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

Katherine E. O. Todd-Brown[1], Rose Z. Abramoff[2, 3], Jeffrey Beem-Miller[4], Hava K. Blair[5], Stevan Earl[6], Kristen J. Frederick[1], Daniel R. Fuka[7], Mario Guevara Santamaria[8], Jennifer W. Harden[9], Katherine Heckman[10], Lillian J. Haren[1], James R. Holmquist[11], Alison M. Hoyt[9], David H. Klinges[12], David S. LeBauer[13], Avni Malhotra[9, 14], Shelby C. McClelland[15], Lucas E. Nave[16], Katherine S. Rocci[17], Sean M. Schaeffer[18], Shane Stoner[4, 19], Natasja van Gestel[20], Sophie F. von Fromm[4, 19], 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]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[4]Max Planck Institute for Biogeochemistry, Jena, Germany
[5]Department of Soil, Water, and Climate, University of Minnesota, St Paul, MN, USA
[6]Global Institute of Sustainability and Innovation, Arizona State University, Tempe, AZ, USA
[7]Virginia Polytechnic Institute and State University, Blacksburg, VA, USA
[8]Centro de Geociencias, Universidad Nacional Autónoma de México, Juriquilla, Querétaro, Mexico
[9]Department of Earth System Science, Stanford University, Stanford, CA, USA
[10]Northern Research Station, USDA Forest Service, Houghton, MI, USA
[11]Smithsonian Environmental Research Center, Edgewater, Maryland, USA
[12]School of Natural Resources and Environment, University of Florida, Gainesville, Florida, USA
[13]Arizona Experiment Station, College of Agriculture and Life Sciences, University of Arizona, Tuscon, AZ, USA
[14]Department of Geography, University of Zürich, Zürich, Switzerland
[15]Department of Soil and Crop Sciences, Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO, USA
[16]Biological Station and Dept. of Ecology and Evolutionary Biology, University of Michigan, Pellston, MI, USA
[17]Natural Resource Ecology Laboratory, Department of Soil and Crop Sciences, Graduate Degree Program in Ecology, Colorado State University, Fort Collins, CO, USA
[18]Biosystems Engineering and Soil Science Department, University of Tennessee, Knoxville, TN, USA
[19]Department of Environmental Systems Science, ETH Zürich, Zürich, Switzerland
[20]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 and richer than ever, but -outside of a few large soil survey resources- remain largely unusable for informing soil management and understanding Earth system processes ~~outside~~ beyond of the orig-

inal 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~~new insight. 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~~availability, input, harmonization, curation, and publication. We then suggest new soil-focused semantic tools to improve existing data pipelines, such as ontologies, vocabulary lists, and community 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 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 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., 2021b), 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 (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, 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", for the purposes of this paper, is defined as a synthetic collection of soil observations, including location, physicochemical properties, units, and methods. Databases can 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 interme-
diate datasets, or large existing databases. Further, these databases are often cross-referenced, for example, by integrating geo-
referenced meteorological data or remote sensing products like net primary production maps. "Harmonization" of data refers
to the 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 systematic 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 practicesand guidance in creating them~~A number of databases have
been compiled in soils around specific themes or measurement types including: soil carbon and nitrogen (Worldwide soil
carbon and nitrogen data (Zinke et al., 1998), International Soil Carbon Network database (Nave et al., 2016)), field based
soil respiration (Soil Respiration Data base (Bond-Lamberty and Thomson, 2010)), lab-based heterotrophic respiration (Soil
Incubation Database (Schädel et al., 2020)), Soil radiocarbon (International Soil Radiocarbon Database (Lawrence et al., 2020)
), and coastal soils (Coastal Carbon Research Coordination Network Database (Holmquist et al., 2018 - 2021)) (see Section
A). The soil resources and data products that make up the World Soil Information Service (WoSIS) curated by International
Soil Reference and Information Centre (ISRIC) (Batjes et al., 2020) is an example of how soil data feed into larger products.
After archival on ISRIC servers, datasets from individual providers are incorporated in the the World Soil Information Servers
workflow of mapping diverse data contributions in to a standard data model, harmonization and distribution (including both
databases and derived data products, such as SoilGrids (Hengl et al., 2017; Hengl et al., 2014)). Harmonized databases can be
a powerful validation tool for Earth System models as well as other modeling and estimation efforts and to construct these
databases the underlying data should adhere to best practices.

