Reviews and syntheses: The promise of big soil data, moving current practices towards future potential

Katherine E. O. Todd-Brown[1], Rose Z. Abramoff[2], Jeffrey Beem-Miller[3], Hava K. Blair[4], Stevan Earl[5], Kristen J. Frederick[1], Daniel R. Fuka[6], Mario Guevara Santamaria[7], Jennifer W. Harden[8], Katherine Heckman[9], Lillian J. Haren[1], James R. Holmquist[10], Alison M. Hoyt[8], David H. Klinges[11], David S. LeBauer[12], Avni Malhotra[8, 13], Shelby C. McClelland[14], Lucas E. Nave[15], Katherine S. Rocci[16], Sean M. Schaeffer[17], Shane Stoner[3, 18], Natasja van Gestel[19], Sophie F. von Fromm[3, 18], and Marisa L. Younger[1]

[1]Department of Environmental Engineering Science, University of Florida, Gainesville, Florida, USA
[2]Laboratoire des Sciences du Climat et de l'Environnement, Gif-sur-Yvette, France
[3]Max Planck Institute for Biogeochemistry, Jena, Germany
[4]Department of Soil, Water, and Climate, University of Minnesota, St Paul, MN, USA
[5]Global Institute of Sustainability and Innovation, Arizona State University, Tempe, AZ, USA
[6]Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[7]Centro de Geociencias, Universidad Nacional Autónoma de México, Juriquilla, Querétaro, Mexico
[8]Department of Earth System Science, Stanford University, Stanford, CA, USA
[9]Northern Research Station, USDA Forest Service, Houghton, MI, USA
[10]Smithsonian Environmental Research Center, Edgewater, Maryland, USA
[11]School of Natural Resources and Environment, University of Florida, Gainesville, Florida, USA
[12]Arizona Experiment Station, College of Agriculture and Life Sciences, University of Arizona, Tuscon, AZ, USA
[13]Department of Geography, University of Zürich, Zürich, Switzerland
[14]Department of Soil and Crop Sciences, Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO, USA
[15]Biological Station and Dept. of Ecology and Evolutionary Biology, University of Michigan, Pellston, MI, USA
[16]Natural Resource Ecology Laboratory, Department of Soil and Crop Sciences, Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO, USA
[17]Biosystems Engineering and Soil Science Department, University of Tennessee, Knoxville, TN, USA
[18]ETH Zürich, Zürich, Switzerland
[19]Department of Biological Sciences & TTU Climate Center, Texas Tech University, Lubbock, Texas, USA

Correspondence: K Todd-Brown (ktoddbrown@ufl.edu)

Abstract.

  In the age of big data, soil data are more available than ever, but -outside of a few large soil survey resources- remain largely unusable for informing soil management and understanding Earth system processes outside of the original study. Data science has promised a fully reusable research pipeline where data from past studies are used to contextualize new findings and reanalyzed for global relevance. Yet synthesis projects encounter challenges at all steps of the data reuse pipeline, including unavailable data, labor-intensive transcription of datasets, incomplete metadata, and a lack of communication between collaborators. Here, using insights from a diversity of soil, data and climate scientists, we summarize current practices in soil data synthesis across all stages of database creation: data discovery, input, harmonization, curation, and publication. We then sug-

gest new soil-focused semantic tools to improve existing data pipelines, such as ontologies, vocabulary lists, and community
10 practices. Our goal is to provide the soil data community with an overview of current practices in soil data and where we need
to go to fully leverage big data to solve soil problems in the next century.

1 Introduction

Soils host a myriad of life forms, from viruses to macro fauna, that govern important ecosystem functions, such as nutrient
cycling, water retention, pollutant remediation and carbon sequestration. Soils play a critical role in human existence, both as
15 a habitat and as a source of food, fuel and fiber. Additionally, soils contain the largest pool of actively cycling carbon (Ciais
et al., 2013) and can contribute to climate change mitigation through soil carbon sequestration management practices (Smith
et al., 2020; Fuss et al., 2018). The health of this critical resource is under threat from mismanaged land-use changes, leading
to reductions in soil organic matter and changes in soil structure that have led to wind and water erosion (Lehmann et al.,
2020). Because of the soil's diverse role in food security, climate change, society, and ecosystem functioning, soils have been
20 the central focus of many studies resulting in a wealth of data.

Soil data are as diverse as soils themselves, reflecting interactions of biological, chemical, hydrological and biophysical
processes. Soil data uses include a broad range of applications such as ecology, biogeochemistry (Iversen et al., 2017; Wieder
et al., 2021), soil engineering, soil taxonomy and classification, geochemistry (Nave et al., 2016; Hengl et al., 2017; Lawrence
et al., 2020), micrometeorology (Cheah et al., 2018), agronomy (Lyons et al., 2020), and geomorphology (see Todd-Brown
25 (2021) for detailed examples of soil databases). Datasets, defined as "a collection of scientific data including primary data and
metadata organized and formatted for a particular purpose" (Löffler et al., 2021) are assembled by an equally diverse range
of organizations. These organizations include government agencies, academic collaborations, nongovernmental organizations,
and industry, reflecting a wide range of generators and users including farmers, land managers, students, technicians, scientists,
and policy makers. Recent efforts have summarized the availability of soil data, database types, their strengths and weaknesses,
30 and how research and data networks can help solve key soil-related research and societal problems (Harden et al., 2018;
Malhotra et al., 2019; Weintraub et al., 2019). These soil databases are a key tool for a wide variety of users, such as the
scientific community, land managers and policy makers. Here we discuss common principles and practices around soil data
harmonization, as well as associated challenges and opportunities.

The creation of multi-sourced data products requires unification of data, through the process of harmonization. "Database",
35 for the purposes of this paper, is defined as a synthetic collection of soil observations, including location, physicochemical
properties, units, and methods. Databases consist of data spanning multiple curation levels - from raw, to processed, to gap-
filled - and are often assembled from different sources including collaborating investigators' datasets, disparate intermediate
datasets, or large existing databases. Further, these databases are often cross-referenced, for example, by integrating georefer-
enced meteorological data or remote sensing products like net primary production maps. "Harmonization" of data refers to the
40 integration process and can include activities like restructuring data tables. Harmonized databases can be a powerful validation
tool for Earth System models. As one example, the International Land Model Benchmarking Project was created as a system-

atic means of Earth System Model evaluation based on observational data, including soil and remote sensing data, which in turn informs model development (Collier et al., 2018). There is a growing need for evaluation data for modeled soil properties as soil components of Earth system models become more complex. This is just one of many potential uses for harmonized soil data, thus there is need for best practices and guidance in creating them.

The FAIR (Findable, Accessible, Interoperable, and Reusable) Data Principles have become a popular shorthand for best practices in scientific data management and stewardship across scientific domains (Wilkinson et al., 2016). Much of the previous focus on FAIR principles has focused on data access (though some work has focused elsewhere (Goble et al., 2020), which can be difficult if it relies on the standard 'available on request' note included at the end of publications (Savage and Vickers, 2009; Vines et al., 2014). While data access has improved due to funding agencies requiring data management plans that include public archiving, more widely available data does not ensure that they are interoperable and reusable. Many of the challenges and best practices we highlight below address this latter half of FAIR data - interoperability and reusability - and are reflective of the diversity of soil data generators, users, and stakeholders. In this work, we hope to guide prospective soil data providers, users, and synthesizers by highlighting both challenges and examples of working toward FAIR data within the soil science community.

While it is not desirable nor tenable to force all soil data to adhere to a single data template, there are shared strategies, semantic tools, community practices, and protocols for data harmonization and integration efforts that would increase transparency and decrease initialization effort of these projects. We set out in this paper to summarize current practices in soil data management and database harmonization, as well as suggest new semantic and community tools for the future of soil harmonization. We offer this summary from the collective viewpoint of over two dozen researchers across North America and Europe, working on a broad array of funded and unfunded projects to construct independent harmonized soil databases. With our perspective and definitions for context, our specific objectives were to define common and aspirational practices in harmonized database development, and envision next steps to improve reproducibility and transparency of harmonized database development for soils.

