Clim. Past Discuss., 11, C1705–C1709, 2015
www.clim-past-discuss.net/11/C1705/2015/

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

Anonymous Referee #1

Received and published: 18 September 2015

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 hence necessary to be compatible with this.
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 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).
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).
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.
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 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.
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]".
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.
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.
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.

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.

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*.

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