Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-195-AC2, 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.

William R. Wieder et al.

wwieder@ucar.edu

Received and published: 24 December 2020

Comments from Anonymous Referee #2 are provided below in normal text. Our responses to each are below each comment in bold with suggested changes to the revised manuscript identified by text in quotes. Note, this is easier to see in the .pdf files we've uploaded as a summlement along with this plain text response.

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. Thanks for this supportive comment. We note that Reviewer #1 raised similar questions, which we clarify: "Data from these kinds of studies (including gradient studies) should be incorporated into existing database structures, like ISCN, but the additional metadata requested as part of SoDaH helps database users understand more information about how data were collected from individual studies."

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. We are happy that you explored and appreciated the "unpublished" features of SoDaH that we have created to facilitate use of and, hopefully, contributions to the database. As much as possible, we have taken the suggestions provided, which are summarized below.

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

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.” We’ve changed the description of ISRaD to include “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. This is tricky, and now define “These primary data may or may not be in a published state but, if not published, would be equivalent to data provided for publication”

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. This is illustrated in Figure 2. These are good questions we seek to clarify with the following revised text.

“To simplify the workflow for data contributors, the metadata template only includes a single tab each for location and profile data. Within these tabs, data contributors are able to add information on metadata (found on the ‘location’ tab) and layer or fraction data (found on the ‘profile’ tab; Fig. 2). Layer data includes information on soil chemical and physical properties that may be measured on bulk soils for defined soil horizons or depth increments. Fraction data would include similar measurements on defined fractions within individual soil layers (e.g. percent soil organic carbon on density fractionated soils). Note, SoDaH currently has sparse data from measured soil fractions, which have therefore been omitted from Fig 2 for simplicity, but the database structure can include information on soil fractions.

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 mul-

tiple 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? Reviewer 1 raised similar concerns. We agree, more examples would help clarify this text: “The metadata template matches site-level information with the detailed measurements collected at each study site. Data on the location tab represents site characteristics for a single site or location (e.g., Prospect Hill Warming experiment at Harvard Forest). Accordingly, the harmonization script broadcasts data provided on the location tab (latitude, longitude, mean annual temperature, etc.) to every row of the harmonized dataset. Data on the profile tab includes profile information about experimental levels (e.g. plots within experimental blocks) and experimental treatments (e.g. +N fertilization) that help clarify how the data were collected. Data on the profile tab should also correspond to columns of variables that are reported in the Level-0 data (e.g., soil organic C measured at different soil layers). Accordingly, the harmonization script copies each unique measurement from the profile tab into a column of data in the harmonized dataset. Data contributors, therefore, can move variables from the location to profile tabs when appropriate. For example, NutNet and NEON data were submitted to SoDaH with information from multiple sites on a single .csv file that provided information about each site as unique columns of data. We, therefore, moved site information (e.g., climate, latitude and longitude) onto the profile tab for these networks. Similarly, gradient studies that report tabular data for individual soil profiles can move information on slope, aspect, vegetation communities or parent material (typically on the location tab) onto the profile tab of the metadata template.”

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. Another good suggestion we seek to clarify: “However, because SoDaH is a flat database values from these different data

files will be stacked, meaning that information from different Level-0 datasets would be recorded in different rows of the aggregated Level-2 database (in the example above, soil properties and productivity will be included, but in different rows). Additional aggregation steps, therefore, may be required to align data within sites. This can be accomplished with information from experimental levels and experimental treatments."

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? Either would be appropriate. "The package includes functions that harmonize Level-0 data into Level-1 data. Data contributors or database managers use the data_harmonizaiton function tools to read and harmonize user-provided raw data that are mapped to a metadata template with controlled vocabulary and standard units (Fig. 3)."

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? This function is not part of the package above, as "This function is intended for use by database managers". But the repository with this function is provided.