2 Current database pipeline

A database pipeline is a series of steps that brings data from an initial format to a harmonized database, with steps including: generation, discovery, acquisition, input and harmonization, curation, gap-filling and pruning, and publication of the database (Figure 1). While these general steps are fairly easy to articulate, their specific implementation varies and is often time consuming. Discovery, access, and "wrangling" can exceed 80% of the effort required for a new scientific discovery (Lohr, 2014; Beno et al., 2017). The extension of this pipeline to create a generalizable solution adds an additional level of difficulty (Furche et al., 2016), but can be essential to ensure that both the data and workflows are FAIR (Goble et al., 2020). For brevity, we assume that datasets of interest have already been identified. However, searchability of data repositories is an active area of research (Pampel et al., 2013; Löffler et al., 2021). We briefly review several common approaches to data acquisition, harmonization, curation, and publication.
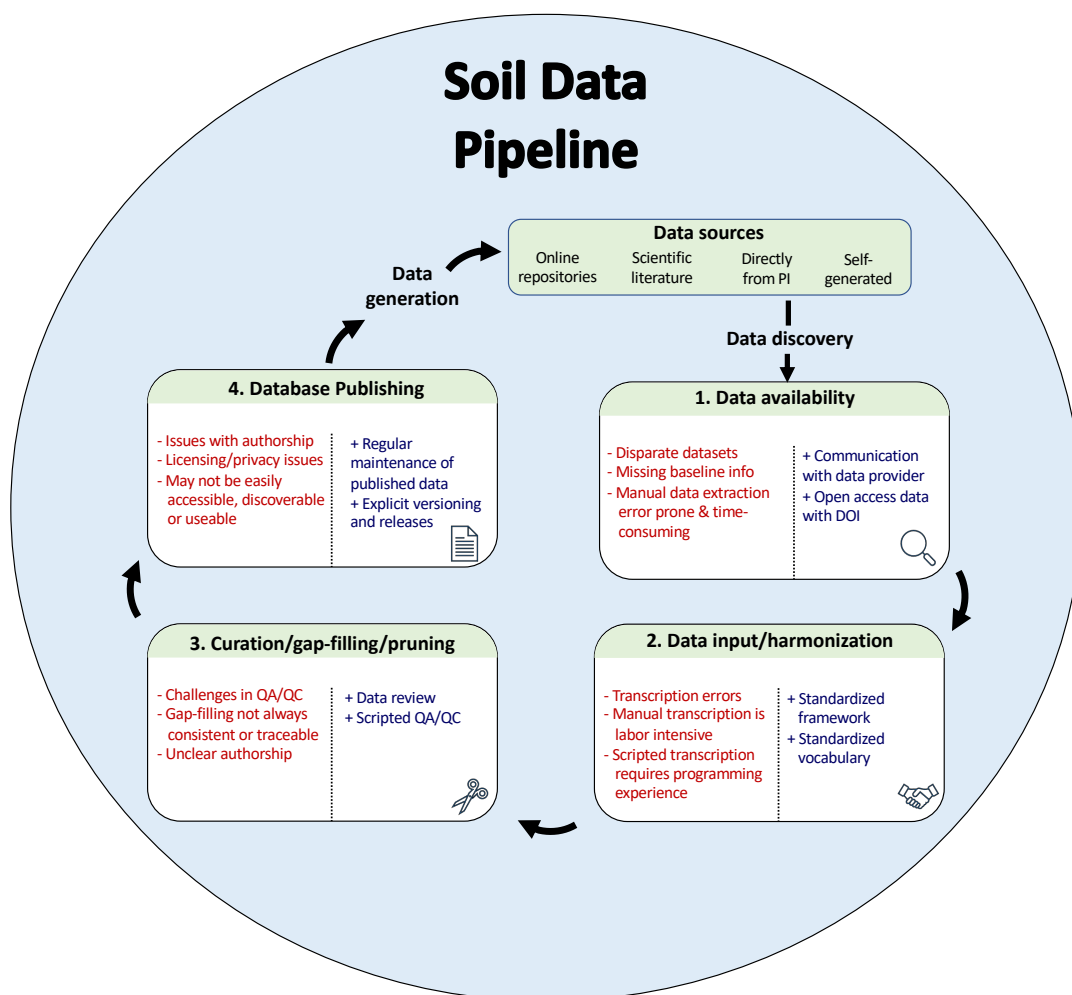
Figure 1. Database pipeline with pain points (indicated by "-") and suggestions for improvement (indicated by "+") to conform more closely to FAIR Data Principles. Data sources can be diverse, including published (online repositories and scientific literature) and unpublished sources (direct from principal investigator; PI). After these sources have been discovered, the data must be accessed, harmonized according to a standard format or data model (internal to the project or a community-driven standard). The aggregated data must then be curated before ultimately being published for reuse.

2.1 Availability

75

Reproducible analysis is fundamental to robust science and data analysis. Naively, a newcomer to the field of 21st century science might be forgiven the assumption that a published peer-reviewed journal article would, by default, also be accompanied by a published dataset in a machine readable format. In the authors' experiences, this is uncommon for a number of reasons.

While some peer-reviewed journals and funding organizations require the data to be deposited in a trusted repository that

80 supports the FAIR principles (Fox et al., 2021), confirming that data meet these high standards is often overlooked during the review process. Indeed, there is often confusion in the field as to what exactly such 'high standards' are. Key contextual data for one study may be mostly irrelevant for a second. Anticipating these contextual data needs is challenging and leads many data providers who would otherwise support data sharing to become frustrated with the existing guidance (Couture et al., 2018).

On the data aggregation side, many data aggregators are challenged by unclear data documentation and metadata. This leads

85 to interactions between data providers and data aggregators that can vary from no contact (e.g. left the field due to career changes, retirement, death, or are unwilling to interact) to high-contact (data providers collaborate with data aggregators to fill out a harmonized template). Intermediates along this gradient could include reaching out to data providers to confirm variable ranges, addressing possible errors, requesting specific unpublished measurements, or clarification of ambiguous descriptions. Data providers have unique knowledge about their systems and can be instrumental in expanding or modifying the scope of

90 the resulting database analysis. Data-centered collaborations can lead to new communities of practice and better science (see Future recommendations section).

In addition to these benefits there are also tradeoffs. Acknowledgement and level of visibility of original data contributions remain an open question. Data providers may expect to be listed as co-authors upon reuse of their data, in recognition of their past effort collecting the data, despite having limited or no engagement in the reuse project. This is often frustrating to a data

95 aggregator who, in turn, is left with an ever expanding list of coauthors with varying levels of involvement (see Publication section). This can lead to conflict and a lack of trust in the community (Longo and Drazen, 2016).

In general however, we feel that direct collaboration between data providers and data aggregators is a critical relationship to nurture. As the community continues to converge on shared tenets of good data governance (FAIR - (Wilkinson et al., 2016) TRUST - (Lin et al., 2020) and CARE - (Carroll et al., 2020)) it is becoming increasingly clear that 'just put it on a repository'

100 is only the beginning.

2.2 Data input and harmonization

Datasets typically reflect the purposes of the original study; however examining those same data in a broader context often requires a different data format. Harmonizing one data contribution with a broader collection entails merging or breaking apart data tables, renaming columns, and occasionally converting units of observation (see Section 2.3 below). Translation of the

105 data table structure (often called the data model) is typically handled one of three ways: manual template transcription, scripted template transcription, or keyed translation. All three methods have justifications for use, and data aggregators may choose

to turn from one method to another as a database grows. Regardless of which method is used, one of the primary goals is to maintain data provenance that allows each data point to be traced to an original study or author.

Manual transcription is the most common method, and typically entails data taken from the original source and entered
110 into a common template by either the data provider or aggregator (see Sections A1 and A2). Asking the data provider to fill out these data templates is often identified as a major hurdle to contribution, yet data aggregators may be unfamiliar with the data provided and thus capture an incomplete or incorrect translation of the original data into the new format without the help of the data provider. Regardless of who fills out the template, human transcription of data is error prone. In some cases this is unavoidable when the data are not available in a machine-readable format. There are a number of software that allow for
115 data extraction from figures (i.e. Web Plot Digitizer (Rohatgi, 2021), Data Thief (Tummers, 2006), and metaDigitise (Pick et al., 2018)) and tables (pdftools (Jeroen Ooms, 2021), tabula (Aristarán et al., 2012 - 2020)), that can reduce human error in transcribing these machine-hostile formats. Despite its flaws, manual transcription is flexible and easy to set up, making it a frequent choice for data aggregation studies with a tight timeline.

