BNL-223193-2022-JAAM

Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats

R. Crystal-Ornelas, A. Rogers

To be published in "Scientific Data"

July 2022

Environmental and Climate Sciences Department

Brookhaven National Laboratory

U.S. Department of Energy

USDOE Office of Science (SC), Biological and Environmental Research (BER) (SC-23)

DISCLAIMER

Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats

Robert Crystal-Ornelas[1*], Charuleka Varadharajan[1], Dylan O'Ryan[1], Kathleen Beilsmith[2], Ben Bond-Lamberty[3], Kristin Boye[4], Madison Burrus[1], Shreyas Cholia[5], Danielle S. Christianson[5], Michael Crow[6], Joan Damerow[1], Kim S. Ely[7], Amy E. Goldman[8], Susan Heinz[6], Valerie C. Hendrix[5], Zarine Kakalia[1], Kayla Mathes[9], Fianna O'Brien[5], Stephanie Pennington[3], Emily Robles[1], Alistair Rogers[7], Maegen Simmonds[1,10], Terri Velliquette[6], Pamela Weisenhorn[2], Jessica Nicole Welch[6], Karen Whitenack[1], Deborah A. Agarwal[5]

[1]Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[2]Argonne National Laboratory, Lemont, IL 60439, USA

[3]Pacific Northwest National Laboratory, Joint Global Change Research Institute at the University of Maryland–College Park, College Park, MD 20740, USA

[4]Geochemistry and Biogeochemistry Group, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA

[5]Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720 USA

[6]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

[7]Environmental and Climate Sciences Department, Brookhaven National Laboratory, Upton, NY 11973, USA

[8]Pacific Northwest National Laboratory, Richland, WA 99354, USA

[9]Integrated Life Sciences, Virginia Commonwealth University, Richmond, VA 23284, USA

[10]Now at: Pivot Bio, 2910 Seventh Street, Berkeley, CA 94710, USA

Corresponding author: Charuleka Varadharajan (cvaradharajan@lbl.gov)

Abstract

Research can be more transparent and collaborative by using Findable, Accessible, Interoperable, and Reusable (FAIR) principles to publish Earth and environmental science data. Reporting formats—instructions, templates, and tools for consistently formatting data within a discipline—can help make data more accessible and reusable. However, the immense diversity of data types across Earth science disciplines makes development and adoption challenging. Here, we describe 11 community reporting formats for a diverse set of Earth science (meta)data including cross-domain metadata (dataset metadata, location metadata, sample metadata), file-formatting guidelines (file-level metadata, CSV files, terrestrial model data archiving), and domain-specific reporting formats for some biological, geochemical, and hydrological data (amplicon abundance tables, leaf-level gas exchange, soil respiration, water and sediment chemistry, sensor-based hydrologic measurements). More broadly, we provide guidelines that communities can use to create new (meta)data formats that integrate with their scientific workflows. Such reporting formats have the potential to accelerate scientific discovery and predictions by making it easier for data contributors to provide (meta)data that are more interoperable and reusable.

Keywords: data standards, metadata, FAIR data, open science, data repository

1. Introduction

Making Earth and environmental science data Findable, Accessible, Interoperable, and Reusable (FAIR)[1,2] contributes to research that is more transparent and reproducible[3]. Search engines and data repositories[2,4,5] have enabled advances in data preservation, findability, and accessibility. However, data interoperability and reuse remain major challenges in part due to the diversity of Earth science data, and because researchers may lack time and funding for data management or awareness of tools and resources to make data more reusable[5,6]. This results in barriers to scientific research and knowledge generation; for example, synthesis of data across different sources can be extremely time-consuming when data and metadata are not standardized in a common, well-defined format.

Standards for data and metadata, hereafter referred to as (meta)data standards, have been proposed as important elements to make Earth and environmental science data easier to find, understand and reuse[7–10]. Formal (meta)data standards are typically accredited by large governing bodies and emphasize making data broadly reusable[11]. For example, the International Organization for Standardization (ISO) 8601 standard provides guidelines for formatting date and timestamps and has been adopted in a wide range of research and business sectors[12]. The Open Geospatial Consortium's Sensor Observation Service standard[13] outlines standardized ways of pulling sensor data from web interfaces. Such accredited standards are extraordinarily useful, but are available only for a few environmental data types and can take over a decade to build governing processes and consensus.

In contrast, reporting formats are community efforts aimed at harmonizing diverse environmental data types without the oversight of the governing protocols or working groups that maintain vocabularies and extensive documentation. There are reporting formats for different research domains and data types including water quality[14] and meteorological data[15]. Reporting formats are typically more focused within scientific domains—for example, marine observations[16] or geoscience[17]. Reporting formats can enable efficient collection and harmonization of information needed to understand and reuse specific types of data within a research community. For example, the use of FLUXNET's half-hourly flux and meteorological reporting format[18] enables both access and reuse of consistently formatted carbon, water, and energy flux data from thousands of sampling locations across the world. However, reporting formats do not exist for most environmental data types, and even if they do, complexity and lack of resources can limit their adoption[9].

There are many scientific benefits when research communities adopt reporting formats, ranging from organizing data collection in the field or lab to more efficient data reuse in

synthesis and modeling efforts. Reporting formats can facilitate data sharing within a group, provide guidelines for consistent data collection, enable streamlined scientific workflows, and enable long-term preservation of knowledge that may not be typically stored or reported with the data[19,20]. Moreover, research disciplines are beginning to operationalize and implement practices[21,22] to achieve the original FAIR guiding principles[21,22]. Reporting formats developed by the research communities for which they are intended are seen as a critical step toward achieving greater data interoperability and reuse[22].

A variety of multidisciplinary data are generated in research sponsored by the U.S. Department of Energy (DOE) and stored in the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) data repository[4,23]. Integration and analysis of diverse data types such as hydrological, geological, ecological, biological, and climatological data is an essential element of complex environmental systems science (ESS) research. However, such interdisciplinary data integration presents unique challenges, such as inconsistent use of terms, formats, and metadata across disciplines[24]. In this manuscript, we describe and harmonize 11 diverse and complementary (meta)data reporting formats that our interdisciplinary team developed for commonly used data types in ESS research to enable their archival following FAIR principles in general purpose repositories such as ESS-DIVE. These include guidelines to format and describe general research elements (e.g., general file metadata, tabular data, physical samples, model data), as well as guidelines developed for more specific data types relevant to interdisciplinary research (e.g., biogeochemical samples, soil respiration, leaf-level gas exchange). As part of this process, we adopted or used components of existing reporting formats or standards to the greatest extent possible, and also developed new reporting formats for some data types. These can be used individually or collectively in scientific workflows, and many of the formats are widely applicable for environmental research. Moreover, the process we used for developing the formats—including our approach to obtain community consensus, mirror documentation across several web platforms, and track community feedback—can be used by other research communities to develop reporting formats for their own purposes.