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 previ-
ous ~~focus~~ work on FAIR principles has focused on data access (though ~~some work has focused elsewhere (Goble et al., 2020)~~
notable counter examples include Goble et al. (2020)), which can be difficult if ~~it relies on the standard~~ data relies on an
'available on request' note included at the end of publications (Savage and Vickers, 2009; Vines et al., 2014). Indeed previous
research has identified challenges with educating and motivating data providers to publish their datasets (Wolkovich et al., 2012)
. It is important to note here that FAIR does not always mean open freely reusable data. Indeed the FAIR Data motto makes
this difference: "as open as possible, as closed as necessary", and this becomes particularly important for data that has possible
economic impact (Luque, 2019). While data access has improved ~~due to~~ as 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~~ reflect 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. The approach and issues outlined in this paper are undoubtedly not unique to soils and are relevant to a wide range of scientific data, particularly environmental data. However we present this as a case study of soil specific database construction. 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.

2.1 Availability

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 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~~ and whom is responsible for ensuring these standards are met. To complicate matters, 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). This is not to say that
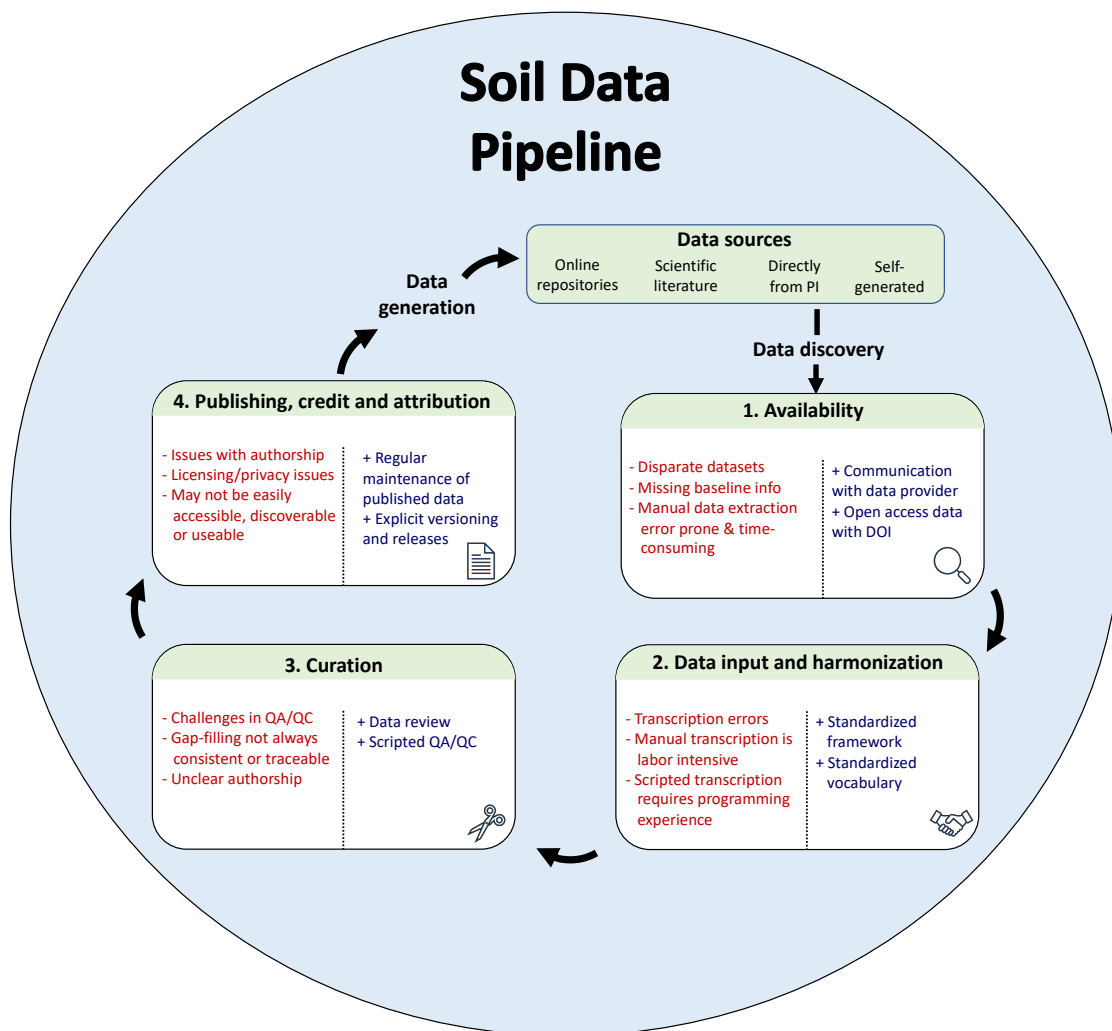
Figure 1. Database pipeline with pain points (indicated by "-") and suggestions for improvement (indicated by "+") to conform more closely to FAIR Data Principles. Data sources can be diverse, including published (online repositories and scientific literature) and unpublished sources (direct from principal 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.