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? This comments to the harmonization package (previous paragraph), which we clarify "These Level-1 data products are stored in the same Google Drive directory as the Level-0 data with resulting output identified with a modified filename. This allows data contributors and database managers to verify the QC report and ensure appropriate data harmonization."

225: I did not see many R scripts in this git repository, which seems to include the main paper. Is this the right address? More scripts are available in the main repository, https://github.com/lter/lterwg-som, but the link provided in the manuscript is intended to include more curated examples, especially from published papers. Since this is the

first SoDaH manuscript we don't have much to share yet.

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? "Users can find this information in the database column labeled merge_align, which is a logical that identifies if multiple data files can be merged. Notes under columns align_1 and align_2 are intended to help communicate what common data fields can help with this alignment (e.g. experimental or treatment levels, L1 and tx_L1, respectively). To help users understand the database column information, the complete database key is provided in the SoDaH online application and gives users descriptions of the column contents."

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? As mentioned in section 2.4, future contributions of code to analyse the SoDaH database are encouraged. These contributions should be made to the LTER SOM GitHub repository, with a priority on developing additional utilities to align and aggregate datasets from individual sites and locations. Contributions will be reviewed by the SoDaH steering committee (currently Wieder, Pierson and Earl) and made publicly available. The committee will continue oversight while new funding options and/or partnerships (e.g., ISCN) are explored

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. This is a good idea, also suggested by Reviewer #1. We've updated the website, as requested (https://lter.github.io/som-website/contribute_data.html) and included more information (see text above) clarifying what is included in Layer data.

Figure 1: Can DIRT and nutnet also be touching the green circle because they are manipulations? This is a good suggestion we can modify in revisions

Figure 2: There are two locations shown here. Do they each get their own Location tabs? Yes, we've clarified this in the caption: "The right side of the figure illustrates data from two hypothetical locations (e.g., a LTER and CZO site, respectively) where Location 1 includes data from two profiles that each have information from one layer. Location 2 provides data from one profile that has information from three layers. Any location may provide data from multiple profiles or layers. With data harmonization data for each profile and layer will inherit metadata and location data that are provided in the location tab." Figure 5: Can the depth axis have units or at least put the units in the caption? This has been done

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. Yes, a copy of the primary data for harmonization are referenced in a google drive folder, which is not a permanent repository. That's why we note in 2.2: "These primary data may or may not be in a published state but, if not published, would be equivalent to data provided for publication. Many of the datasets in SoDaH were already published in public repositories like EDI, the repository for LTER data, or available through the NEON data portal. Other datasets that we wanted to include in SoDaH, however, had not been published or were difficult to find or identify (mainly data from CZO sites and the DIRT network, but also some LTER data). Publishing these primary data remains an active priority for our working group."

From a reproducibility standpoint, we probably should be storing the completed metatemplates in Zenodo, or add them to the current database repository in EDI. We wonder, however, how much value would be gained from such an effort?

And in section 3.4: "We ask that new contributions of primary data that are harmonized into SoDaH be published with a unique DOI".

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? The authorship policy was mainly for our working group as we developed SoDaH. Now that the dataset is published "We encourage users of SoDaH data to cite both this publication and the dataset citation provided by the EDI data portal in their products"

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. The "data summary" tab has data " by site" that includes 'notes.pdf' information. These were used by the database managers to check the data harmonization process. Now that the database is published, this information isn't really needed and we have removed these links from the Shiny app.

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?

Yes, "the data table on the Query page of the SoDaH Shiny application is responsive to the filter options at the top of the Query page. When users click the "Download data" button next to the table, the downloaded .csv file will contain the same data shown in the application table at that time. Code examples for working with the database, including how to filter by specific column values, are provided in the GitHub repository (https://github.com/lter/lterwg-som/data-processing/Tarball_v2 scripts)."

Please also note the supplement to this comment:
https://essd.copernicus.org/preprints/essd-2020-195/essd-2020-195-AC2-supplement.pdf

---

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