2. Results

Our community-centric approach to developing reporting formats had four key outcomes that are broadly important to making scientific data more reusable. First, the teams reviewed a total of 112 pre-existing data standards and other data resources (data repositories, data systems, datasets, projects) to create (meta)data crosswalks (Supplementary Files 1-20). Such

crosswalks provide a tabular map of existing resources related to each data type, allowing the teams to identify gaps in existing standards, and determine which variables, terms, and metadata were essential to harmonize and incorporate into their reporting formats. At the onset of the review process, ESS-DIVE recommended adopting existing standards to the extent possible. However, we found that for all 11 data types, none entirely met ESS research community needs, and this necessitated development of all 11 reporting formats.

Second, we created 11 reporting formats (Table 1) that encompass a range of complex and diverse ESS (meta)data fields that can be used when researchers upload data to ESS-DIVE. Six of the reporting formats created by our community of scientists are cross-domain reporting formats (Figure 1a), which apply broadly to data across different scientific disciplines. These reporting formats were developed to help researchers more consistently format their (meta)data for interdisciplinary science applications and include basic dataset metadata for citation and findability[25], file-level metadata[26], guidelines for formatting Comma Separated Value (CSV) files[27], sample metadata[28], terrestrial model data archiving guidelines[29], and research locations metadata[30]. The remaining five reporting formats apply to different domain data types (Figure 1b) and include microbial amplicon abundance tables[31], leaf-level gas exchange[32], soil respiration[33], sample-based water and soil chemistry measurements[34], and water level and sonde-based hydrologic measurements[35]. All reporting formats have a minimal set of required metadata fields necessary for programmatic data parsing and optional fields that provide detailed spatial/temporal context about the sample useful to downstream scientific analyses. Throughout development, we aimed to strike a balance between pragmatism for the scientists reporting data and machine-actionability that is emblematic of FAIR data. A comparison between FAIR guiding principles and our reporting formats (Table 3) highlights how a community-centric effort like ours can move data archiving towards achieving many FAIR data principles (though see discussion for limitations).

Together, these 11 reporting formats are part of a flexible, modular, and integrated framework (Figure 1) that can accommodate new reporting formats in the future, and enable their findability and accessibility individually or collectively. As part of the framework development, all teams created templates with harmonized terms and formats to be internally consistent as much as possible. For example, dates are always reported in YYYY-MM-DD format. Whenever reporting formats include spatial data, the variables are harmonized as “latitude” and “longitude” and reported in decimal degrees with common bounds (-90 to 90 and -180 to 180, respectively). All formats that require CSV files adopted as many recommendations from the CSV reporting format as possible. Data collected using the water and soil chemistry,

and amplicon reporting formats have an option to report a persistent identifier for associated samples [International General Sample Number (IGSN)], to enable effective tracking across online data systems, as outlined in the Sample ID reporting format.

The third outcome is related to how we shared and archived all reporting formats in three ways, each with a distinct use. First, all reporting formats are published as datasets in the ESS-DIVE repository, which enables direct, public download and citation upon use. Second, each reporting format is hosted on the version control platform GitHub, which enables ongoing edits and versioning while also allowing users to provide feedback[36]. Third, the most up-to-date reporting format content from GitHub is rendered as a project website through the service GitBook[37]. We mirrored the reporting format instructions and templates across several web platforms to ensure the documentation is available in a variety of digital formats to serve the needs of various user groups and stakeholders. GitHub is likely a more familiar platform and user interface for software engineers and informatics specialists, for example, while GitBook websites may be preferred by Earth science researchers.

Lastly, we formulated guidelines (Box 1) for research communities that want to replicate our model of community-centric (meta)data reporting format development. We encourage (1) reviewing existing standards, (2) developing a crosswalk of terms across relevant standards or ontologies of interest, (3) iteratively developing templates and documentation with feedback from prospective users, (4) assembling a minimum set of (meta)data required for reuse, and (5) hosting finalized documentation on platforms that can be publicly accessed and updated easily.

3. Discussion

Many scientific journals and funders require data deposition in long-term repositories. However, in many cases, data are submitted to repositories in bespoke formats with little (meta)data standardization[5]. Community-led (meta)data reporting formats like the set described in this paper can enable archived data to be more reusable and interoperable[21,22]. Our scientist-centric approach to creating the formats helped to determine workflows that are most useful and practical for researchers to adopt. Here we discuss important aspects that need to be considered in development and use of such reporting formats.

Reporting formats can help researchers organize and synthesize their own (meta)data for their research purposes. It can be challenging for small teams, or even individuals to keep track of data collected over multi-year field campaigns or laboratory experiments[19,20]. Early adoption of a consistent way of compiling data can help individuals or research teams avoid ad

hoc data collection practices and also help researchers efficiently integrate their data, particularly when multiple analyses or teams are involved.

Moreover, community reporting formats can lead to greater data accessibility and reuse. For example, researchers in the Ameriflux network[38] organize flux data in the Flux Processing (FP-in) reporting format[18]. When participants in the network agree to provide their flux data in this format[39], benefits include: 1) access to data services such as automated QA/QC of datasets and value-added ONEFlux data processing[40], 2) Digital Object Identifier assignment which helps to track dataset citation and reuse, and 3) potential to increase findability of their data. Similarly, when contributors upload datasets on ESS-DIVE, they are offered automated metadata quality assessments, and published data are assigned DOIs and made searchable across the DataONE network. In another example, the Watershed Function Scientific Focus Area project[41] adopted ESS-DIVE's water and soil chemistry reporting format as an initial step towards establishing a field data workflow in a community observatory where diverse hydrological, geochemical, geophysical, ecological, and remote sensing datasets are collected[42]. The use of the reporting format will make it possible for researchers to synthesize data on chemical concentrations both within and across field locations.