An alternative approach to manual transcription is scripted transcription. Scripted template transcription involves writing a
120 computer program, customized to the specific data being ingested, to reformat the data tables and column names to match a target data standard or template (see Section A3). This approach requires familiarity with both soil science (to understand the measurements) and programming (to write the scripts). In practice, the authors find such a skill combination unusual for any individual researcher, necessitating the use of interdisciplinary teams and adding organizational complexity. The codebase can also become unwieldy if written on a case-by-case basis for each input dataset. These costs are countered by an increase in
125 accuracy, transparency, and reproducibility when compared with manual transcription.

Keyed translation is the most general approach and, as a result, requires the most extensive informatics work. Keyed translation is related to scripted transcription but uses a dictionary to define relationships between the input data format and target data format; for example mapping Column 1 to Column A or Table A to Table B. Keyed translation combines metadata about each dataset with a generalized conversion script to generate a harmonized database (see A4). Such a generalized approach can
130 be more easily extended to expand the number of data sources. However there is currently no broadly agreed upon annotation vocabulary making it necessary to annotate each dataset individually within each project. In addition, the computational expertise needed for this approach is the highest of the three outlined here. While we feel this holds great promise for future studies, this is an uncommon approach due to these challenges.

2.3 Curation

135 Data that are in an integrated database often still need to be curated to ensure accuracy, convert units, address missing data or gaps, and reduce or aggregate data to derive relevant data products. While scripts are often used extensively at this phase, expert interpretation and review is a critical component. Finally, reuse of databases often requires a repeat of this curation phase – what may be appropriate for one question or purpose may not be appropriate for another.

Scripting is heavily utilized to augment expert review to quality control data. These scripts both automate and document
140 quality control criteria, however setting those criteria often requires an extensive knowledge of the system and measurement

methods. For example, the ISRaD database has an automated quality control protocol accessible via web interface which ensures that values are within reasonable ranges, and checks that records and critical metadata are appropriately linked across database tables (Lawrence et al., 2020). Following this initial filter, a manual individual 'expert review' is conducted by a trained ISRaD volunteer (see A2). These extensive quality control procedures require time and diverse expertise, making them

145 unattractive for many open source database projects without broad recognition of service by the field.

Gap-filling expands data coverage, and can include a number of strategies to fill in missing data at both the layer or horizon level as well as the profile or site level. Strategies include linear interpolation, pedotransfer functions, georeferenced data extraction, or more sophisticated machine learning algorithms. Given the wide variability in gap-filling practices and objectives, these methods must be extensively documented, clearly state use case restrictions, and estimate uncertainties.

150 Finally, a common step is to remove unnecessary data from the data product via pruning. Pruning removes samples based on location or type of measurement after database compilation, typically to cater to specific data user needs, and reduces the size of the data product. Both pruning and gap-filling highlights the importance of maintaining an intermediary harmonized database as well as the final data product both to preserve the original contextual data and reuse of that data for alternative projects.

155 **2.4 Publication, credit and attributions**

Authorship issues are common in data aggregation projects due to unclear expectations and conflicting conventions across what are often large teams of collaborators. Indeed these problems are common in large collaborative projects (Cooke and Hilton, 2015). This can be mitigated with an expanded co-author list to include both data providers, aggregators, and re-analysis teams, but requires significant project management and organizational overhead. As always with larger team science,

160 we highly recommend a formal authorship policy prior to beginning database compilation including what the role of each contributor is and who are in lead manuscript positions, listed as co-authors, and listed in acknowledgements. While it can be tempting in a data aggregation project to fall back on what you are legally entitled to (i.e. if the data are there, use them), we strongly feel that collaborative projects build trust within the scientific community, leading to better data interpretation and seeding future collaborations.

165 Related to issues of credit and co-authorship described above are issues with data licenses. Similar to manuscripts, data are often released under a specific legal license with requested reuse considerations, which may hinder the inclusion of otherwise 'public' data in a data synthesis. There is a tension in choosing between an adequately restrictive license, which can help ensure that a specific project and data providers are given credit, and a permissive license, which can increase data reuse. The Creative Commons provides a framework to examine these considerations but there are many other standard and custom licenses. The

170 most permissive is CC-0 or public domain license that puts no restriction on data use. A 'By-Acknowledgment' (BY) rider requests that the original data source be acknowledged in the derivative product in some way (sometimes this acknowledgement is specified, sometimes not). A 'Non-Commercial Use' clause restricts the sale of the data for commercial purposes. Finally a 'Share-Alike' or copy-left clause says that the data may be reused if it is released under the same license. A 'CC-BY' license

is probably the closest to the traditional academic practice of research citation and many scientific repositories including the
175 Environmental Data Initiative and Pangaea encourage data providers to select this option.

In all cases database creation does not have to be a single push, but is ideally part of an ongoing synthesis effort, leading to
the need for database versioning. The COSORE database is an example for such an approach (Bond-Lamberty et al., 2020).
After each major change (release) the database receives a new doi and is permanently archived on a repository. This allows
maximal transparency, allowing data users to reproduce an analysis from a given version and making it easy to find the newest
180 version of the database.

3   Future of soil data

The hope of big data is to have any data collected at any time anywhere in the world at the tip of your fingers (see Section
B). For soil science the potential for long-term (multi-decadal) understanding is particularly exciting. Long-temporal coverage
of soil data could lead to a better understanding of soil carbon sequestration potential to mitigate climate change, or better
185 management of soils for crops. How do we attain these above futures, where data reuse is equally as valued as data production?

We recommend implementing a core set of measurements and processes to facilitate soil data reuse. The recommendations
in section 3.1 are aimed at researchers collecting soil data who wish to ensure the long-term value and reusability of their
datasets. These recommendations are also relevant for journals and peer reviewers of soil science research as a short checklist
of key details that should be reported or addressed. Section 3.2 outlines recommendations for researchers who wish to partic-
190 ipate in the data harmonization process. These recommendations encompass both technical and social considerations for data
harmonization efforts and focus on what can be done right now to further soil data exchange.

3.1   What to measure and report?

Soils are inherently rooted in time and space, making high resolution spatial and temporal information (including sampling
date, latitude, longitude, and geographic datum, and depth of sample) critical for building context and data reuse. Data providers
195 will often ask 'what should I measure' to be relevant to data aggregation efforts, and there are efforts to provide such guid-
ance (Billings et al., 2021). We have chosen instead to focus on critical temporal-location information to allow data to be
expanded, contextualized, and annotated. The issue is not that researchers do not know how to record this information, but
rather conflicting objectives may prevent its recording.

Geospatial metadata may present a privacy concern, for example when the soil measurements are tied to the economic
200 valuation of the land as in agricultural systems. For data collected on privately owned land, such as on-farm research and
observations, researchers may not be at liberty to release detailed location information publicly in order to protect landowner
privacy (Richardson et al., 2015). There are efforts to bridge data sharing and data privacy. For example, the platform under
development by the International Agroinformatics Alliance will integrate secure data storage, granular data permissions, and
options to register privately hosted data to facilitate data discovery and sharing while protecting privacy (Gustafson et al.,
205 2017). Clearly this is an ongoing discussion that will require more research and conversations with stakeholders.

The advantages of high precision geolocations are significant; and regardless of the precision, the level of uncertainty in the provided geolocation is critical and often missing in archived datasets. Location information enables soil data to be joined with the growing number of gridded global datasets that can provide key contextual information for interpretation and modeling. While there are privacy concerns in some locations, not reporting the location of a sample collection should be the exception

210   and not the rule, especially in publicly funded research data.

In addition to location, sample depth is also critical due to the variation of soil properties and processes with depth. Unfortunately over 60% of studies fail to report measurements of sample depth (i.e. layer defining upper and lower bounds) associated with soil data (Yost and Hartemink, 2020). This is particularly critical to advancing our understanding of deeper soil properties and functions, but also relevant for the effects of surface tillage and grazing on managed lands.

