Project Details

Name of project: Open source data harmonization: beyond data entry
Project lead and contact details: Kathe Todd-Brown (University of Florida, ktoddbrown@ufl.edu)
Project partners and contact details: William Teng (NASA-ADNET, william.l.teng@nasa.gov) and Rustem A. Albayrak (NASA-ADNET, rustem.a.albayrak@nasa.gov)
Proposed start and end date: March 1, 2020 - September 1, 2020
Budget Requested: $9701
Budget Summary: 1 UF undergraduate student + 1% faculty support ($7701) + travel for student to ESIP 2021 Winter Meeting ($2000)

Project Outline

Project description:

Summary: ***In this project, we will develop, implement, and test best practices for compiling transparent, reproducible, harmonized, and extendable data collections for meta-analysis.*** To do this, we will examine six current meta-analysis efforts in the soil community. Through one-on-one interviews with the PIs and developers as well as a broader community surveys, we will identify the strengths and weaknesses of the approaches used by the individual projects. We will generate a white paper outlining suggested best practices based on our findings. We will then use best practices to prioritize ongoing development of the Soil Organic Carbon Data Rescue and Harmonization (SOC-DRaH), an open community project started by the International Soil Carbon Network (ISCN). The results from this project would provide a solid basis for seeking future funding to benchmark soil carbon dynamics in Earth system models, generate soil maps, and gap fill missing data using machine learning algorithms.

Background: Reanalysis of data from individual projects in a meta-analysis is commonly done by researchers. However, aggregating and preparing data for such an analysis is generally difficult and sometimes unreproducible, because the various data sources are not harmonized. In this context, harmonization means uniform application of a single control vocabularies and data model. In many fields, the main problem is no longer data availability; data can be found in many public online repositories. Nor is the difficulty findability; repositories have expanded their data search capacities through efforts like DataOne, and data accessibility is better now than it has ever been historically. The main problem is harmonizing the various data sources, a problem that urgently needs community guidance.

Similar to other research communities, soil scientists have diverse data sets which would benefit from a more global data context. Several efforts to provide this global context by harmonizing soil data have emerged over recent years, a subset of which is presented in Table 1. These meta-analysis efforts reflect a common increase in complexity, as researchers try to build on other meta-analysis (Crowther 2016 to van Gestel 2018 is one example of this) or broaden data aggregation efforts to expand from a single lab effort (Crowther 2016 and van Gestel 2018) to a mult-lab effort (CC-RCN, ISRaD, LTER-SOM, and ISCN3) requiring more active workflow management practices. We also see two aggregation practices emerging, templated vs scripted data ingestion (See “Method” section). Examining the workflow of these meta-analysis efforts could provide valuable insights into the process and be used to generate best practices recommendations.

Table 1. Selected current meta-analysis efforts to harmonize soil data. Single lab or multi-lab refers to whether the compilation team comprised a single PI or multiple PIs. Templated ingest uses a common template; scripted ingest uses unique templates. Citation number references either a published product or a link to ongoing effort.

Name	Type of soil data	Single lab or multi-lab effort	Templated or scripted ingest	Citation
Crowther 2016	field warming experiments	single	template	1
van Gestel 2018	field warming experiments	single	template	2
CC-RCN	coastal wetland with focus on high resolution core dating	multi	script	3
ISRaD	fraction and radiocarbon isotope	multi	template	4
LTER-SOM	long term ecological studies	multi	script	5
ISCN3	regional surveys	multi	template	6

[1] Crowther, T., Todd-Brown, K., Rowe, C. et al. Quantifying global soil carbon losses in response to warming. Nature 540, 104–108 (2016) doi:10.1038/nature20150. [2] van Gestel, N., Shi, Z., van Groenigen, K. et al. Predicting soil carbon loss with warming. Nature 554, E4–E5 (2018) doi:10.1038/nature25745 [3] Coastal Carbon Research Coordination Network https://serc.si.edu/coastalcarbon [4] Lawrence, C. R., Beem-Miller, J., Hoyt, A. M., et al.: An open-source database for the synthesis of soil radiocarbon data: International Soil Radiocarbon Database (ISRaD) version 1.0, Earth Syst. Sci. Data, 12, 61–76, https://doi.org/10.5194/essd-12-61-2020, 2020. [5] Longterm ecological research soil organic matter group https://github.com/lter/lterwg-som [6] International Soil Carbon Network https://iscn.fluxdata.org/

Method: We will evaluate each of the soil data harmonization efforts listed in Table 1 and implement the resulting best practices to prioritize ongoing development of SOC-DRaH. Dr Todd-Brown has been involved in a number of these efforts to varying degrees and is uniquely positioned to carry out this research. While the exact outline will change, we expect to find the following trends in these projects. To carry out this evaluation, we will (1) review publications within and beyond the soil community, as well as publicly available code and workflows, (2) contact PIs for one-on-one interviews, and (3) conduct a survey of the soil science community to analyse how other meta-analyses are commonly done and solicit more general feedback on how they should be done.