Application of the reporting formats also allows for the use of tools and services that enhance data curation, findability and reuse. As an example, some of the fields in ESS-DIVE's dataset metadata reporting format[25] allow programmatic metadata quality validation, which checks for field presence, format, and length. Because these metadata can be mapped to a variety of machine-readable metadata formats including JSON-LD and the U.S. Department of Energy's Office of Scientific and Technical Information (OSTI) reporting formats[43]. This further enabled transforming and disseminating ESS-DIVE datasets across other platforms such as Google Dataset Search, DataONE, OSTI and DataCite.

The development of these reporting formats was driven by the scientific need for practical tools for data management, while improving the potential for data reuse achieving many of the FAIR guiding principles (Table 3). We made several pragmatic choices to ensure that the reporting formats would have a low barrier to adoption by time-limited researchers. This included investigating whether using pre-existing reporting formats "off the shelf" would meet project and researcher's scientific needs and workflows. Although it is desirable to use existing formats whenever possible, we found that there were many circumstances when they do not directly apply to a scientific community's research (meta)data needs. For example, although the Water Quality Exchange format[14] is used within the United States to report water quality monitoring data by local, state, and federal agencies, the format was not entirely suitable for

ESS-DIVE's purposes. Some of the concerns raised by the community included: 1) the structure of the data and metadata templates that are used for regulatory reporting were considered to be cumbersome and inefficient for scientific use (e.g., containing redundant elements of sampling and analytical methodology along with the data) and 2) the required vocabularies (as specified in the template dictionary) were found to be difficult to use because they included several terms that were unnecessary, while missing terms for specific analytes of interest to the community.

To address these concerns, we developed the ESS-DIVE reporting format for sample-based water and soil chemistry[34] that is more suitable for files typically generated in scientific laboratories. It borrows elements from the WQX standard, but provides flexibility in format and terminology, while capturing sufficient metadata and vocabularies to enable data exploration and reuse including the ability to use scripts to compare and combine different datasets[34]. In this way, the water and soil chemistry reporting format achieves some component of FAIR guiding principle "I2" that suggests using ontologies, while still being responsive to a research community that desired flexibility in research terminology (Table 3).

Similarly, when creating the sample ID metadata reporting format, we decided to extend the existing IGSN sample identifier template and guidelines in ESS-DIVE's Sample ID reporting format to meet researchers' need to link interdisciplinary environmental and biological samples, and to minimize effort in providing information for sample collections[44]. In this case, incorporating IGSNs ensures that researchers using this format achieve FAIR principle "F3" and have globally unique identifiers for their data products, which facilitates tracking associated sample data across multiple online data systems. In an effort to be pragmatic, we decided to lower the threshold for adoption of the sample ID reporting format (and nearly all others; Table 3) by compromising on elements that would achieve FAIR principle "I3" related to machine readable knowledge representation. All reporting formats encourage users to define variables in a data dictionary. Though this may not be fully machine readable according to the FAIR principles[21], this method of defining variables is a key step toward reusable and machine actionable data. The feedback gathered when creating our Sample ID reporting format was then provided to the broader IGSN community to help improve the IGSN metadata template for interdisciplinary science[45,46].

Through the process, we learned that many (meta)data standards are not accessible to a typical researcher and require a significant learning curve to become fluent in the informatics terminology used by established data standards. For example, the Open Geospatial Consortium's data standard for environmental sensors[13] is a detailed schema described over 100 pages, which is challenging for a typical scientific researcher to understand and implement.

Hence, we had to make several pragmatic choices to ensure that the reporting formats would be amenable to adoption by time-limited researchers. Once choice involved replacing terms in existing standards with words that were more intuitive to scientists. For example, whilst there was no reporting format for leaf-level gas exchange data, a crosswalk of the instrument output from a relatively small number of instrument manufacturers quickly identified a common terminology that already had broad acceptance and use by the scientific community (Supplementary File 7). By using crosswalks (Supplementary Files 1-10) our teams were able to map ESS-DIVE's reporting formats to many existing (meta)data standards and other data resources, and, in the future, will allow building tools that enable interoperability with different systems. We also simplified the reporting format templates and instructions to the greatest extent possible by specifying a few required fields and several more optional fields to provide additional details.

Our model and guidelines of supporting and empowering the scientific community to develop (meta)data reporting formats that meet their needs can enable other communities to undertake these internal data standardization efforts that make their data even more useful beyond the purpose for which they were collected (Box 1). We acknowledge that other research infrastructures have made important strides toward data standardization within research communities though they can still take dozens of years to manifest[17]. We found value in including a broad range of stakeholders in the process, and included field personnel who make the measurements, instrument manufacturers, and scientists who use the data in models or synthesis activities[47].

There are incentives that can help promote widespread adoption of these or other formats to justify the time investment required for individual researchers or teams into scientific workflows. First, involving data collectors and reusers at the core of the development process makes the resulting formats more pragmatic and scientifically useful. Importantly, the domain scientists involved in the reporting format development became community ambassadors and helped engage their use by fellow researchers through conference presentations and peer-reviewed papers[44,47–49]. Second, we expanded our user community by sharing information about the reporting formats through a series of webinars, documentation, tutorials, and personalized community outreach. These incentives have had some success, as evidenced by the datasets submitted to ESS-DIVE using one or more of the reporting formats within a few months after they were finalized (Table 2).

We identify some future work that can potentially lower the barrier to adopting reporting formats, provide added benefits to those who use the formats, and make (meta)data FAIRer[10].

Currently, ESS-DIVE applies a set of manual checks to datasets uploaded to ESS-DIVE that follow the reporting format. However, development of automated formatting checkers[50] would help users instantly validate their datasets against reporting format guidelines. Other types of software can also be built around the reporting formats. For example, software could be developed to automatically convert sensor or instrument-derived data into the units requested by a reporting format. As a starting point for this work, the file-level metadata reporting format already includes an open-source script[51] that enables reading and parsing data files submitted in that format. The leaf-level gas exchange reporting format includes a detailed translation table matching the reporting format data variables with standard outputs from 10 commonly used, commercially produced instruments. This could provide the foundation for development of conversion software to automatically format data with the recommended variable names and units. ESS-DIVE is also planning a data integration and fusion component of the repository that will facilitate synthesizing and analyzing datasets that adhere to any of the 11 ESS-DIVE reporting formats. Enabling advanced queries within the files will require development of software and data parsers so that a great number of reporting formats achieve FAIR principle "F4" which calls for data to be fully searchable.