archiving data for the purpose of meeting funder requirements or reproducing the associated analysis can not be useful in and of itself; however this does not automatically lend the data to integration in a database.

On the data aggregation side, many data aggregators are challenged by unclear data documentation and metadata. This ~~leads to~~ ambiguity can lead to a range of 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 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~~ trade-offs with contacting the original data providers. 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 aggregator who, in turn, is left with an ever expanding list of coauthors with varying levels of involvement (see ~~Publication section). This~~ Section 2.4). Differing expectations on acknowledgements can lead to conflict and result in a lack of trust in the community (Longo and Drazen, 2016).

~~In general however, we feel that direct~~ Direct collaboration between data providers and data aggregators is a critical relationship to nurture~~.~~, but exists in a broader context of good data stewardship practices. The FAIR Principles for good data practices have been thoughtfully extended via TRUST and CARE. The TRUST Principles (Transparency; Responsibility; User focus; Sustainability; Technology) articulate key features for trustworthy digital repositories, which are essential for preserving data access and reuse over time (Lin et al., 2020). The CARE Principles for Indigenous Data Governance (Collective benefit, Authority to control; Responsibility; Ethics) position decisions related to data management and reuse in the context of Indigenous culture and knowledge systems, highlighting actions that ultimately support community-based data sovereignty (Carroll et al., 2020). As the community continues to ~~converge on~~ discuss 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' is only the beginning.

2.2 Data input and harmonization

~~Datasets~~ Data formats typically reflect the purposes of the ~~original study; however examining those same data in a broader study; thus~~ examining data to address a different question or 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 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 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~~ manual 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 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 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 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). While this approach has the most explicit need for clear semantic resources, these are also essential for creating effective manual transcription templates and protocols. Such a generalized approach can be more easily extended to expand the number of data sources. However there is currently no broadly agreed upon annotation vocabulary in the field of soil science, 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

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 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
180 database tables (Lawrence et al., 2020). Following this initial filter, a manual individual 'expert review' is conducted by a trained ISRaD volunteer (see Section A2). These extensive quality control procedures require time and diverse expertise, making them 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~~ For example, expanding the environmental context for a particular soil sample
185 location, such as extracting net primary productivity and land use classification from satellite products, is one example of gap-filling soil-relevant observations that may not have been collected at the time of the soil sampling. Strategies for gap-filling more broadly 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.

190 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~~ and reduce 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.

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~~ authorship issues are common in large collaborative projects (Cooke and Hilton, 2015). This can be mitigated with an ~~expanded~~ inclusive co-author list to include both data providers, aggregators, and re-analysis teams, but requires significant project management and organizational overhead. As always with
200 larger team science, 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 entitledto (i.e. if the data are there, use them),~~ is accessible and legally entitled, we strongly feel that ~~collaborative~~ more inclusive projects build trust within the scientific community, leading to better data interpretation and seeding future collaborations.

205 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.
210 The most permissive is CC-0 or public domain license that puts no restriction on data use. A 'By-Acknowledgment' (BY) rider ~~requests~~ requires 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 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 version of the database.

3 Future of soil data

The hope of big open data is to have any data collected at any time anywhere in the world at the tip of your fingers (see Section B). To reach this hope it's important not just to work with large volumes of data but also diverse observations and measurements in a way that is trusted. 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 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 participate 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 will often ask 'what should I measure' to be relevant to data aggregation efforts, and there are efforts to provide such guidance (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 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., 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 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.