215   Soils are temporally dynamic and time of collection can provide key insights into decadal level changes in soils. Soils change over time owing to pedogenesis, historical land use and increasingly, global climate change (Tugel et al., 2005; Richter et al., 2011; Ellis, 2011; Harden et al., 2018). Recording the time of collection for modern datasets can produce valuable returns in future reanalysis. Older datasets, consisting of historic measurements and archived samples, are increasingly valuable to track soil responses to global change. Such datasets can provide a window into the dynamics of how soil properties change, and

220   should be a high priority for data rescue and documentation.

Decadal scale soil records provide valuable information for the study of global change and land management, and therefore sites associated with older observations should be prioritized for re-survey efforts (Hawkins et al., 2013). Decadal scale datasets have their own challenges. These data are often associated with a particular researcher or group that are typically released at the retirement of the primary investigator, and likely have been reformatted multiple times across several generations of storage

225   systems and lab staff and may even be on analog storage and not digitized. Thus, these re-survey efforts can be combined with data rescue to be an incredibly valuable data product involving digitization of paper copies, or conducting structured interviews of personnel to enrich metadata of prior observations (Karasti et al., 2006). In the opinion of the authors, data rescue efforts are an underutilized resource in the field.

By adopting these recommendations to record geolocation, depth of sample, and collection date we can greatly increase the

230   value of soil data, extending the measurement reusability for future analysis.

3.2   How to harmonize?

We touched on several common approaches to data harmonization in this paper. Often driven by a single research question or objective, data harmonization has historically been a laborious process carried out by a single or small group of researchers for a specific project. Based on our experiences in various harmonization projects, we propose a more community-centered approach

235   moving forward, founded on the principles of open and transparent science. Outputs from these groups should include semantic tools like ontologies and shared vocabulary lists with clear and transparent governance, as well as a new community-centered approach to the practice of data harmonization and the resulting databases.

3.2.1 Community tools

There is an understandable tendency by many scientists towards standards. If all soil data adhered to a common template, or
240 data model, with uniform tables and column names, then it would be trivial to append the data from one study with the data
from a second. Unfortunately, due to the diversity of soil types and methodologies, as well as ever evolving measurement
technologies, we feel that this is impractical for soil research data, although several valiant efforts are underway to do this
(Nave et al., 2016; Lawrence et al., 2020). In practice, researchers will continue to develop their own data tables and internal
conventions that make sense for their experimental structure, location, and measurement type. However, semantic tools and
245 standards are still useful (Onerhime, 2021).

Annotating datasets with a common vocabulary forms the theoretical backbone of all data harmonization work. Whether this
is a manual copy-paste from a source data table to a common data template, or creating a thesaurus that cross-references given
data columns to some internal standard name, both processes rely on a vocabulary. Classical soil glossaries and lab manuals
have been printed as dictionary style references that are difficult to transform into digital resources for both copyright reasons
250 and technical ones. This vocabulary could be a valuable community resource but would require ongoing engagement with the
research community to remain accessible, relevant, and up to date. Further extending this vocabulary into an ontology, that
captures the relationship between terms in addition to their definitions, could drive the next generation of data-driven machine
learning. Community developed ontologies and vocabulary lists like ENVO (Buttigieg et al., 2016), CSDMS (CSDMS, 2019),
GLOSIS (Palma et al., 2020), and CF (Hassell et al., 2017) could provide reusable resources that are currently missing and
255 underutilized in the soil community. The soil community as a whole needs to engage with these broader resources to ensure
the informatics reflects new developments in the understanding of soil science and measurements being made.

3.2.2 Community practice

Before data can talk, communities need to talk. Based on the experience of the authors, developing, adopting, and maintaining
semantic resources is beyond the scope of any one lab or organization and requires community. In the development phase, a
260 diverse community can ensure that the broadest possible needs are being addressed. Adoption is more likely if the resource
addresses the needs of the community and that community has ownership over the resource. Finally, maintaining semantic
resources requires ongoing updates and revisions as methods shift. All of this requires a new type of community, one centered
on data and tools to support the interoperability of that data.

Successful data centered communities are open, transparent, diverse, and rewarding (Cooke and Hilton, 2015). They are
265 open, in the sense that anyone can join or contribute and are empowered through educational activities to participate. Trans-
parency ensures it is easy to contribute and to understand decision making processes. Diverse communities can draw on a
wide range of skill sets, from experience in soil processes to knowledge representation. They are also rewarding, furthering
members' careers through creation of tangible products (for example citations or grant dollars) and opportunities for scientific
leadership and service. While there are several approaches to achieve this, one possible workflow for the establishment of a
270 new harmonized database might look something like the following:

1. Open application period and selection of team

2. Purpose setting and establishing a common vocabulary

3. Curation and use of data product

A mix of backgrounds in software development, knowledge engineering (the design of ontologies and knowledge graphs), classical soil pedology, and specific measurement methodology (e.g. microbial characterizations, physical chemistry, and vadose zone hydrology) is required for a data harmonization project. It is often challenging to recruit such an interdisciplinary team that requires such a high level of diversity beyond traditional domain boundaries. A well-advertised open application period is a critical first step, and may be supplemented with targeted solicitations. Teams representing a range of perspectives, backgrounds, and career stages are better able to envision a wider range of end uses (and users) of soils data. These efforts are likely to result in better outcomes- improving research products and advancing the careers of people previously excluded from existing synthesis efforts, power structures, or established communities.

One of the first challenges with an interdisciplinary team is establishing agreement on goals and methods. This requires developing a shared understanding and vocabulary (e.g. through educational activities on computational tools or soils surveys and measurements). In an academic community, shared purpose is most easily motivated by a synthesis paper or research question. Data may provide a clear shared motivation but it's uses and governance processes need to be clearly identified and revisited regularly.

Sustainable creation and curation of the harmonized database is essential to create relevant data products to serve a database's specific purpose and enable future reuse to address a variety of questions. Accessing, annotating and merging the datasets is a well-established technical process once the community tools and community of practice are in place. Curation of databases could be patterned after the manuscript review process, where domain researchers review proposed database additions to ensure the accuracy of new contributions. This review process should keep the diverse needs and practices of the soil community in mind, including soil surveyors, field/lab experimentalists, and land managers. In the end, the synthesis of existing data is not the goal: it is their application to scientific problems. In that regard, successful product development from a database can encourage growth and adoption of the data resource by others.

4   Conclusions

Soils are the foundation for our food and fiber system as well as a significant component of the global carbon cycle. As such, information and measurements of the soil system, from hydrological conductivity to soil carbon stocks to changes in nutrient content, are a key public good for a varied group of users. However, this valuable scientific resource is currently under-utilized due to many of the issues outlined above. We suggest data use and reuse could be facilitated by addressing issues along the database construction pipeline.

We outlined database creation as a common set of steps: generation, discovery, access, curation, and publication. While this pipeline can look different depending on the skill sets, timeline, and funding structure of the researchers involved, we

summarized common pain points throughout this process, which can reduce the accuracy and usability of a database. Data collection, synthesis, and use are inherently human endeavors, and as such, breaks in this pipeline are often driven by lack of
305 community awareness and practices.

We put forth recommendations ranging from measurement prioritization to data harmonization decisions that can help move forward community practices around soil data. We recommend that contextual information like geolocation, depth of sample, observation time, and management history all be reported with soil measurements. Soil data harmonization requires the development of new semantic tools like vocabulary lists and ontologies that are co-produced by data and soil scientists. Building
310 the capacity to create and maintain these tools require communities of practice including: open application periods to recruit diverse participants, established goals, and clear outcomes. The creation of such communities are not an easy task, but are a needed one.

Ultimately, soil data are an invaluable resource generated and used by a diversity of groups. Given this value, we hope that the work of advancing soil information systems will increasingly be recognized and rewarded as a critical component of the
315 research process. To achieve this, we need not only new tools and practices, but shifts in the broader incentive structure for conducting this kind of work. Our review provides a path forward to enhance community practice around soil data so that we can begin to tackle the vast array of research and management problems, and their solutions, that lay beneath our feet.

Appendix A: Current soil projects

Below are a series of snapshots compiled to represent the range of approaches groups currently take to aggregating soil
320 databases. These four snapshots include a manually compiled database of field warming experiments (Section A1: Crowther et al. (2016)), a database using manual transcription combined with scripted curating of soil radiocarbon measurements (Section A2 - Lawrence et al. (2020)), a manual-scripted combination of coastal soils (Section A3 - Holmquist (2021)), and a keyed translation database of long-term observation (Section A4 - Wieder et al. (2021, 2020)). This is appendix not meant to be exhaustive. For a living list of researcher-driven soil databases please see Todd-Brown (2021).