With more data being generated than ever, reusable data can have substantial societal, economic, and scientific impacts. But for Earth and environmental science data, which are complex and heterogeneous, achieving reusability will require concentrated effort at (meta)data standardization within research communities. Our work to develop 11 community (meta)data reporting formats is a critical step to making Earth and environmental science data more reusable because we emphasize human readability that is compatible with machine readability. We hope that our model of empowering research communities to self-organize and create their own (meta)data reporting formats will enable other communities to undertake these internal data standardization efforts that make their data even more useful beyond the purpose for which they were collected.

4. Methods

Earth and environmental science data are complex, multi-scale, and span diverse research domains such as geology, hydrology, climate, ecology, and biology. At ESS-DIVE, we initiated a community-centric model that engaged domain scientists to develop formats for common Earth science data types. The objective was to create formatting guidelines and templates that would gather the minimum but sufficient metadata or data necessary for data interpretation and reuse.

4.1 Reviewing existing standards and feedback on drafts

Each team conducted a review of existing standards (Table 1), involving both literature searches and exploring resources from informatics groups (e.g., Research Data Alliance and Earth Science Information Partners) or agencies working with similar data, to identify whether any existing data standards or conventions could be used directly or to inform their reporting format. Based on this review, each team created tabular 'crosswalks' (Supplementary Files 1-10) to map related terminology from relevant standards. This process helped identify gaps in existing standards, and determine important elements that had to be present, and variations in terminology used across different standards that required harmonization. For example, some existing standards report date and time under the column name 'datetime' while another reports the same information, as 'ValueDateTime' (see example of a terminology crosswalk[35]). Here, we provide a brief narrative of methods for each reporting format with details on existing data standards and other data resources reviewed during reporting format development. For further details on the technical aspects of each reporting format, please refer to ESS-DIVE's community space on GitHub[36] or view the datasets for each reporting format submitted to ESS-DIVE (Table 1).

4.2 Obtaining community consensus

Each team created instructions and (meta)data templates for their reporting formats. The teams piloted the formats within their research groups and communities to ensure the templates were practical and useful for scientists who collect and reuse data (Figure 2). In total, 247 individuals representing 128 institutions provided input at various stages of the reporting format development process. As the reporting format instructions and templates reached a final stage, they published the "ready-to-use" reporting formats in three locations each with distinct benefits for the end-users: GitHub[37], GitBook, and the ESS-DIVE data repository to enable findability and long-term preservation.

4.3 Cross-domain reporting formats

4.3.1 Dataset metadata

The goal for creating the dataset metadata reporting format was to ensure that any dataset submitted to ESS-DIVE would have complete and descriptive metadata to enable its findability and citation upon use. The ESS-DIVE team reviewed machine and human-readable metadata standards including the Ecological Metadata Language[52] as well as JSON for Linking Data[53].

The ESS-DIVE metadata reporting format follows existing metadata standards as much as possible (e.g., 'title' in Ecological Metadata Language is also 'title' for ESS-DIVE's metadata).

4.3.2 File-level metadata

The file-level metadata reporting format was developed for users to provide details about the individual files contained within a dataset. The review of existing standards[26] included file-level metadata used across 6 organizations (e.g., USGS, NEON).

4.3.3 CSV file formatting guidelines

The CSV reporting format was developed to provide guidelines for more consistently formatting tabular data[27]. The intention was to make this a domain agnostic set of guidelines so that anyone who works with tabular data can use the format in their research to make tabular data more interoperable and machine-readable. The team reviewed existing standards and guidelines (Table 1) including recommendations from the Environmental Data Initiative (e.g., do not mix data types in a column) and the ORNL DAAC (e.g., indicating missing numeric values with -9999).

4.3.4 Sample IDs and metadata

The ESS-DIVE Sample ID reporting format[28] aligns as much as possible with extensive work on IGSN[54] with the goal of standardizing sample collection metadata and more efficiently tracking physical samples sent to different collaborators, labs, data systems, etc. This work also reviewed 12 different standards and data resources to provide recommendations for improving interoperability of biological[8,55] and environmental samples[14].

4.3.5 Terrestrial model data archiving guidelines

The model data archiving reporting format[29] was informed by input from the DOE's land modeling community and other guidelines from the American Geophysical Union and National Science Foundation Earthcube communities. In developing the guidelines[49], the goal was to help modelers make decisions about which components of their terrestrial models should be archived in a long-term data repository. The guidelines were developed with input on which model data were most useful to archive, how long they remained useful, and what scientific purpose they would serve.

4.3.6 Location metadata

The goal of developing the location metadata reporting format was to provide generalized guidelines for describing locations used in research. The review of existing standards included metadata templates from specific projects at some of the DOE's National Labs to understand the different field sampling strategies of large interdisciplinary projects. The review also included known standards and guidelines for recording locations such as Climate and Forecast Conventions[56], the Federal Geographic Data Committee's Content Standard for Digital Geospatial Metadata[57] and the Open Geospatial Consortium[58].

4.4 Reporting formats for domain-specific data types

In addition to the set of 6 cross-domain reporting formats described above, we also developed 5 formats that are tailored to specific data types commonly used in the terrestrial and subsurface ecosystem research community. ESS-DIVE's goal was to engage Earth and environmental scientists to determine practical reporting formats that data contributors are willing to use while at the same time ensuring a high potential for data reuse.

4.4.1 Amplicon abundance table metadata

The reporting format for amplicon abundance table metadata was developed to facilitate consistent reporting of microbiome sample data with the format of these tables following ESS-DIVE's CSV file guidelines. Required data (e.g., representative sequences) were chosen to support comparisons of abundance tables across studies. The reporting format distinguishes between sequencing metadata and bioinformatic processing metadata for amplicon abundance tables. As much as possible, the team aligned recommendations for sequencing metadata with the existing standards developed by the Genomic Standards Consortium for minimum information about a marker gene sequence and minimum information about any (x) sequence[55] (Supplementary File 6). In the absence of an existing standard for bioinformatic processing metadata, the reporting format contains a minimal set of fields to capture the data processing steps most relevant for comparing and combining amplicon counts across studies (Table 1). The final set of sequencing and bioinformatic metadata fields selected were informed by a community of scientists involved with either the development of microbiome data pipelines or conducting microbiome studies in both field and lab settings.