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 and depending on the measurement the exact resolution will vary (for example, parent material may just need the decade of collection while soil respiration may need a minute resolution). 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 should be a high priority for data rescue and documentation.

~~Decadal scale~~Long-term (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~~ Often referred to as data 'rescuing', there is an staggering potential of decadal-scale data sitting in labs across the world but best practices on how to use this data are lacking. Decadal data are often associated with a particular researcher or group ~~that are typically released at the retirement of the primary investigator, and~~ and may represent an entire career of data collection that likely have been reformatted multiple times across several generations of storage 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~~ Augmenting (or extending) rescuing data from a single group is resurveying older sites and 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 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 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.

The uses of soil databases for research context are varied (for example: Earth system model benchmarking (Collier et al., 2018)) but there are other private economic impacts of having soil data available. Soil health metrics in public databases could impact land fertility evaluation and there is increasing interest in soil carbon data by carbon markets for offsetting CO_2 emissions. As mentioned in Section 3.1, specific information on nutrient and water retention of a soil can make it more or less valuable leading to landowner reluctance to release data. More recently an increasing interest in generating carbon offsets by increasing soil carbon sequestration has lead to a proliferation of new venture corporations that either generate new or use available soil data in order to define land management practices (e.g. IndigoAg, CIBO Technologies, Seqana, Regrow, Nori, LoamBio). Industry companies generally treat data that they collect or process as part of their intellectual property, which is pent private. While there is clearly scientific value in these data, it is unclear how researchers, landholders, and private companies will negotiate the use and integration of these data into research output. Nonetheless, privately held data would also benefit from and contribute to connection with community developed tools.

3.2.1 Community tools

There is an understandable tendency by many scientists towards ~~standards~~data model standards (for example Rüegg et al. (2014) and as proposed for soil respiration data (Bond-Lamberty et al., 2021)). If all soil data adhered to a common template, or 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 standards ~~are still useful (Onerhime, 2021)~~as well as software development practices like versioning could improve interoperability (Onerhime, 2021; Crystal-Ornelas et al., 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~~ This digital vocabulary has roots in classical soil glossaries and lab manuals that have been printed as dictionary style references~~that are difficult to transform into digital resources for both copyright reasons 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, in addition to addressing legacy copyright issues. 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 underutilized in the soil community. The soil research 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~~ the data community needs 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 a diverse community. In the development phase, a diverse community with a range of expertise, stakeholders, and career stages 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 open, in the sense that anyone can join or contribute and are empowered through educational activities to participate. Transparency 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~~ Finally, 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 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, short-term shared purpose is most easily motivated by a synthesis paper or research question, but longer-term motivators are unclear. 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~~availability, input, harmonization, 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 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 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

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 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. (2021b, 2020)). ~~This is appendix not meant to be exhaustive. For a living list of researcher-driven soil databases please see Todd-Brown (2021)~~Finally we also include a broader list of soil data collections to illustrate more generally how the field approaches these projects (Section A5).

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 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 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

The International Soil Radiocarbon Database (ISRaD) consists of an open-source database of soil and soil-related radiocarbon 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 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 (based on templates designed in the International Soil Carbon Network (Nave et al., 2016)) for data ingestion which each contributor fills out and submits to a 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 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

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 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 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 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 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 that, as of this writing, includes data from over seventy unique studies (Wieder et al., 2020, 2021b). 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.

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 and based off of templates designed in the International Soil Carbon Network (Nave et al., 2016). 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 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.

A5 Data collection list

This is appendix is meant to be representative, not exhaustive.

1. SOils DAta Harmonization (SoDaH) & Synthesis (Wieder et al., 2020). Data available at https://lter.github.io/som-website/database.html (Wieder et al., 2021a). See Section A4

    – **Personnel**: 5 people over 4 yrs

    – **Description**: Layer-level soil data includes time series and experimental data.

    – **Pipeline**: Aggregation of data from broad scale research networks. After level-0 data was contributed (both raw data +metadata template) SoDah supplies a script. First harmonization of raw data into a common format takes place and then it is aggregated into an a flat csv file.

    – **Size**: 303.6 MB

2. Soil warming meta-anlaysis (Crowther et al., 2016). Data available on request. See Section A1

    – **Personnel**: 2 data aggregators with multitude of contributors and coauthors over 4 yrs

    – **Description**: Soil C in ambient and warmed soils. Field soil warming experiments

    – **Pipeline**: Used a template approach. The data template only asked for a brief description of the site location (latitude, longitude, etc.), the measurement they are looking for in their project, and how the author collected that measurement. This very restricted data model is then much easier (and more satisfying) to collaborate with data products to fill out.

