Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-195-RC2, 2020

Open Access 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.

Anonymous Referee #2

Received and published: 3 December 2020

I very much enjoyed reading about and exploring the new database: The Soils Data Harmonization (SoDaH) database. SoDaH is a valiant effort to combine the soil carbon data from three massive scientific efforts (LTER's, CZO's, and NEON) and to create a database structure that allows for time series and experimental data. Gradient data were also mentioned as something new to include though I do not see where gradients would have had trouble fitting into existing database structures like ISCN. The database uses a similar hierarchal structure to existing databases such as ISCN and ISRaD, and the familiarity should help the greater soil carbon community both contribute data and use the database.

Overall the description of the database in this manuscript was pretty clear in terms of how the database is structured (with the exception of layers). I did find, however, that more information is needed about the expectations of data contributors and users. I will go into more detail on that below. Lastly, I applaud the inclusion of the web-based shiny app. I enjoyed exploring the data with it, and I think it will help people easily see whether the data that they seek exists in the database and if it does, what the data coverage is. I recently spent a long time struggling to access and understand the data from a certain plant trait database, and I could see how the experience would have been much better with a shiny app. I will warn the authors that my comments go beyond the paper, to the webpage, shiny and git repository. With these ESSD papers that one has to evaluate the whole package.

Line Edits

Line 55: Get rid of comma on after "Synthesizing these data"

Line 135-140: The description of ISRaD makes is sound like 13C was a goal of ISRaD, though in reality ISRaD focuses on radiocarbon and includes 13C data if available, but datasets with only 13C data were not targeted. Furthermore, ISRaD includes 14CO2 data from gas wells, incubations, and fluxes. I think a more accurate description would be "radiocarbon from bulk soils, soil fractions, and soil gases."

Line 167: Is raw data the correct term here? To me raw data implies that the data is straight from an instrument and may still be in peak heights or areas and not corrected to actual carbon values. However, I am not sure what would be better to call it.

Line 171: I think what you mean by "layer" should be described here. It is also unclear how the layer fits in within the profile tab, or is it its own tab? It is hard to tell because it is a different color than profile in figure 2. I guess if there is no fraction data, then layer does not need to be its own tab but there did seem to be fraction data included based on the fields in the shiny app.

Line 182-189: More concrete examples might be helpful here as it seems to me that some studies will only have a single location to describe (an experiment) and then the treatments would be described in the profile tab, but a gradient study might have multiple location tabs or would the lat and long fields have to be moved to the profile tab in that case? I think the latter is described on line 189, but clarification would be good when it comes to gradients. For NEON data is every terrestrial site in its own google drive folder as single locations or are they all combined into one folder?

194: Can you define what you mean by "stacked". I am pretty sure it means that the from the same experiment the soil carbon and nitrogen data would each get its own line if they were on separate raw data files. This seems to be another case where a description of a concrete example would help.

197: It is unclear who the intended users of the soilHarmonization R package are. Is it the database managers or are the data contributors expected to use this package?

210: Why is the dataHarvest function not part of the above R package? Or is it? Again, is the data contributor expected to use this function after submitting data via their google drive folder? If they are not, who views the QC? Would it be best for the data contributor to view it since they know their data best?

225: I did not see many R scripts in this git repository, which seems to include the main paper. Is this the right address?

240: For users of this database, how can they access the grouping variable information? It does not seem like users can view these templates directly? Or maybe they can, and I just could not find that info?

279: Future contributions from who? Who will be overseeing this database? Is there a steering committee or manager? How will succession in such positions be handled?

280: It was hard to find how to contribute data on the website since it was towards the bottom of the database tab, maybe make it its own link at the top like Authorship is?

Also looking through the instructions it was not clear how to handle layer. Maybe it's just me, but a description of a study and an example of a filled-out template could be helpful here. I am really stuck on how layers should be described.

Figure 1: Can DIRT and nutnet also be touching the green circle because they are manipulations?

Figure 2: There are two locations shown here. Do they each get their own Location tabs?

Figure 5: Can the depth axis have units or at least put the units in the caption?

Other questions:

Where are these level 0 data stored? It seems like the contributions are given via users' own google drive folders, so that does not seem very permanent.

The authorship process is very clear on the website and seems to pertain to future users of the data, but the policy is not mentioned at all in this paper. Should it be?

For the Shiny app, I wanted more information on how to interpret each dataset's (level 1) QAQC. I looked at data I am familiar with and could not really understand what the graphs were trying to show.

Is there a way to only download the data that you query in the shiny app? Or could the shiny app show the code used for a certain query to help the user subset the downloaded database in R?

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