4.4.2 Leaf-level gas exchange

The team working on this reporting format[32] reviewed existing conventions used in plant trait databases, large data collections developed for synthesis papers, and the variable descriptions

that are part of standard instrument outputs in order to determine the most suitable variable names to use to report leaf-level gas exchange data. Templates for formatting metadata about the methods and sample materials used in an experiment, as well as details on the instrumentation involved in collecting data were developed through an iterative process of input and review open to all interested stakeholders. The reporting format is designed to be flexible and modular, provides guidelines on the archive of raw and processed data, and seeks to capture experimental metadata needed to interpret and reuse these data types[47].

4.4.3 Soil respiration

To create the soil respiration reporting format, this team reviewed and integrated recommendations from 9 existing guidelines and standards (Table 1)[33]. The review captured an array of how different standards format their general metadata and data (e.g., formatting date and timestamps) and also accounted for a range of soil-atmosphere gas exchange data types (e.g., GHGs or radiocarbon)[48].

4.4.4 Sample-based water and soil chemistry measurements

The goal in creating a reporting format for water-soil-sediment data was to harmonize chemical concentration data that span several measurement types. The review included 15 standards (Table 1) for related environmental chemistry measurements including metadata elements from the EPA's WQX[14] as well as EarthChem[59]. Based on input from the potential ESS user community that included both data collectors, managers, and modelers, we developed a reporting format based on community input[34].

4.4.5 Water level and sonde-based hydrologic monitoring

This reporting format harmonizes variables common to sonde-based hydrologic monitoring research including water level, temperature, and pH data. The existing standards and/or data sources included in the crosswalk for the hydrologic monitoring reporting format (Table 1) were chosen for inclusion given their common use in the scientific community. They aligned generally on the types of hydrologic metadata to record (e.g., information about dates and times as well as information about data collection sites) but had different terminology across each of the resources[35]. The development of the reporting format included a review of additional data sources and standards beyond those listed in the crosswalk (Table 1).

Acknowledgements

Robert Crystal-Ornelas, Charuleka Varadharajan, Dylan O'Ryan, Madison Burrus, Shreyas Cholia, Joan Damerow, Valerie C. Hendrix, Zarine Kakalia, Fianna O'Brien, Emily Robles, Maegen Simmonds, Karen Whitenack, and Deborah A. Agarwal were funded through the ESS-DIVE repository by the U.S. DOE's Office of Science Biological and Environmental Research under contract number DE-AC02-05CH11231. Kim S. Ely and Alistair Rogers were supported through the US Department of Energy contract number DE-SC0012704 to Brookhaven National Laboratory. We acknowledge the work of Diana Swantek in producing the Figure 2 illustration. Reporting format development was supported by ESS-DIVE's Community Funds, through the Office of Biological and Environmental Research in the Department of Energy, Office of Science.

Competing interests

The authors of this manuscript have no competing interests to declare.

Data availability

Each data reporting format and all supporting documentation are hosted on our GitHub Community Space[36] and archived in the ESS-DIVE data repository[25–35]. The supplementary information for this manuscript is also archived in ESS-DIVE[60].

Code availability

We have made code available which enables file-level metadata extraction[51] for files that adhere to the reporting format.

Author contributions (according to CRediT taxonomy)

Conceptualization: DAA, CV

Data curation: RCO, CV, DO, KB, BBL, KB, MC, JD, KSE, AEG, SH, KM, SP, AR, MS, TV, PW, JNW, DAA

Funding Acquisition: DAA, CV

Methodology: DAA, CV, RCO, JED, KB, BBL, KB, MC, JD, KSE, AEG, SH, KM, SP, AR, MS, TV, PW, JNW

Project Administration: DAA, KW

Resources: DAA

Software: MC

Supervision: DAA, CV

Visualization: RCO, CV

Writing - original draft: RCO, CV, JED

Writing - review and editing: RCO, CV, DO, KB, BBL, KB, MB, SC, DSC, MC, JD, KSE, AEG, SH, VCH, ZK, KM, FO, SP, ER, AR, MS, TV, PW, JWN, KW, DAA

References

1. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
2. Stall, S. *et al.* Advancing FAIR data in Earth, space, and environmental science. *Eos, Earth and Space Science News* **99**, (2018).
3. Toelch, U. & Ostwald, D. Digital open science-Teaching digital tools for reproducible and transparent research. *PLoS Biol.* **16**, e2006022 (2018).
4. Varadharajan, C. *et al.* Launching an Accessible Archive of Environmental Data. *EOS* **100**, (2019).
5. Tenopir, C. *et al.* Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide. *PLoS One* **15**, e0229003 (2020).
6. Perrier, L., Blondal, E. & MacDonald, H. The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis. *PLoS One* **15**, e0229182 (2020).
7. Sansone, S. A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* **37**, 358–367 (2019).
8. Wieczorek, J. *et al.* Darwin Core: an evolving community-developed biodiversity data standard. *PLoS One* **7**, e29715 (2012).
9. Poisot, T., Bruneau, A., Gonzalez, A., Gravel, D. & Peres-Neto, P. Ecological Data Should Not Be So Hard to Find and Reuse. *Trends Ecol. Evol.* **34**, (2019).
10. Michener, W. K. Ecological data sharing. *Ecol. Inform.* **29**, 33–44 (2015).
11. Lin, D. *et al.* The TRUST Principles for digital repositories. *Sci Data* **7**, 144 (2020).
12. ISO. Date and Time Format (ISO Standard Number 8601-1:2019). (2019).
13. Bröring, A., Stasch, C. & Echterhoff, J. OGC Sensor Observation Service Interface Standard, Version 2.0. (2012).