    – **Size**: 114 KB

3. vanGestel - field treatments van Gestel et al. (2018). Data available in supplementary information and raw data available on request

    – **Personnel**: 2 over 3 years

    – **Description**: Soil C in ambient and warmed soils. Field treatments and soil warming

    – **Pipeline**: Excel was used as the repository for raw data (no unit conversion, no gap-filling etc.), with different sheets comprised of different aspects of the C cycle (soil C, plant C etc), and 1 sheet including meta-data. All computations/harmonization/QA/QC etc. done in R.

    – **Size**: 1 MB

4. World Soil Information Service (WoSIS) (Batjes et al., 2020). Data available at https://www.isric.org/explore/wosis (Batjes and Calis .

    – **Personnel**: Institutional, 20 years and ongoing

    – **Description**: Focuses on bulk soil characteristics. To serve the user with a selection of standardised/ harmonised soil profile data. These quality-assessed data may be used to underpin digital soil mapping and a range of global assessments.

    – **Pipeline**: Submitted data is first preserved in the ISRIC WDC-Soils Data Repository. It is then quality-assessed, standardised, and imported into the WoSIS data model itself. It is harmonized where possible using consistent procedures.

    – **Size**: Varies based on portion of database

5. SoilHealthDB (Jian et al., 2020). Data available at https://github.com/jinshijian/SoilHealthDB (Jian et al.).

    – **Personnel**: 3 people, initial publication 2020 with ongoing data contributions

    – **Description**: 42 soil health indicators and 45 background indicators. The primary goal is to enable the research community to perform comprehensive analyses of soil health changes related to cropland conservation management.

- **Pipeline**: Data extracted to Excel template; basic scripted QA/QC scripts; no unit conversion/harmonization (leaves this to future users); database and code available on github
- **Size**: 64MB including metadata and documentation

6. International Soil Carbon Network 3 (ISCN3). Data available at https://iscn.fluxdata.org/data/ (ISC).

- **Personnel**: 5 leads over 4 years to first publication and ongoing
- **Description**: Soil C stocks and associated site and bulk soil characteristics. To provide a science-based network that facilitates data sharing, assembles databases, identifies gaps in data coverage, and enables spatially explicit assessments of soil carbon in context of landscape, climate, land use, and biotic variables
- **Pipeline**: SQL-based ingest of Excel templates or Access tables, hybrid scripted-manual QA/QC, rules-based computation of soil C stocks
- **Size**: 300 Mb

7. Soil Incubation Database (SIDb) (Schädel et al., 2020). Data available at https://soilbgc-datashare.github.io/sidb/ (Sierra)

- **Personnel**: 8 people over 4 years to first publication and ongoing
- **Description**: Soil incubated CO_2 and CO_2 isotope data along with ancillary variables. Time series data of soil incubations and soil respiration.
- **Pipeline**: Templates for data in csv format and metadata in ymal format. The data is processed by an R package that checks for consistency between data and metadata. The R package also provides functions to access and manipulate data.
- **Size**: 1 MB

8. Coastal Carbon Research Coordination Network (CC-RCN). Data available at https://serc.si.edu/coastalcarbon/about (Holmquist, 2018). See Section A3.

- **Personnel**: 3 PIs, 5 steering group members, 2 staff over 5 yrs
- **Description**: Tidal wetland soil C stocks and ancillary information. Provide access to data, analysis tools and synthesis products to accelerate the pace of discovery in coastal wetland C science.
- **Pipeline**: Once the data is attained, the technician codes out the methods metadata, then does an initial data inspection by going through one or two drafts of a data release, and finally uploads it to Figshare. Custom R-functions and hook scripts are used for preparing the actual data releases. They do things like take quick looks at the range of the datasets to make sure that everything is how it should be. When preparing data releases, they prepare 'EML' style metadata.

- **Size**: 14.3 MB

9. International Soil Radiocarbon Database (ISRaD) (Lawrence et al., 2020). Data available at https://soilradiocarbon.org/ (Beem-Miller et al., 2021). See Section A2.

