Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC2
https://doi.org/10.5194/bg-2021-323-RC2, 2022

Comment on bg-2021-323

William Wieder (Referee)

Referee comment on "Reviews and syntheses: The promise of big diverse soil data, moving current practices towards future potential" by Katherine E. O. Todd-Brown et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-323-RC2, 2022

This is solid, carefully written, succinct perspective on the landscape of soil data and harmonization efforts. I commend the authors for putting together clear, thoughtful state of the science.

This said, I would also encourage the authors to think a bit more broadly about the context in which they are laying out these suggestions. Specifically, with the rapidly evolving soil C sequestration landscape and an infusion of private interests into soil C world (e.g. https://seqana.com/, IndigoAg, etc), how should academics, industry, ngo's and government agencies maintain data access and communication in what's potentially a more crowded, active (and presumably better funded) field? I appreciate this touches on ideas that are broader and somewhat more existential than the soils data challenge the paper more narrowly addresses- but it seems relevant to contextualize the broader landscape of *who* and *why* harmonize soils data- beyond the *how* it can be done better.

My remaining comments are relatively minor, and largely intended to clarify aspects of the text.

I'm not sure I agree with the statement in Line 43-45. Modeled soil properties (here I'm

thinking of hydraulic and thermal properties) rely on pedotransfer functions that use input data of soil physical characteristics (texture and organic matter content). None of these 'soil properties' are used for benchmarking or evaluation, making me wonder what the growing need for more data are really needed for- especially if ILAMB already uses information on soil C stocks and inferred turnover times?

Moreover, data products like SoilGrids already exist, which seems to have a wealth of data that can be used as inputs for or evaluation of Earth system models. Are you suggesting new efforts should go into recreating or augmenting the data processing wheel that informs ISRIC data products (SoilGrids and the Harmonized World Soils Database)? I don't get the sense this is what the authors are envisioning? I also appreciate that "This is just one of many potential uses for harmonized soil data", but I do worry that as written the authors are implying that the harmonized datasets we do have somehow do not reflect FAIR principles.

I really like the tone of the last paragraph of the introduction, which seems constructive and positive.

I also like the preview for what's ahead in in section 2 (lines 73-74) and wonder if the subheadings for section two and headings in Fig 1 should use identical language (acquisition, harmonization, curation, and publication.

Line 79-84, I appreciate the challenge you're trying to articulate- but it kinds of seems like you're suggesting reviewers or journals need better evaluation of data publishing standards. I wonder I this is really where the responsibility should lie, specifically because I don't think as a community, we're well trained in best management of data practices.

I think given better information, data providers would happily provide more useful datasets to repositories, but don't know how. Maybe this is what's implied in line 83 with data providers who 'become frustrated'? I realize you're trying to be brief here- and maybe a solution is articulated in Section 3- but I do worry that the takeaway message from this paragraph is 'currently archived data are incomplete and therefore useless, and we're not really going to tell you how to make them better'.

Line 87, what's a harmonized template?

Line 99 What are TRUST and CARE? If an aim of this manuscript is to broadly educate soil-minded scientists on best data practices describing features of these practices should be briefly articulated (not just referenced).

Line 105. These different transcription / translation methods are nicely described in the text, with examples in Appendix A. Would a table help emphasize similarities and differences of databases listed in Appendix A?

Building on this table idea, I don't think Avni's 2019 paper really provides much depth on these features of the databases. It seems like some A2-A4 all have some high-level similarities- e.g. R code provided on github. A3 and A4 both have a Shiny App (although not listed for A4). Does ISRAD have a shiny app too, or just it's own R library? For what it's worth I feel like some of these back end usability features are helpful if we want people to engage with the harmonized datasets.

Finally, both ISRAD and SODAH were organized with the nested hierarchy established with ISCN. Should this be mentioned? Should ISCN be highlighted in the text (a number of co-

authors have contributed to this effort)? This hierarchical organization of the data is implied, but maybe not explicitly established in the metadata and data models we are or should be using.

Section 2.2, It seems like scripted transcription requires clear dictionaries, vocabulary and metadata to be successful, but based on text in 2.1 this is not common, OR is this just happening in keyed translation?

Section 2.3 is pretty brief Would additional examples be helpful here to illustrate how different efforts have gapfilled or pruned their data? How do these databases expand- which seems important aspect of curation (although discussed in 2.4 for COSORE).

Line 275, I may add something aboveground to this list (as vegetation, land use, productivity and climate are also important for belowground measurements, but rarely co-located with belowground measurements being collected).

Section 3.2 (or in the introduction). Are there successful examples we can learn from elsewhere (e.g. fluxnet, TRY, or FRED) how can these other database models be translated for soils? What unique challenges do the landscape of soils data provide?

I'm 100% behind the suggestions and vision the authors laid out, but I do wonder a bit about to what end? What are the pressing questions that a massive new soils database will let us address? Given the diversity of soil uses, measurements, and communities is a database of databases really what we need? OR, is the soil science community well enough served by individual collections of data that are more focused on more topical areas like radio carbon, respiration fluxes, spectral databases, or bulk C stocks? I realize

this isn’t you’re grant proposal- but presumably it’s heading that way. The text clearly delineates data providers and data aggregators, but who are the data users that will ultimately do something with these datasets once they’re wrangled into something more useful?

Finally, appologies on my delay in posting this review.