14. Read, E. K. *et al.* Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resour. Res.* **53**, 1735–1745 (2017).
15. AmeriFlux. *BADM: Biological, Ancillary, Disturbance, and Metadata* https://ameriflux.lbl.gov/data/badm/ (2020).
16. Dañobeitia, J. J. *et al.* Toward a Comprehensive and Integrated Strategy of the European Marine Research Infrastructures for Ocean Observations. *Frontiers in Marine Science* **7**, (2020).
17. Cocco, M. *et al.* The EPOS Research Infrastructure: a federated approach to integrate solid Earth science data and services. *Ann. Geophys.* **65**, DM208–DM208 (2022).
18. Flux Processing (FP-In). *Half-Hourly / Hourly Data Upload Format* https://ameriflux.lbl.gov/half-hourly-hourly-data-upload-format/ (2017).
19. Goodman, A. *et al.* Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput. Biol.* **10**, e1003542 (2014).
20. Lowndes, J. S. S. *et al.* Our path to better science in less time using open data science tools. *Nat Ecol Evol* **1**, 0160 (2017).
21. Jacobsen, A. *et al.* FAIR principles: Interpretations and implementation considerations. *Data Intellegence* **2**, 10–29 (2020).
22. Bailo, D. *et al.* Perspectives on the Implementation of FAIR Principles in Solid Earth Research Infrastructures. *Front Earth Sci. Chin.* **8**, (2020).
23. Environmental System Science Data Infrastructure for a Virtual Ecosystem. *ESS-DIVE* https://data.ess-dive.lbl.gov/data.
24. Hills, D. J. *et al.* Earth and space science informatics perspectives on integrated, coordinated, open, networked (ICON) science. *Earth Space Sci.* **9**, (2022).
25. Agarwal, D. *et al.* ESS-DIVE Reporting Format for Dataset Package Metadata. *ESS-DIVE* https://www.doi.org/10.15485/1866026 (2022).
26. Velliquette, T. *et al.* ESS-DIVE Reporting Format for File-level Metadata. *ESS-DIVE*

https://www.doi.org/10.15485/1734840 (2021).

27. Velliquette, T. *et al.* ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure. *ESS-DIVE* https://www.doi.org/10.15485/1734841 (2021).
28. Damerow, J. *et al.* ESS-DIVE global sample numbers and metadata reporting format for Environmental Systems Science (IGSN-ESS). *ESS-DIVE* https://www.doi.org/10.15485/1660470 (2020).
29. Simmonds, M. B. *et al.* ESS-DIVE guidelines for archiving terrestrial model data. *ESS-DIVE* https://www.doi.org/10.15485/1813868 (2021).
30. Crystal-Ornelas, R. *et al.* ESS-DIVE Reporting Format for Location Metadata. *ESS-DIVE* https://www.doi.org/10.15485/1865730 (2022).
31. Weisenhorn, P. & Beilsmith, K. ESS-DIVE Reporting Format for Amplicon Abundance Tables. *ESS-DIVE* https://www.doi.org/10.15485/1865729 (2022).
32. Ely, K. S., Rogers, A. & Crystal-Ornelas, R. ESS-DIVE reporting format for leaf-level gas exchange data and metadata. *ESS-DIVE* https://www.doi.org/10.15485/1659484 (2020).
33. Bond-Lamberty, B., Christianson, D. S., Crystal-Ornelas, R., Mathes, K. & Pennington, S. C. ESS-DIVE reporting format for field measurements of soil respiration. *ESS-DIVE* https://www.doi.org/10.15485/1798520 (2021).
34. Boye, K. *et al.* ESS-DIVE Reporting Format for Sample-based Water and Soil Chemistry Measurements. *ESS-DIVE* https://www.doi.org/10.15485/1865731 (2022).
35. Goldman, A. E., Ren, H., Torgeson, J. & Zhou, H. ESS-DIVE Reporting Format for Hydrologic Monitoring Data and Metadata. *ESS-DIVE* https://www.doi.org/10.15485/1822940 (2021).
36. ESS-DIVE Community Space. *ESS-DIVE Community Space* https://github.com/ess-dive-community (2021).
37. Crystal-Ornelas, R. *et al.* A guide to using GitHub for developing and versioning data standards and reporting formats. *Earth Space Sci.* **8**, (2021).

38. Novick, K. A. *et al.* The AmeriFlux network: A coalition of the willing. *Agric. For. Meteorol.* **249**, 444–456 (2018).
39. AmeriFlux Data Policy. *AmeriFlux Data Policy* https://ameriflux.lbl.gov/data/data-policy/ (2021).
40. Onboarding and Orientation for new site teams. https://ameriflux.lbl.gov/sites/onboarding-and-orientation-for-new-site-teams/ (2017).
41. Hubbard, S. S. *et al.* The East River, Colorado, watershed: A mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics. *Vadose Zone J.* **17**, 1–25 (2018).
42. Kakalia, Z. *et al.* The Colorado East River community observatory data collection. *Hydrol. Process.* **35**, (2021).
43. OSTI. *Instructions for announcement of U.S. Department of Energy (DOE) publicly available scientific research datasets* https://www.osti.gov/elink/F2416instruct.jsp (2017).
44. Damerow, J. *et al.* Sample identifiers and metadata to support data management and reuse in multidisciplinary ecosystem sciences. *Data Science Journal* **20**, 1–19 (2021).
45. Klump, J. *et al.* Towards globally unique identification of physical samples: Governance and technical implementation of the IGSN global sample number. *Data Sci. J.* **20**, (2021).
46. Richard, S. M. *et al.* Internet of samples. *Proc. Assoc. Inf. Sci. Technol.* **58**, 813–815 (2021).
47. Ely, K. S. *et al.* A reporting format for leaf-level gas exchange data and metadata. *Ecol. Inform.* (2021).
48. Bond-Lamberty, B., Christianson, D. S., Crystal-Ornelas, R., Mathes, K. & Pennington, S. C. A reporting format for field measurements of soil respiration. *Ecol. Inform.* (2021).
49. Simmonds, M. B. *et al.* Guidelines for Publicly Archiving Terrestrial Model Data to Enhance Usability, Intercomparison, and Synthesis. *Data Sci. J.* **21**, (2022).
50. Fowler, D., Barratt, J. & Walsh, P. Frictionless data: Making research data quality visible.

Int. J. Digit. Curation **12**, 274–285 (2018).