  - **Personnel**: 30+ people, 6 yrs to first publication and ongoing
  - **Description**: Soil physicochemical data, isotopic data, fractionation data, incubation data. Aggregation and harmonization of radiocarbon and fraction data from previously published work.
  - **Pipeline**: Standardized Excel template; scripted QA/QC; github repository
  - **Size**: 31 MB

10. SoilTemp (Lembrechts et al., 2020). Data available at https://soiltemp.weebly.com/ (Soi)

  - **Personnel**: Core team of 5-10 developers, 100s of data contributors. 5 yrs and ongoing
  - **Description**: In-situ soil temperature and other microclimate time series, soil surface vegetation characteristics. Building a global database of soil and near-surface temperatures, and associated species occurrence data for use in ecological modelling.
  - **Pipeline**: Provides a CSV-format template to data users, and tools in development to facilitate submission via an R package. Some curation and harmonization is performed via a scripted approach, although the workflow is in active development.
  - **Size**: 30 GB

Appendix B: Future Dream

It is often difficult to envision how inter-connected soil data could impact science. Here we have 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 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.

565 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
570 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
575 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

580 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 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.

585 ~~Big corporations and governments have taken on the task of parcelling out~~ Corporations and governments increasingly recognize the value of legacy datasets and actively participate in disseminating data to the ~~general public, stripping out sensitive information reduces the utility of the data but is seen as a necessaryevil. Ethicists are still debating whether this gives the corporations~~public, though redaction of sensitive information is necessary. While the impact of 'good'/~~governments too much knowledge, and~~ bad' data metrics on land valuation are still a concern, double-blind methods are being developed to ~~obfuscate~~
590 ~~sensitive data even from the data holders~~increase the security of data sharing and automate the obfuscation of sensitive data.

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 management 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 been a few day workshop into a longer ~~slow- burn~~ 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 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.

Competing interests. The authors declare no competing interests.

610   *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 the Department of Energy's Biological and Environmental Research (DOE-BER) program (DE-SC0010632). Rose Abramoff was funded by
615   the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 834169 and the United States Department of Energy, Office of Science, Office of Biological and Environmental Research. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the United States Department of Energy under contract DE-AC05-00OR22725.

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),
620   Chelsea Carey (Point Blue Conservation Science), Gabriel Reuben Smith (Stanford University; ETH Zurich) for their contribution to conversations on this topic.

The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

References

625   ISCN - Data, https://iscn.fluxdata.org/data/.

SOILTEMP: Towards a global database of microclimate, https://soiltemp.weebly.com/.

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

Batjes, N. and Calisto, L.: ISRIC - WoSIS Soil Profile Database, https://www.isric.org/explore/wosis.

Batjes, N. H., Ribeiro, E., and van Oostrum, A.: Standardised soil profile data to support global mapping and modelling (WoSIS snapshot

630   2019), Earth System Science Data, 12, 299–320, https://doi.org/10.5194/essd-12-299-2020, 2020.

Beem-Miller, J., Monroe, J. G., Hoyt, A. M., Stoner, S., von Fromm, S. F., 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
for E-Democracy and Open Government (CeDEM), pp. 69–81, https://doi.org/10.1109/CeDEM.2017.22, 2017.

635   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 or-
ganic 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.

Bond-Lamberty, B. and Thomson, A.: A global database of soil respiration data, Biogeosciences, 7, 1915–1926, https://doi.org/10.5194/bg-
640   7-1915-2010, 2010.

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,
645   D., Macdonald, C. A., Maseyk, K., Mathes, K., Mauritz, M., Mayes, M. A., McNulty, S., Miao, G., Migliavacca, M., Miller, S., Miniat,
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,
650   T., Zeng, J., Zha, T., Zhang, Q., and Zou, J.: COSORE: A community database for continuous soil respiration and other soil-atmosphere
greenhouse gas flux data, Global Change Biology, 26, 7268–7283, https://doi.org/https://doi.org/10.1111/gcb.15353, 2020.

Bond-Lamberty, B., Christianson, D. S., Crystal-Ornelas, R., Mathes, K., and Pennington, S. C.: A reporting format for field measurements
of soil respiration, Ecological Informatics, 62, 101 280, https://doi.org/https://doi.org/10.1016/j.ecoinf.2021.101280, 2021.