325 ## A1 Field-warmed soils

The template-driven approach to data harmonization is exemplified by (Crowther et al., 2016). In this study, individual researchers who collected data of interest (in this case soil field-warming manipulations) were contacted directly and invited to collaborate in a meta-analysis. A post-doc was tasked with creating a data template, and working with those collaborators to capture a representation of their study. These data were then appended into an integrated set of data tables and analyzed. By
330 working with researchers directly, this approach captured both published and unpublished data and ensured a nuanced interpretation of the study results. This careful one-on-one approach combined with co-authorship on a high profile journal ensured that researchers were comfortable sharing data that they might have otherwise withheld from a joint publication.

One challenge with this approach is patchy secondary data. Secondary data like climate and soil physical chemical characteristics may not be critical to a small study at a single site but become fundamental to a larger cross-site analysis. Crowther

335 et al. (2016) addressed this by extracting site-level environmental covariates from gridded geospatial files generated from
global modeled predictions covering an array of climate (e.g. WorldClim (Fick and Hijmans, 2017)) and soil physiochemical
characteristics (e.g. SoilGrids (Hengl et al., 2017)). Although these global predictions can be characterized by considerable
uncertainty – especially at the local scale – these global products at least ensure a full set of standardized meta-data from every
single location.

A2 ISRaD

340

The International Soil Radiocarbon Database (ISRaD) consists of an open-source database of soil and soil-related radiocar-
bon data. Additionally, ISRaD provides a continually-developing library of tools for data access and manipulation based in
R (Lawrence et al., 2020). Radiocarbon data production is an expensive and time-consuming process but provides unique in-
formation on longer-timescale in situ processes. In addition, a global pulse in atmospheric radiocarbon content in the 1960's

345 provides unique analytical power for data collected in the intervening decades to constrain models.

Initial funding was provided by the USGS Powell Center and the USDA NIFA FACT program. Currently, the Max Planck
Institute for Biogeochemistry provides both ongoing funding and staffing. Since 2015, over 300 studies have been compiled,
though the data collection process is active and ongoing, and user-submitted data are also welcomed.

The project utilizes a standardized Excel template for data ingestion which each contributor fills out and submits to a

350 designated ISRaD coordinator. The core unit of a data entry is the "profile", which is a unique spatial AND temporal identifier.
All data must be matched to a profile, which is in turn matched to a "site", and the uppermost level of "entry", which identifies
the publication from which the data originates. Such hierarchy is best preserved through vetting with both an automated and
human-led QA/QC process. Therefore, prior to ingestion, data undergo both automated quality check and expert review for
metadata consistency and data quality. All database and data handling tools are built in the open-access R computational

355 language, and an official ISRaD code library is available through the R library repository, CRAN. All code and data are
available in an open Git repository (Beem-Miller et al., 2021). New functions and explanatory vignettes can be submitted
by users for inclusion in the R package. The project website contains information, links, guides, and updates on the project
(ISRaD, 2018-2021).

A3 Coastal Carbon Research Coordination Network

360 The Coastal Carbon Research Coordination Network (CCRCN) was formed to accelerate the pace of discovery in coastal
wetland carbon science by providing the community with access to data, analysis tools, and synthesis opportunities. Funded by
a National Science Foundation Research Coordination Network, the project's primary staff includes a funded research scientist
as well as several part-time data technicians. Besides organizing topical working groups and communities events, one of the
primary engagements of the CCRCN is the development and maintenance of a database of carbon stock and sequestration in

365 coastal marshes, mangroves, swamps, scrub/shrub, and seagrass.

Both the database and its software is hosted on GitHub (Holmquist et al., 2018 - 2021), and its structure and naming
conventions (Holmquist, 2018) were based upon a series of iterative conversations with a range of experts. Integration of new

datasets into the database is via scripted transcription, by which curation to CCRCN standards is performed in a unique "hook" script tailored for each dataset. A suite of helper tools aids unit conversion, quality control, and spatiotemporal processing
370   specific to soil carbon data. Datasets are joined together to construct the multi-level database, partitioned primarily by scale of observation (depth series, core/plot, site, and methods levels). An automatically-generated bibliography tracks primary citations of data contributors alongside the secondary citation of the database itself. Post synthesis QA/QC script identifies possible duplicate plot-level entries between datasets. Internally-facing visuals and reports are generated via Markdown implementation to track database growth as well as geographic/biophysical gaps in the database. Finally, the online version of the database feeds
375   into the backend of its primary public interface, the Coastal Carbon Atlas (Holmquist, 2021). This R Shiny App allows anyone to explore global representation, as well as query desired according to a variety of environmental/methodological parameters, then download the data and (importantly) the corresponding citations.

Parallel to synthesizing the database has been a concerted effort to generate data releases, each with its own DOI, as a service to data submitters as well as the coastal carbon community. This has included dedicated staff time towards outreach, formatting
380   of datasets, generating metadata (based upon the Ecological Metadata Language standards) and assigning DOIs to datasets, which has so far resulted in the public release of 22 datasets so far on the Smithsonian Figshare repository.

A4   SoDaH

The Soils Data Harmonization (SoDaH) and Synthesis project features a tool suite for harmonizing soil organic matter data from disparate sources into a common data model, and a database of harmonized soil organic matter data and related variables
385   that, as of this writing, includes data from over seventy unique studies (Wieder et al., 2020, 2021). The product of a Long-Term Ecological Research (LTER) synthesis working group (LTER Soil Orgnaic matter Working Group, a), the project brought together soil scientists with diverse backgrounds and affiliations with scientific research networks to refine and evaluate theories of soil organic matter dynamics, and to produce a soil organic matter dataset that spans a wide range of environmental and experimental conditions.
390   SoDaH employs a keyed-translation approach using metadata about the data combined with conversion scripts to translate contributed data tables into the common data model. Metadata are organized at multiple levels, including around the study (e.g., study location, data provider) and data variables, which are subdivided into profile, layer, and fraction categories. Additional metadata fields facilitate the identification of experiment details and study design, allowing users to, for example, query data associated only with specific manipulations or control conditions. The harmonization script (LTER Soil Orgnaic matter
395   Working Group, b) maps user-provided metadata and data resulting in new flat file(s) in which the variable names and units, if relevant, are standardized in the output along with appropriate quality control. All output conforms to the specifications of the SoDaH data model thereby enabling the aggregation of output data from disparate studies into a single data file.

Appendix B: Future Dream

It is often difficult to envision how inter-connected soil data could impact science. Here we've taken the liberty of a bit of creative writing to envision what the future of soil science could be in a data connected world.

B1 A new data-savvy world

In the late 21st century, soil data are is effortlessly collected and collated. There are still divisions between groups on what exactly the correct way to fill in certain missing data is, but in general most instruments come with their own connection interfaces that ensure interoperability between sensor data. It took the field scientists a little longer to get behind standardized digital records but most researchers now collaborate with their data libraries and archivists to adhere to established data standards.

Occasionally a researcher will design a completely novel method or experimental treatment that will require additions or modifications to soil semantic tools. These cases are highly sought after by informaticians and extending an existing international standard has been known to launch the careers of young researchers. More often, researchers will review new data archives to vet their annotations and report this as they do in manuscript reviews. Researchers of course complain about this additional workload, but recognize that if they are going to generate data, then there is an obligation to review their colleagues' data. Once or twice in their career a researcher will serve as a domain expert on a relevant ontology board, providing updates and revisions to these key international resources.

Digital data archives are now entirely annotated and the idea of putting data online without reference to an international ontology is considered irresponsible. Data are annotated with one or more ontologies from different sub-domains and there are a range of AI/ML tools that can leverage these annotations to create an integrated database. Model development and meta-analysis studies now spend the bulk of their time honing hypotheses instead of cleaning data.

The great data rescue projects of decades past are sadly, mostly done now. An entire generation of researchers cut their teeth combing through old paper archives and fighting with optical recognition characters. New researchers lament that this highly fruitful line of 'new' research data is now mostly spent. Designing a new sensor processing pipeline just isn't as romantic as speculating on the nature of that old coffee stained field journal.