51. McNelis, J., Crow, M. & Devarakonda, R. ESS-DIVE File Level Metadata Extractor. *DOE Code* https://www.doi.org/10.11578/DC.20201103.5 (2020).
52. Jones, M. *et al.* Ecological Metadata Language version 2.2.0. *KNB Data Repository* https://www.doi.org/10.5063/F11834T2 (2019).
53. Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. & Lindström, N. JSON-LD 1.0. *W3C recommendation* **16**, 41 (2014).
54. Lehnert, K. A., Klump, J., Wyborn, L. & Ramdeen, S. IGSN: Trustworthy and Sustainable Services for FAIR Samples. in *Geophysical Research Abstracts* vol. 21 (2019).
55. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, 415–420 (2011).
56. CF Conventions and Metadata. https://cfconventions.org/index.html (2020).
57. Federal Geographic Data Committee. *Content Standard for Digital Geospatial Metadata*. https://ci.nii.ac.jp/naid/10016800076/ (1998).
58. Observations and Measurements. *Observations and Measurements* https://www.ogc.org/standards/om (2011).
59. Walker, J. D., Lehnert, K. A., Hofmann, A. W., Sarbas, B. & Carlson, R. W. EarthChem: International Collaboration for Solid Earth Geochemistry in Geoinformatics. in vol. 2005 IN44A-03 (2005).
60. Crystal-Ornelas, R. *et al.* Data from: “Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats.” *ESS-DIVE* https://www.doi.org/10.15485/1866606 (2022).
61. JSON for Linking Data. *JSON for Linking Data* https://json-ld.org/ (2016).
62. Brase, J. DataCite - A Global Registration Agency for Research Data. in *2009 Fourth International Conference on Cooperation and Promotion of Information Resources in*

Science and Technology 257–261 (2009).

63. Evans, K. *et al. ASCII File Format Guidelines for Earth Science Data*. (2016).
64. Federal Geographic Data Committee. *National Geospatial Data Assets (NGDA) Metadata Guidelines*. (2016).
65. Shafranovich, Y. *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. (2005).
66. USGS Data Dictionaries. https://www.usgs.gov/data-management/data-dictionaries (2020).
67. EDI. *Five phases of data publishing - Phase 2: Format and QC data tables* https://environmentaldatainitiative.org/five-phases-of-data-publishing/phase-2/ (2019).
68. Pepler, S. & Parton, G. BADC-CSV Format for Data Exchange. https://help.ceda.ac.uk/article/105-badc-csv (2009).
69. Hsu, L. How “clean” should an Excel file be to be considered machine readable. https://my.usgs.gov/confluence/pages/viewpage.action?pageId=559852026 (2016).
70. NEON. NEON file naming conventions. https://data.neonscience.org/file-naming-conventions (2022).
71. Tarboton, D. G., Horsburgh, J. S. & Maidment, D. R. CUAHSI community Observations Data Model (ODM) version 1.1 design specifications. *Des Doc* (2008).
72. StreamPulse uploading data. http://pulseofstreams.weebly.com/uploading_data.html (2022).
73. Cerf, V. *ASCII format for Network Interchange*. (1969).
74. Newell, D. B. & Tiesinga, E. *The International System of Units (SI) (2019 Edition)*. (National Institute of Standards and Technology, 2019).
75. EPSG. *WGS 84*. (1984).
76. ORNL DAAC CSV Standards. https://daac.ornl.gov/submit/csvstandards/ (2018).
77. Data Quality Review Checklist. *https://daac.ornl.gov/submit/qachecklist/* https://daac.ornl.gov/submit/qachecklist/ (2022).

78. USGS Data Templates. https://www.usgs.gov/products/data-and-tools/data-management/data-templates (2022).
79. National Archives. https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#structuredata (2022).
80. Klyne, G. & Newman, C. *Date and Time on the Internet: Timestamps*. (2002).
81. Loescher, H. AmeriFlux BASE CR-Lse La Selva. *AmeriFlux* https://www.doi.org/10.17190/AMF/1246013 (2016).
82. Torn, M. & Dengel, S. AmeriFlux US-NGB NGEE Barrow. *AmeriFlux* https://www.doi.org/10.17190/AMF/1436326 (2018).
83. Brokaw, N. Luquillo Forest Dynamics Plot (LFDP) Liana Data. *Environmental Data Initiative* http://www.doi.org/10.6073/PASTA/EABE6E15723324EA3938B456D5BB02C2 (2017).
84. Bueno de Mesquita, C. P. Plant colonization of moss-dominated soils in the alpine: Microbial and biogeochemical implications. *Environmental Data Initiative* https://www..doi.org/10.6073/PASTA/C0CACD100CD89DA258B40A77FBB2FD4C (2019).
85. Brooks, S. East Fork Poplar Creek sonde data at kilometer 16.2 water year 2019. *Mercury Critical Interfaces SFA Data Search* https://www.doi.org/10.12769/1569821 (2019).
86. Riscassi, A. & Brooks, S. East Fork Poplar Creek discharge at kilometer 5.4 water year 2012. *Mercury Critical Interfaces SFA Data Search* https://www.doi.org/10.12769/1489524 (2019).
87. National Ecological Observatory Network (NEON). Coarse downed wood bulk density sampling (DP1.10014.001). *NEON* https://www.doi.org/10.48443/Z0TG-QS14 (2020).
88. Salmon, V., Iversen, C., Childs, J. & VanderStel, H. NGEE arctic plant traits: Soil cores, Kougarok road mile marker 64, Seward peninsula, Alaska, 2016. *NGEE Arctic Data Search* https://www.doi.org/10.5440/1346200 (2019).
89. Philben, M. *et al.* Results of experimental additions of organic nitrogen on soil organic matter decomposition, teller road site, Seward peninsula, 2017 and 2018. *NGEE Arctic*

Data Search https://www.doi.org/10.5440/1454263 (2019).

90. Yaffar, D., Lugo, A. E., Silver, W., Cuevas, E. & Molina Colon, S. Plant root trait measurements raw data, 1962-2018, Island of Puerto Rico. *Next-Generation Ecosystem Experiments Tropics; Oak Ridge National Laboratory* https://www.doi.org/10.15486/NGT/1558773 (2019).
91. Norby, R. *et al.* Root-soil depth profile in Luquillo Experimental Forest, Puerto Rico, February, 2019. *Next-Generation Ecosystem Experiments Tropics; Oak Ridge National Laboratory* https://www.doi.org/10.15486/NGT/1574087 (2019).
92. Griffiths, N. & Sebestyen, S. SPRUCE porewater chemistry data for experimental plots beginning in 2013. *Oak Ridge National Lab's Terrestrial Ecosystem Science Scientific Focus Area (ORNL TES SFA)* https://www.doi.org/10.3334/CDIAC/SPRUCE.028 (2016).
93. McPartland, M. Y. *et al.* SPRUCE: NDVI data from selected SPRUCE experimental plots, 2016-2018. *Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States)* https://www.doi.org/10.25581/SPRUCE.057/1490190 (2019).
94. Croteau, M. N., Sikder, M., Poulin, B. A. & Baalousha, M. Laboratory data to assess the effect of nanoparticle size and natural organic matter composition on the bioavailability of platinum nanoparticles to a model freshwater invertebrate species. *U.S. Geological Survey* https://www.doi.org/10.5066/P9G18URX (2020).
95. Danczak, R. E. *et al.* WHONDRS 48 Hour Diel Cycling Study at the Altamaha River in Georgia, USA. *Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS); ESS-DIVE* https://www.doi.org/10.15485/1577263 (2019).
96. Stegen, J. C. *et al.* WHONDRS 48 hour Diel cycling study at HJ Andrews Experimental Forest Watershed 1 (WS1). *Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS); ESS-DIVE* http://www.doi.org/10.15485/1509695 (2019).
97. SESAR. SESAR Batch Registration Quick Guide. *Zenodo*

https://www.doi.org/10.5281/zenodo.3874923 (2020).