Buttigieg, P. L., Pafilis, E., Lewis, S. E., Schildhauer, M. P., Walls, R. L., and Mungall, C. J.: The environment ontology in 2016: bridging do-
655   mains 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-
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.

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, 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, 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, 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.

Crystal-Ornelas, R., Varadharajan, C., Bond-Lamberty, B., Boye, K., Burrus, M., Cholia, S., Crow, M., Damerow, J., Devarakonda, R., Ely, K. S., Goldman, A., Heinz, S., Hendrix, V., Kakalia, Z., Pennington, S. C., Robles, E., Rogers, A., Simmonds, M., Velliquette, T., Weierbach, H., Weisenhorn, P., Welch, J. N., and Agarwal, D. A.: A Guide to Using GitHub for Developing and Versioning Data Standards and Reporting Formats, Earth and Space Science, 8, e2021EA001 797, https://doi.org/https://doi.org/10.1029/2021EA001797, e2021EA001797 2021EA001797, 2021.

CSDMS: CSN Searchable List — Community Surface Dynamics Modeling System (vs 0.8.3), wiki, https://csdms.colorado.edu/mediawiki/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.

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.

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, 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.

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., 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-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., de Jesus, J. M., MacMillan, R. A., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Walsh, M. G., and Gonzalez, M. R.: SoilGrids1km — Global Soil Information Based on Automated Mapping, PLOS ONE, 9, 1–17, https://doi.org/10.1371/journal.pone.0105992, 2014.

Hengl, T., Mendes de Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., 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.

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.

735   Iversen, C. M., McCormack, M. L., Powell, A. S., Blackwood, C. B., Freschet, G. T., Kattge, J., Roumet, C., Stover, D. B., Soudzilovskaia, 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.

740   Jian, J., Stewart, R. D., and Du, X.: SoilHealthDB, https://github.com/jinshijian/SoilHealthDB.

     Jian, J., Du, X., and Stewart, R. D.: A database for global soil health assessment, Scientific Data, 7, 16, https://doi.org/10.1038/s41597-020-0356-3, 2020.

     Karasti, H., Baker, K. S., and Halkola, E.: Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network, Computer Supported Cooperative Work (CSCW), 15, 321–358,
745   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., 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
750   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.

     Lembrechts, J. J., Aalto, J., Ashcroft, M. B., Frenne, P. D., Kopecký, M., Lenoir, J., Luoto, M., Maclean, I. M. D., Roupsard, O., Fuentes-
755   Lillo, E., García, R. A., Pellissier, L., Pitteloud, C., Alatalo, J. M., Smith, S. W., Björk, R. G., Muffler, L., Cesarz, S., Gottschall, F., Backes, A. R., Okello, J., Urban, J., Plichta, R., Svátek, M., Phartyal, S. S., Wipf, S., Eisenhauer, N., Pușcaș, M., Turtureanu, P. D., Varlagin, A., Dimarco, R. D., Jump, A. S., Randall, K., Dorrepaal, E., Larson, K., Walz, J., Vitale, L., Svoboda, M., Higgens, R. F., Halbritter, A. H., Curasi, S. R., Klupar, I., Koontz, A., Pearse, W. D., Simpson, E., Stemkovski, M., Graae, B. J., Sørensen, M. V., Høye, T. T., Calzado, M. R. F., Lorite, J., Carbognani, M., Tomaselli, M., Forte, T. G. W., Petraglia, A., Haesen, S., Somers, B., Meerbeek, K. V., Björkman, M. P.,
760   Hylander, K., Merinero, S., Gharun, M., Buchmann, N., Dolezal, J., Matula, R., Thomas, A. D., Bailey, J. J., Ghosn, D., Kazakis, G., Pablo, M. A. d., Kemppinen, J., Niittynen, P., Rew, L., Seipel, T., Larson, C., Speed, J. D. M., Ardö, J., Cannone, N., Guglielmin, M., Malfasi, F., Bader, M. Y., Canessa, R., Stanisci, A., Kreyling, J., Schmeddes, J., Teuber, L., Aschero, V., Čiliak, M., Máliš, F., Smedt, P. D., Govaert, S., Meeussen, C., Vangansbeke, P., Gigauri, K., Lamprecht, A., Pauli, H., Steinbauer, K., Winkler, M., Ueyama, M., Nuñez, M. A., Ursu, T.-M., Haider, S., Wedegärtner, R. E. M., Smiljanic, M., Trouillier, M., Wilmking, M., Altman, J., Brůna, J., Hederová, L., Macek, M.,
765   Man, M., Wild, J., Vittoz, P., Pärtel, M., Barančok, P., Kanka, R., Kollár, J., Palaj, A., Barros, A., Mazzolari, A. C., Bauters, M., Boeckx, P., Alonso, J. L. B., Zong, S., Cecco, V. D., Sitková, Z., Tielbörger, K., Brink, L. v. d., Weigel, R., Homeier, J., Dahlberg, C. J., Medinets, S., Medinets, V., Boeck, H. J. D., Portillo-Estrada, M., Verryckt, L. T., Milbau, A., Daskalova, G. N., Thomas, H. J. D., Myers-Smith, I. H., Blonder, B., Stephan, J. G., Descombes, P., Zellweger, F., Frei, E. R., Heinesch, B., Andrews, C., Dick, J., Siebicke, L., Rocha, A., Senior, R. A., Rixen, C., Jimenez, J. J., Boike, J., Pauchard, A., Scholten, T., Scheffers, B., Klinges, D., Basham, E. W., Zhang, J., Zhang,
770   Z., Géron, C., Fazlioglu, F., Candan, O., Bravo, J. S., Hrbacek, F., Laska, K., Cremonese, E., Haase, P., Moyano, F. E., Rossi, C., and Nijs, I.: SoilTemp: a global database of near-surface temperature, Global Change Biology, 00, 1–14, https://doi.org/10.1111/gcb.15123, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gcb.15123, 2020.