Contrary to popular belief, the 'traditional' skills of soil observation (hand texturing soils, matching horizon colors) are more in demand than ever. The ongoing climate crisis has now defined several generations of researchers and reignited interest in soils beyond an agricultural context. Knowing your soil and how you impact that soil is as important as water quality, soil science is a fundamental curriculum element for not only foresters and conservation majors but also urban planners and backyard gardeners. New passionate generations of young students have grown up on soil judging competitions and soil reports are common for any land or home purchase.

B2 The era of big data rescue

In the middle of the 21st century, we are in the heyday of the Big Data Rescue of the 21st century. Driven by the need to understand the impact of legacy management and new data scrubbing technologies, researchers have dived deep into the filing

430 cabinets and paper archives of the past century. In addition to the traditional literature review, new graduate students now do targeted data rescue chapters as part of their dissertation. These data rescue projects have also drawn new researchers from the library, data, and other sciences into soils and is also used as a common undergraduate research project.

Big corporations and governments have taken on the task of parcelling out data to the general public, stripping out sensitive information reduces the utility of the data but is seen as a necessary evil. Ethicists are still debating whether this gives the

435 corporations/governments too much knowledge, and double-blind methods are being developed to obfuscate sensitive data even from the data holders.

Ontologies and other semantic resources are increasingly being adapted and extended by domain scientists. Unfortunately, there are several competing standards reflecting national, domain, and general political divisions in the research community. However there are several clearly identified mature semantic resources that most disciplines agree are pretty good. Data man-

440 agement plans from funders now require identifying semantic resources in addition to the final data archive.

B3 A post-pandemic, better-connected world

Over the next few years, the soils community has fully recognized that our we have a data problem, that is actually a community problem. The collecting and publishing science in isolated labs has become increasingly frustrating to new researchers used to instantaneous web results. The COVID-19 Pandemic forced a rapid shift in how science is done, moving what might have

445 been a few day workshop into a longer slow- burn virtual collaboration over months. This led to a new kind of decentralized project management where most projects are now interconnected to similar researchers in regular virtual seminars and working groups.

This increase in researcher interactions has led to an increase in data interactions. As research interact more online, there was a correlating increase in comparing data from their study with their colleagues results. This led to an informal common

450 vocabulary and data methodologies that increasingly showed up on newly archived data. Some graduate students are starting to dive into data rescue operations and further expanding these vocabularies to include older methodologies. Combining automated optical character recognition of scanned documents with manual corrections, these older data are providing valuable insights into climate change.

Author contributions. All authors contributed to the writing on of this manuscript. Facilitator and organizational lead was K Todd-Brown.

455 *Competing interests.* The authors declare no competing interests.

Acknowledgements. We would like to thank several anonymous survey respondents for their contributions.

This work is based on materials provided by the ESIP Lab with support from the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA) and the United States Geologic Survey (USGS).

James Holmquist was funded by the Coastal Carbon Research Coordination Network (DEB-1655622). Natasja van Gestel was funded by
460 the Department of Energy's Biological and Environmental Research (DOE-BER) program (DE-SC0010632).

We would like to thank Christina Schädel (Center for Ecosystem Science and Society, Northern Arizona University), Tom Crowther (ETH Zurich), Adrien C. Finzi (Boston University), Karis McFarlane (Lawrence Livermore National Laboratory), Corey R. Lawrence (USGS), Chelsea Carey (Point Blue Conservation Science), Gabriel Reuben Smith (Stanford University; ETH Zurich) for their contribution to conversations on this topic.

465 References

Aristarán, M., Tigas, M., Merrill, J. B., and Das, J.: Tabula, Accessed: Nov. 2021, https://github.com/tabulapdf/tabula, 2012 - 2020.

Beem-Miller, J., Monroe, J. G., Hoyt, A. M., Stoner, S., van Fromm, S., Lawrence, C. R., and Sierra, C. A.: ISRaD (vs 1.7.8), Release date: 2021-Jan-07; Accessed: Nov. 2021, https://github.com/International-Soil-Radiocarbon-Database/ISRaD, 2021.

Beno, M., Figl, K., Umbrich, J., and Polleres, A.: Open Data Hopes and Fears: Determining the Barriers of Open Data, in: 2017 Conference
470 for E-Democracy and Open Government (CeDEM), pp. 69–81, https://doi.org/10.1109/CeDEM.2017.22, 2017.

Billings, S. A., Lajtha, K., Malhotra, A., Berhe, A. A., de Graaff, M.-A., Earl, S., Fraterrigo, J., Georgiou, K., Grandy, S., Hobbie, S. E., Moore, J. A. M., Nadelhoffer, K., Pierson, D., Rasmussen, C., Silver, W. L., Sulman, B. N., Weintraub, S., and Wieder, W.: Soil organic carbon is not just for soil scientists: measurement recommendations for diverse practitioners, Ecological Applications, 31, e02 290, https://doi.org/https://doi.org/10.1002/eap.2290, 2021.

475 Bond-Lamberty, B., Christianson, D. S., Malhotra, A., Pennington, S. C., Sihi, D., AghaKouchak, A., Anjileli, H., Altaf Arain, M., Armesto, J. J., Ashraf, S., Ataka, M., Baldocchi, D., Andrew Black, T., Buchmann, N., Carbone, M. S., Chang, S.-C., Crill, P., Curtis, P. S., Davidson, E. A., Desai, A. R., Drake, J. E., El-Madany, T. S., Gavazzi, M., Görres, C.-M., Gough, C. M., Goulden, M., Gregg, J., Gutiérrez del Arroyo, O., He, J.-S., Hirano, T., Hopple, A., Hughes, H., Järveoja, J., Jassal, R., Jian, J., Kan, H., Kaye, J., Kominami, Y., Liang, N., Lipson, D., Macdonald, C. A., Maseyk, K., Mathes, K., Mauritz, M., Mayes, M. A., McNulty, S., Miao, G., Migliavacca, M., Miller, S., Miniat,
480 C. F., Nietz, J. G., Nilsson, M. B., Noormets, A., Norouzi, H., O'Connell, C. S., Osborne, B., Oyonarte, C., Pang, Z., Peichl, M., Pendall, E., Perez-Quezada, J. F., Phillips, C. L., Phillips, R. P., Raich, J. W., Renchon, A. A., Ruehr, N. K., Sánchez-Cañete, E. P., Saunders, M., Savage, K. E., Schrumpf, M., Scott, R. L., Seibt, U., Silver, W. L., Sun, W., Szutu, D., Takagi, K., Takagi, M., Teramoto, M., Tjoelker, M. G., Trumbore, S., Ueyama, M., Vargas, R., Varner, R. K., Verfaillie, J., Vogel, C., Wang, J., Winston, G., Wood, T. E., Wu, J., Wutzler, T., Zeng, J., Zha, T., Zhang, Q., and Zou, J.: COSORE: A community database for continuous soil respiration and other soil-atmosphere
485 greenhouse gas flux data, Global Change Biology, 26, 7268–7283, https://doi.org/https://doi.org/10.1111/gcb.15353, 2020.

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., and Mungall, C. J.: The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation, Journal of Biomedical Semantics, 7, 57, https://doi.org/10.1186/s13326-016-0097-6, 2016.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-
490 Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M.: The CARE Principles for Indigenous Data Governance, Data Science Journal, 19, 43, https://doi.org/10.5334/dsj-2020-043, 2020.

Cheah, Y., Christianson, D., Chu, H., Pastorello, G., O'Brien, F., Ong, Y., Ingen, C., Torn, M., and Agarwal, D.: AmeriFlux BADM: Implementing lessons from 12 years of long-tail data management into next generation earth science systems, in: AGU Fall Meeting Abstracts, vol. 2018, pp. IN34A–03, 2018.

495 Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneny, R. B., Piao, S., and Thornton, P.: Carbon and Other Biogeochemical Cycles, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T., D., Q., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., chap. Carbon and, pp. 465–570, Cambridge University Press, Cambridge, UK, and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.014,
500 2013.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, Journal of Advances in Modeling Earth Systems, 10, 2731–2754, https://doi.org/10.1029/2018MS001354, 2018.

Cooke, N. J. and Hilton, M. L., eds.: Enhancing the Effectiveness of Team Science, The National Academies Press, Washington, DC,
505   https://doi.org/10.17226/19007, 2015.