New projects are immediately confronted with the need to develop a control vocabulary and data model. Some groups will expend considerable effort at this stage to include a large diverse group of researchers before aggregating data (ISCN3), while others will extend their data model only as needed (Crowther 2016). *We expect there is a clear need in the community for a common control vocabulary and mechanism to regularly revise and update this vocabulary to reflect current methods.*

Next, projects need an organized workflow to merge and harmonize data sets from different sources into a common data model. Manual entry transcribing the data from source to a common template is expected to be the most common. The templated approach is flexible in the sense that the data source could be encoded in a graph, text from a manuscript, or tabular format and manually transcribed with minimal training. However, once that data is transcribed, revision to the data model often requires manually revisiting the original data source. In addition, the templated approach does not take direct advantage of digitized data tables that are becoming more and more common on online repositories. There are a few projects (LTER-SOM and SOC-DRaH) that instead focus on developing a scripted approach where a piece of code is written uniquely for each data set to convert the data automatically. This scripted approach requires more technical skill but is more robust to changes in the data model, because the original data format is preserved. *We expect that, while a scripted approach is more technically robust, it would make it more difficult to recruit community contributions.*

Finally, projects need to extract insight from these diverse data sets. Larger richer data sets offer more options for analysis. However, the format of the data can also impact usability for final analysis. *We expect that many of these studies rely heavily on Excel for data storage and may not be aware of better alternative formats.*

All code, data products, and white paper generated by this project will be made available via a repository under the ESIP organization GitHub under an MIT or CC-BY license. We anticipate the following deliverables: (1) a general white paper on best practices for data harmonization, supported by (2) survey data and interview transcriptions, and (3) specific recommendations to SOC-DRaH for future feature development to increase usability. Scientifically, we are interested in comparing the control vocabularies and data models developed by each group, specifically, what kind of measurements are common across all projects and is there any consensus on vocabulary.

Project objectives, significance, and impact:

The project objective is to ***develop, implement, and test best practices for compiling transparent, reproducible, harmonized, and extendable data collections for meta-analysis.*** This project is an example of modernization of Earth science workflows using community-recommended best practices and extension of open source software critical to collecting, distributing, or analyzing Earth science data. We will also set guidance for and development of one of the largest soil carbon databases, SOC-DRaH.

Description of key project steps and timeline:

	March	April	May	June	July	Aug.	Sept.	Oct.
Review existing projects	**X**	**X**						
Survey the community and conduct one-on-one interviews		**X**	**X**	**X**				
Compile survey results and transcriptions				**X**	**X**	**X**		
Whitepaper and code preparation					**X**	**X**	**X**	**X**

Description of additional funding currently supporting this work: Todd-Brown start-up funding (additional student researcher time)

Outreach

What groups/audiences will be engaged in the project?

International Soil Carbon Network and ESIP (Agriculture and Climate Cluster and other relevant collaboration groups) as well as all meta-analysis groups listed in Table 1 will be invited to engage in this project.

How will you judge the project's impact?

Level of engagement from the broader community in the evaluation process, specifically participation in one-on-one interviews and in the community surveys.

How will you share the knowledge generated by the project?

Code, survey data, and whitepapers generated by this project will be released on GitHub under the ESIP organization. We will also present these results at the ESIP Winter meeting.

Description of *who* (agencies/individuals) should be aware of this project, i.e. potential outreach targets:

Soils and biogeochemical communities. Specific agencies include FACT-NIFA-USDA, Macrosystems-NSF, Geoinformatics-NSF, TES-DOE, EESM-DOE. ROSES-NASA, CUAHSI (https://data.cuahsi.org/)

Project Partners

Description of project partners (agencies/individuals) and their involvement:

Kathe Todd-Brown (University of Florida) - supervise student, lead interviews and surveys
William Teng (NASA-ADNET) - engagement with the ESIP-Agriculture and Climate Cluster, advise on data format and storage recommendations
Rustem A. Albayrak (NASA-ADNET) - advise on data format reusability recommendations with targeted use for machine learning analysis

How will this project engage members of the ESIP community: A soils data product will be of broad interest to the Agriculture and Climate Cluster, and the team will provide regular updates to this group and recruit interested parties for further involvement. We will also present results at the ESIP 2021 winter meeting.