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.

775   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, 780   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.

785   Luque, C.: Open Data and FAIR Data: differences and similarities, https://www.ogoov.com/en/blog/open-data-and-fair-data-differences-and-similarities/, 2019.

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., Vin- 790   dušková, 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, 795   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., Schirm-bacher, P., and Dierolf, U.: Making Research Data Repositories Visible: The re3data.org Registry, PLoS ONE, 8, e78 080, 800   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.

805   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, 810   https://doi.org/https://doi.org/10.2136/sssaj2011.0124, 2011.

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

Rüegg, J., Gries, C., Bond-Lamberty, B., Bowen, G. J., Felzer, B. S., McIntyre, N. E., Soranno, P. A., Vanderbilt, K. L., and Weathers, K. C.: Completing the data life cycle: using information management in macrosystems ecology research, Frontiers in Ecology and the Environment, 12, 24–30, https://doi.org/https://doi.org/10.1890/120375, 2014.

815   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.

Schädel, C., Beem-Miller, J., Aziz Rad, M., Crow, S. E., Hicks Pries, C. E., Ernakovich, J., Hoyt, A. M., Plante, A., Stoner, S., Treat, C. C., and Sierra, C. A.: Decomposability of soil organic matter over time: the Soil Incubation Database (SIDb, version 1.0) and guidance for incubation procedures, Earth System Science Data, 12, 1511–1524, https://doi.org/10.5194/essd-12-1511-2020, 2020.

820   Sierra, C. A.: SIDB, https://soilbgc-datashare.github.io/sidb/.

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.

825   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.

van Gestel, N., Shi, Z., van Groenigen, K. J., Osenberg, C. W., Andresen, L. C., Dukes, J. S., Hovenden, M. J., Luo, Y., Michelsen,
830   A., Pendall, E., Reich, P. B., Schuur, E. A. G., and Hungate, B. A.: Predicting soil carbon loss with warming, Nature, 554, E4–E5, https://doi.org/10.1038/nature25745, 2018.

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.

835   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., Earl, S., and Pierson, D.: SoDaH: Database, https://lter.github.io/som-website/database.html, 2021a.

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.,
840   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.,
845   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, 2021b.

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.

Wolkovich, E. M., Regetz, J., and O'Connor, M. I.: Advances in global change research require open science by individual researchers,
Global Change Biology, 18, 2102–2110, https://doi.org/https://doi.org/10.1111/j.1365-2486.2012.02693.x, 2012.

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.

Zinke, P., Stangenberger, A., Post, W., Emanuel, W., and Olson, J.: Global Organic Soil Carbon and Nitrogen (Zinke et al.),
https://doi.org/10.3334/ORNLDAAC/221, 1998.