited file format. Most of the databases (NOAA, PANGAEA) are used Tab Separated file formats, or TSV. Though being not better defined than CSV in principle, it has the massive advantage of using a simple character "tabulation" that is not used in common writing language for unique separation of fields. From a computer language perspective, it is far easier to use. Adopting in within a utf-8 coding norm is sufficient and universal enough for the underlying data.

As written above, the CSV "problem" is a red herring, so there is no need to fix it. Although we are not convinced that TSV is fundamentally better than CSV given the discussion above (and the reviewer's own admission that tabs are not universally parsed, either), this could be simply updated if its superiority can be unequivocally proven - an example of how the structure can evolve if need be.

14 A better metadata format. The most common format used for metadata description is XML. This format has the massive advantage of being readily compatible and translatable in HTML format (an enormous advantage for user-parsing of databases), is supported by office suite software for creating standard forms and templates, both as input and output format. It is already used in many large databases in the world and hence its wide support is not to be proved. It would solves many of the problems outlined here.

XML is of course another option – indeed, one of the authors (Emile-Geay) originally thought it would be suitable (see Emile-Geay & Eshleman, G^3, 2013). However, the simplicity and human-readability of JSON are causing it to become a leading format for data exchange on the web, with extensive community support for tools on all software platforms. The importance of XML's support in HTML

CPD
11, C1763–C1773, 2015

and Office is overstated, as users are not expected to be editing the documents directly. Moreover, editing such documents in Office would not be an appealing or practical option, in addition to the fact that Office is not open source, and that its attendant formats keep changing (e.g. xls vs xlsx), preventing backcompatibility.

15 On the linked data approach. I have to admit I did not understand the point that the authors are trying to make on that aspect. Though I perfectly understand the advantages of linked data (like in HTML hyperlinking or XML relationships) I do not get what Figure 2 is trying to show us. The manuscript is particularly not-explicit in that regard. It is meant to "illustrate the standard more concretely". Apparently it just confuses some readers.

Figure 2 is not meant to illustrate the linked data concept, rather to show the hierarchical structure of the metadata for a real-world example. We will make this more clear. For more information about Linked Open Data, we recommend the following resources:

http://linkeddata.org/

http://www.mkbergman.com/447/what-is-linked-data/

https://vimeo.com/36752317

http://www.ted.com/talks/tim_berners_lee_on_the_next_web?language=en

16 Following all the deficiencies of the approach listed above, I can only but recommend straight rejection of the manuscript. The problem stated is very important though and should be taken up by the community, but as a community-wide approach.

Again this is a misconception, that we need to make more clear in the manuscript. The Authorea comments provide examples of real community feedback. We also suggest that progress and innovation is not always most efficient in the largest groups; and large committees are often far better at rejecting ideas

CPD

11, C1763–C1773, 2015

than developing solutions. We suggest that this is a case where a proposal needs to be put out to the community. Indeed, the extant literature on the diffusion of innovations shows that innovations are proposed by individuals or small groups, and are either adopted or rejected by the community. We are hereby putting this field-tested (cf PAGES2k) innovation to our community members, and are developing tools so they can try it in their own workflow. This is "a community-wide approach", although we acknowledge that it is (and argue that it needs to be) a different community-wide approach than the reviewer suggests.

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