Couture, J. L. C. J. L. C. J. L., Blake, R. E., McDonald, G., and Ward, C. L.: A funder-imposed data publication requirement seldom inspired data sharing, PLoS ONE, 13, e0199 789, https://doi.org/10.1371/journal.pone.0199789, 2018.

Crowther, T., Todd-Brown, K., Rowe, C., Wieder, W., Carey, J., Machmuller, M., Snoek, B., Fang, S., Zhou, G., Allison, S., Blair, J., Bridgham, S., Burton, A., Carrillo, Y., Reich, P., Clark, J., Classen, A., Dijkstra, F., Elberling, B., Emmett, B., Estiarte, M., Frey, S., Guo,
510   J., Harte, J., Jiang, L., Johnson, B., Kröel-Dulay, G., Larsen, K., Laudon, H., Lavallee, J., Luo, Y., Lupascu, M., Ma, L., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L., Schmidt, I., Sistla, S., Sokol, N., Templer, P., Treseder, K., Welker, J., and Bradford, M.: Quantifying global soil carbon losses in response to warming, Nature, 540, 104–108, https://doi.org/10.1038/nature20150, 2016.

CSDMS: CSN Searchable List — Community Surface Dynamics Modeling System (vs 0.8.3), wiki, https://csdms.colorado.edu/mediawiki/
515   index.php?title=CSN_Searchable_List&oldid=227158, [Online; accessed 6-October-2021], 2019.

Ellis, E.: Anthropogenic transformation of the terrestrial biosphere, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 369, 1010–35, https://doi.org/10.1098/rsta.2010.0331, 2011.

Fick, S. E. and Hijmans, R. J.: WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas, International Journal of Climatology, 37, 4302–4315, https://doi.org/https://doi.org/10.1002/joc.5086, 2017.

520   Fox, P., Erdmann, C., Stall, S., Griffies, S. M., Beal, L. M., Pinardi, N., Hanson, B., Friedrichs, M. A. M., Feakins, S., Bracco, A., Pirenne, B., and Legg, S.: Data and Software Sharing Guidance for Authors Submitting to AGU Journals, https://doi.org/10.5281/zenodo.5124741, We are thankful to Dr. Peter Fox and his initial draft contribution to this guidance document while also saddened by his recent passing. Peter's contributions to the community continue to have a great impact as represented in this document. Thank you, Peter, you are missed., 2021.

525   Furche, T., Gottlob, G., Libkin, L., Orsi, G., and Paton, N.: Data Wrangling for Big Data: Challenges and Opportunities, in: Advances in Database Technology — EDBT 2016, Advances in Database Technology, pp. 473–478, University of Konstanz, https://doi.org/10.5441/002/edbt.2016.44, 19th International Conference on Extending Database Technology, EDBT 2016 ; Conference date: 15-03-2016 Through 18-03-2016, 2016.

Fuss, S., Lamb, W. F., Callaghan, M. W., Hilaire, J., Creutzig, F., Amann, T., Beringer, T., de Oliveira Garcia, W., Hartmann, J., Khanna,
530   T., Luderer, G., Nemet, G. F., Rogelj, J., Smith, P., Vicente, J. L. V., Wilcox, J., del Mar Zamora Dominguez, M., and Minx, J. C.: Negative emissions — Part 2: Costs, potentials and side effects, Environmental Research Letters, 13, 063 002, https://doi.org/10.1088/1748-9326/aabf9f, 2018.

Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M. R., Peters, K., and Schober, D.: FAIR computational workflows, Data Intelligence, 2, 108–121, 2020.

535   Gustafson, A., Erdmann, J., Milligan, M., Onsongo, G., Pardey, P., Prather, T., Silverstein, K., Wilgenbusch, J., and Zhang, Y.: A platform for computationally advanced collaborative agroinformatics data discovery and analysis, in: Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact, pp. 1–4, Association for Computing Machinery, New York, NY, United States, 2017.

Harden, J. W., Hugelius, G., Ahlström, A., Blankinship, J. C., Bond-Lamberty, B., Lawrence, C. R., Loisel, J., Malhotra, A., Jackson, R. B.,
540   Ogle, S., Phillips, C., Ryals, R., Todd-Brown, K., Vargas, R., Vergara, S. E., Cotrufo, M. F., Keiluweit, M., Heckman, K. A., Crow,
S. E., Silver, W. L., DeLonge, M., and Nave, L. E.: Networking our science to characterize the state, vulnerabilities, and management
opportunities of soil organic matter, Global Change Biology, 24, e705–e718, https://doi.org/https://doi.org/10.1111/gcb.13896, 2018.

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., and Taylor, K. E.: A data model of the Climate and Forecast metadata conventions (CF-
1.6) with a software implementation (cf-python v2.1), Geoscientific Model Development, 10, 4619–4646, https://doi.org/10.5194/gmd-
545   10-4619-2017, 2017.

Hawkins, S. J., Firth, L. B., McHugh, M., Poloczanska, E. S., Herbert, R. J. H., Burrows, M. T., Kendall, M. A., Moore, P. J., Thompson,
R. C., Jenkins, S. R., Sims, D. W., Genner, M. J., and Mieszkowska, N.: Data rescue and re-use: Recycling old information to address new
policy concerns, Marine Policy, 42, 91–98, https://doi.org/10.1016/j.marpol.2013.02.001, 2013.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N.,
550   Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler,
I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLoS ONE, 12, e0169 748,
https://doi.org/10.1371/journal.pone.0169748, 2017.

Holmquist, J.: Coastal Carbon Database Structure: Naming Conventions for Attributes and Variables (Version 1), Accessed: Nov. 2021,
https://serc.si.edu/coastalcarbon/database-structure, 2018.

555 Holmquist, J.: Coastal Carbon Atlas, Accessed: Nov. 2021, https://ccrcn.shinyapps.io/CoastalCarbonAtlas/, 2021.

Holmquist, J., Klinges, D., Lonneman, M., and Wolfe, J.: Coastal Carbon Research Coordination Network Data Library, Accessed: Nov.
2021, https://github.com/Smithsonian/CCRCN-Data-Library, 2018 - 2021.

ISRaD: International Soil Radiocarbon Database: project webpage, Accessed: Nov. 2021, https://soilradiocarbon.org, 2018-2021.

Iversen, C. M., McCormack, M. L., Powell, A. S., Blackwood, C. B., Freschet, G. T., Kattge, J., Roumet, C., Stover, D. B., Soudzilovskaia,
560   N. A., Valverde-Barrantes, O. J., van Bodegom, P. M., and Violle, C.: A global Fine-Root Ecology Database to address below-ground
challenges in plant ecology, New Phytologist, 215, 15–26, https://doi.org/https://doi.org/10.1111/nph.14486, 2017.

Jeroen Ooms: pdftools: Text Extraction, Rendering and Converting of PDF Documents, https://CRAN.R-project.org/package=pdftools, R
package version 3.0.1, 2021.

Karasti, H., Baker, K. S., and Halkola, E.: Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infras-
565   tructuring in the Long Term Ecological Research (LTER) Network, Computer Supported Cooperative Work (CSCW), 15, 321–358,
https://doi.org/10.1007/s10606-006-9023-2, 2006.

Lawrence, C. R., Beem-Miller, J., Hoyt, A. M., Monroe, G., Sierra, C. A., Stoner, S., Heckman, K., Blankinship, J. C., Crow, S. E., McNicol,
G., Trumbore, S., Levine, P. A., Vindušková, O., Todd-Brown, K., Rasmussen, C., Hicks Pries, C. E., Schädel, C., McFarlane, K., Doetterl,
S., Hatté, C., He, Y., Treat, C., Harden, J. W., Torn, M. S., Estop-Aragonés, C., Asefaw Berhe, A., Keiluweit, M., Della Rosa Kuhnen, A.,
570   Marin-Spiotta, E., Plante, A. F., Thompson, A., Shi, Z., Schimel, J. P., Vaughn, L. J. S., von Fromm, S. F., and Wagai, R.: An open-source
database for the synthesis of soil radiocarbon data: International Soil Radiocarbon Database (ISRaD) version 1.0, Earth System Science
Data, 12, 61–76, https://doi.org/10.5194/essd-12-61-2020, 2020.

