Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-195-RC1, 2020

Earth System Science Data Discussions

Interactive comment on "SoDaH: the SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0" *by* William R. Wieder et al.

Jeffrey Beem Miller (Referee)

jbeem@bgc-jena.mpg.de

Received and published: 10 November 2020

General comments

The authors present a new database (SoDaH) of soil observations synthesized from datasets curated by five well-known research networks: LTER, CZO, NEON, NutNet, and DIRT. Two key aspects of SoDaH make it unique among the new generation of soil databases: 1) flexibility in its approach to harmonizing data from diverse sources, and 2) the development of a powerful web-based tool for querying, visualizing, and extracting subsets of data. These two features of SoDaH make it a valuable addition to the growing pool of soil databases available to the soil science community.

The manuscript is generally well written, and clearly presents the need for more accessible sources of compatible soil data to facilitate broad-scale syntheses, as well as the challenges of providing such a resource. However, the description of the process of data harmonization and aggregation is somewhat confusing in the text (Fig. 3 provides an excellent visual summary). Please see the specific comments for more details. Additionally, the metadata template only provides sparse instructions for how to properly fill it out. I had to carefully compare a few of the source data files with the accompanying (filled out) metadata template in order to understand exactly what data was required for the metadata template and in what format it should be provided. A more thorough guide or an additional supporting document would improve this process. For example, the ISRaD database (Lawrence et al., 2020) mentioned in this manuscript provides both a template file and a supporting "template information file" to facilitate data entry. I would recommend implementing a solution along these lines.

One important feature of a synthesized database is data transparency, which SoDaH excels at in some ways, but falls short in others. The workflow for aggregating data preserves the raw data, which is the gold standard. However, with the way the metadata template is structured there is no clear way to document the data source for site level data, which data contributors enter manually. As multiple contributors could provide data from the same site, it seems possible that conflicting data could be reported for, say, mean annual temperature. How would users distinguish which reported value is more appropriate for their analysis?

Finally, a critical feature of SoDaH is the web-based tool for querying and generating reports from the database (Shiny app), but unfortunately the use of this tool is neither well documented in the text, nor by the supporting resources (with the exception of an hour-long webinar available as a downloadable video clip, which while very useful, is not very user-friendly). Underselling this extremely powerful feature of SoDaH is in my view the biggest shortcoming in SoDaH as presented here. Providing a simplified overview or vignette that gives an example of the kinds of queries that can be made

(filters, etc.) and the reports that can be generated (visualizations, downloadable .csv tables, maps, etc.) would greatly help with reception and use of SoDaH within the community.

With minor improvements to the clarity of the text, and some additional documentation of the usage of the Shiny app, I think the manuscript is an excellent candidate for publication in ESSD.

Specific comments

Lines 58-60: Terms such as “harmonize” and “automate” would benefit from explicit definitions, although understandably this may not be possible with the word limits of the abstract (perhaps in the main text?).

Line 83: It would be helpful to expand on or quantify exactly what you mean by “similar data products”. Expanding on the importance of public availability of these databases would also be helpful.

Lines 123-130: Fine as is, but perhaps this information could be simplified in the text and expanded on in a table?

Line 145: While the framework for reporting data from experimental manipulations is a key asset of SoDaH, it is not clear why SoDaH allows for a greater range of spatial and temporal data than other databases.

Line 168: Suggest moving “Fig. 3” to the end of the sentence (after “...the structure of SoDaH”)

Line 180: The use of “ontologies” in this context is not entirely unclear and sounds like jargon. Additionally, what does “automatic harvesting of data” mean? Do you mean something more along the lines of automating the process of data acquisition? I realize those are similar, but the meaning is not clear from how it is written.

Lines 182-185: It seems like the “site” (or “location”?) is the fundamental organizational

unit of SoDaH. It might be helpful to state that more clearly and expand on the example provided in order to help readers understand how to define a site/location and how that definition relates to data organization at each level of SoDaH, i.e. from raw data to querying the aggregated database.

Lines 186-189: This section is not clear to me. What do you mean by the statement that data on the location tab are “broadcast to every row of the harmonized dataset”? The analogy that clarified this somewhat for me (used elsewhere in the manuscript) is the idea that the profile tab is a “map” for matching variables in the raw data onto the standardized variables in SoDaH. Additionally it is not clear why or when it would be appropriate to move data from the site to the profile tab.

Line 194: Can you expand upon (either in the text or by providing an example in the supporting information) how one would go about describing additional aggregation steps and how that would be implemented in the data aggregation process? This seems like a very messy and case-by-case basis, but also like a problem that would be encountered fairly frequently.

Line 200: When and how (what platform) would users “point to the Google Drive directory”? I assume this means when running the function in R?

Line 201: Suggest “...generates a new flat file(s) in which the relevant variable names and units are standardized...”

Line 202-206: If possible, it would be helpful to define or clarify some of the terms you use throughout this section in advance, e.g. “harmonized dataset”, “Level-1” data products”. The process is very clearly shown in Fig. 3 along with the terminology, so perhaps you could give a one-sentence description of the workflow in which you name the outputs of each step of the process?

Lines 208-212: This may not be the best place for it, but some discussion of data transparency would help to showcase the strengths SoDaH. Aside from the issue of

site-level data lacking a clear source, the preservation of raw data in SoDaH is a valuable feature. However, it is not clear from the website how to access the raw data files (I was able to find them, but it wasn't simple). Perhaps this could be clarified or stated explicitly somewhere?

Line 239: Clarify in the template how to specify these grouping variables.

Lines 251-255: Perhaps histograms of these data could replace Fig. 5? In its current format Fig. 5 is completely illegible.

Table 1: Consider providing some examples of gradient studies or time series.

Figure 3: Excellent figure, very clearly describes the data workflow.

Figure 5: This figure is illegible and it is unclear what it shows. Suggest replacing with histograms of site characteristics (see comment for lines 251-255).

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-195, 2020.