Lehmann, J., Bossio, D. A., Kögel-Knabner, I., and Rillig, M. C.: The concept and future prospects of soil health, Nature Reviews Earth &
Environment, 1, 544–553, https://doi.org/10.1038/s43017-020-0080-8, 2020.

575 Lin, D., Crabtree, J., Dillo, I., Downs, R. R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., et al.: The
TRUST Principles for digital repositories, Scientific Data, 7, 1–5, 2020.

Löffler, F., Wesp, V., König-Ries, B., and Klan, F.: Dataset search in biodiversity research: Do metadata in data repositories reflect scholarly information needs?, PLOS ONE, 16, e0246 099, https://doi.org/10.1371/journal.pone.0246099, 2021.

Lohr, S.: For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights, New York Times, https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html, 2014.

Longo, D. L. and Drazen, J. M.: Data Sharing, New England Journal of Medicine, 374, 276–277, https://doi.org/10.1056/nejme1516564, 2016.

LTER Soil Orgnaic matter Working Group: soilHarmonization: Homogenize LTER Soil Organic Matter Working Group data and notes, Accessed: Nov. 2021, https://github.com/lter/soilHarmonization, 2018b.

LTER Soil Orgnaic matter Working Group: Advancing soil organic matter research: Synthesizing multi-scale observations, Manipulations and Models, Accessed: Nov. 2021, https://lternet.edu/working-groups/soil-organic-matter/, 2019a.

Lyons, S. E., Osmond, D. L., Slaton, N. A., Spargo, J. T., Kleinman, P. J., Arthur, D. K., and McGrath, J. M.: FRST: A national soil testing database to improve fertility recommendations, Agricultural & Environmental Letters, 5, e20 008, 2020.

Malhotra, A., Todd-Brown, K., Nave, L., Batjes, N., Holmquist, J., Hoyt, A., Iversen, C., Jackson, R. B., Lajtha, K., Lawrence, C., Vindušková, O., Wieder, W., Williams, M., Hugelius, G., and Harden, J.: The landscape of soil carbon data: Emerging questions, synergies and databases, Progress in Physical Geography: Earth and Environment, 43, 707 – 719, 2019.

Nave, L., Johnson, K., Van Ingen, C., Agarwal, D., Humphrey, M., and Beekwilder, N.: International Soil Carbon Network (ISCN) Database v3-1, https://doi.org/10.17040/ISCN/1305039, 2016.

Onerhime, E.: Data Standards for Soil: Why aren't they taking root? [version 1; not peer reviewed], document 5:74, Gates Open Research, 2021.

Palma, R., Janiak, B., Reznik, T., Schleidt, K., Kozel, J., Sousa, L. D., Egmond, F., Mouazen, A. M., and Moshou, D.: Global Soil Information System (GloSIS) Ontology, Tech. rep., SIEUSOIL project., http://w3id.org/glosis/model, 2020.

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., and Dierolf, U.: Making Research Data Repositories Visible: The re3data.org Registry, PLoS ONE, 8, e78 080, https://doi.org/10.1371/journal.pone.0078080, 2013.

Pick, J. L., Nakagawa, S., and Noble, D. W.: Reproducible, flexible and high-throughput data extraction from primary literature: The metaDigitise R package, bioRxiv, https://doi.org/10.1101/247775, 2018.

Richardson, D. B., Kwan, M.-P., Alter, G., and McKendry, J. E.: Replication of scientific research: addressing geoprivacy, confidentiality, and data sharing challenges in geospatial research, Annals of GIS, 21, 101–110, 2015.

Richter, D. d., Bacon, A. R., Mobley, M. L., Richardson, C. J., Andrews, S. S., West, L., Wills, S., Billings, S., Cambardella, C. A., Cavallaro, N., DeMeester, J. E., Franzluebbers, A. J., Grandy, A. S., Grunwald, S., Gruver, J., Hartshorn, A. S., Janzen, H., Kramer, M. G., Ladha, J. K., Lajtha, K., Liles, G. C., Markewitz, D., Megonigal, P. J., Mermut, A. R., Rasmussen, C., Robinson, D. A., Smith, P., Stiles, C. A., Tate III, R. L., Thompson, A., Tugel, A. J., van Es, H., Yaalon, D., and Zobeck, T. M.: Human–Soil Relations are Changing Rapidly: Proposals from SSSA's Cross-Divisional Soil Change Working Group, Soil Science Society of America Journal, 75, 2079–2084, https://doi.org/https://doi.org/10.2136/sssaj2011.0124, 2011.

Rohatgi, A.: Webplotdigitizer: Version 4.5, https://automeris.io/WebPlotDigitizer, 2021.

Savage, C. J. and Vickers, A. J.: Empirical Study of Data Sharing by Authors Publishing in PLoS Journals, PLoS ONE, 4, e7078, https://doi.org/10.1371/journal.pone.0007078, 2009.

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D. P., Batjes, N. H., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J. E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., and Klumpp, K.: How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal, Global Change Biology, 26, 219–241, https://doi.org/https://doi.org/10.1111/gcb.14815, 2020.

Todd-Brown, K.: An open list of soil databases, Embargo-ed repository: Nov 2021, https://github.com/ktoddbrown/soil_database_list, 2021.

Tugel, A., Herrick, J., Brown, J., Mausbach, M., Puckett, W., and Hipple, K.: Soil Change, Soil Survey, and Natural Resources Decision Making, Soil Science Society of America Journal, 69, https://doi.org/10.2136/sssaj2004.0163, 2005.

Tummers, B.: DataThief software, version 3.0, https://datathief.org/, 2006.

Vines, T. H., Albert, A. Y., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., and Rennison, D. J.: The Availability of Research Data Declines Rapidly with Article Age, Current Biology, 24, 94–97, https://doi.org/https://doi.org/10.1016/j.cub.2013.11.014, 2014.

Weintraub, S. R., Flores, A. N., Wieder, W. R., Sihi, D., Cagnarini, C., Gonçalves, D. R. P., Young, M. H., Li, L., Olshansky, Y., Baatz, R., Sullivan, P. L., and Groffman, P. M.: Leveraging Environmental Research and Observation Networks to Advance Soil Carbon Science, Journal of Geophysical Research: Biogeosciences, 124, 1047–1055, https://doi.org/https://doi.org/10.1029/2018JG004956, 2019.

Wieder, W. R., Pierson, D., Earl, S. R., Lajtha, K., Baer, S., Ballantyne, F., Berhe, A. A., Billings, S., Brigham, L. M., Chacon, S. S., Fraterrigo, J., Frey, S. D., Georgiou, K., De Graaff, M.-A., Grandy, A. S., Hartman, M. D., Hobbie, S. E., Johnson, C., Kaye, J., Snowman, E., Litvak, M. E., Mack, M. C., Malhotra, A., Moore, J. A. M., Nadelhoffer, K., Rasmussen, C., Silver, W. L., Sulman, B. N., Walker, X., and Weintraub, S.: SOils DAta Harmonization database (SoDaH): an open-source synthesis of soil data from research networks, https://doi.org/10.6073/PASTA/9733F6B6D2FFD12BF126DC36A763E0B4, 2020.

Wieder, W. R., Pierson, D., Earl, S., Lajtha, K., Baer, S. G., Ballantyne, F., Berhe, A. A., Billings, S. A., Brigham, L. M., Chacon, S. S., Fraterrigo, J., Frey, S. D., Georgiou, K., de Graaff, M.-A., Grandy, A. S., Hartman, M. D., Hobbie, S. E., Johnson, C., Kaye, J., Kyker-Snowman, E., Litvak, M. E., Mack, M. C., Malhotra, A., Moore, J. A. M., Nadelhoffer, K., Rasmussen, C., Silver, W. L., Sulman, B. N., Walker, X., and Weintraub, S.: SoDaH: the SOils DAta Harmonization database, an open-source synthesis of soil data from research networks, version 1.0, Earth System Science Data, 13, 1843–1854, https://doi.org/10.5194/essd-13-1843-2021, 2021.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data, 3, https://doi.org/10.1038/sdata.2016.18, 2016.

Yost, J. L. and Hartemink, A. E.: How deep is the soil studied–an analysis of four soil science journals, Plant and soil, 452, 5–18, 2020.