98. IGSN Descriptive Metadata Schema. *IGSN metadata* https://github.com/IGSN/metadata (2017).

99. Weibel, S., Kunze, J., Lagoze, C. & Wolf, M. Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC* **2413**, 132 (1998).

100. ISO 19156:2011. *Geographic information — Observations and measurements* https://www.iso.org/standard/32574.html (2011).

101. Joint Genome Institute Genome Online Database. https://genome.jgi.doe.gov/portal/ (2022).

102. USGS National Digital Catalog. https://www.usgs.gov/programs/national-geological-and-geophysical-data-preservation-program/national-digital-catalog (2022).

103. Geologic Materials Repository Working Group. *The U.S. Geological Survey Geologic Collections Management System (GCMS)—A master catalog and collections management plan for U.S. Geological Survey geologic samples and sample collections*. (2015).

104. NEON Biorepository Data Portal. https://biorepo.neonscience.org/portal/ (2022).

105. Hanson, B. Data policies and practices for AGU publications for models and model output. (2020).

106. Williams, D. N., Lawrence, B. N., Lautenschlager, M., Middleton, D. & Balaji, V. The Earth System Grid Federation: Delivering globally accessible petascale data for CMIP5. *Proceedings of the Asia-Pacific Advanced Network* vol. 32 121 (2011).

107. Dryad. https://datadryad.org/ (2022).

108. Zenodo. https://zenodo.org/ (2022).

109. DAAC. https://earthdata.nasa.gov/eosdis/daacs (2021).

110. NCAR. https://rda.ucar.edu/ (2022).

111. EOL. https://www.eol.ucar.edu/about-eol (2022).

112. Arctic Data Center. https://arcticdata.io/ (2016).

113. Federal Geographic Data Committee. Content Standard for Digital Geospatial Metadata Part 1: Biological Data Profile. (1999).
114. Christianson, D. S. *et al.* A metadata reporting framework (FRAMES) for synthesis of ecohydrological observations. *Ecol. Inform.* **42**, 148–158 (2017).
115. USGS. https://dashboard.waterdata.usgs.gov/ (2022).
116. SNOTEL. https://wcc.sc.egov.usda.gov/reportGenerator/ (2022).
117. Earth Microbiome Project. https://earthmicrobiome.org/protocols-and-standards/metadata-guide/ (2022).
118. National Center for Biotechnology Information. SRA Metadata and Submission Overview. https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/ (2019).
119. Kattge, J. *et al.* TRY plant trait database - enhanced coverage and open access. *Glob. Chang. Biol.* **26**, 119–188 (2020).
120. LeBauer, D. *et al.* BETYdb: a yield, trait, and ecosystem service database applied to second generation bioenergy feedstock production. *Glob. Change Biol. Bioenergy* **10**, 61–71 (2018).
121. Maitner, B. S., Boyle, B., Casler, N. & Condit, R. The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods Ecol. Evol.* (2018).
122. ICOS. https://www.icos-cp.eu/ (2022).
123. Pastorello, G. *et al.* The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci Data* **7**, 225 (2020).
124. Dorich, C. D. *et al.* Improving N2O emission estimates with the global N2O database. *Curr Opin Environ Sustain* **47**, 13–20 (2020).
125. Lawrence, C. R. *et al.* An open-source database for the synthesis of soil radiocarbon data: International Soil Radiocarbon Database (ISRaD) version 1.0. *Earth Syst. Sci. Data* **12**, 61–76 (2020).
126. Schädel, C. *et al.* Decomposability of soil organic matter over time: the Soil Incubation

Database (SIDb, version 1.0) and guidance for incubation procedures. *Earth Syst. Sci. Data* **12**, 1511–1524 (2020).

127. Jian, J. *et al.* A restructured and updated global soil respiration database (SRDB-V5). *Earth Syst. Sci. Data* **13**, 255–267 (2021).

128. Ojima, D., Mosier, A., Del Grosso, S. & Parton, W. J. TRAGNET analysis and synthesis of trace gas fluxes. *Global Biogeochem. Cycles* **14**, 995–997 (2000).

129. Hibbard, K. A., Law, B. E., Reichstein, M. & Sulzman, J. An analysis of soil respiration across northern hemisphere temperate ecosystems. *Biogeochemistry* **73**, 29–70 (2005).

130. Horsburgh, J. S. *et al.* Observations Data Model 2: A community information model for spatially discrete Earth observations. *Environ. Model. Soft.* **79**, 55–74 (2016).

131. NEON Protocols & Standardized Methods. https://www.neonscience.org/data-collection/protocols-standardized-methods (2020).

132. EarthChem Data Templates. https://www.earthchem.org/ecl/templates/ (2021).

133. SLAC. https://www-ssrl.slac.stanford.edu/sfa/ (2022).

134. WFSFA. https://eesa.lbl.gov/projects/watershed-function-sfa/ (2022).

135. River Corridor SFA. https://www.pnnl.gov/projects/river-corridor (2022).

136. AWHSFA. https://www.anl.gov/bio/subsurface-biogeochemical-research (2022).

137. LLNL Seaborg. https://seaborg.llnl.gov/research/environmental-radiochemistry (2022).

138. Mercury SFA. https://www.esd.ornl.gov/programs/rsfa/ (2021).

139. WHONDRS. https://www.pnnl.gov/projects/WHONDRS (2022).

140. Ameriflux. https://ameriflux.lbl.gov/ (2022).

141. NEON. *https://www.neonscience.org/* (2020).

142. Vogel, T. M. *et al.* TerraGenome: a consortium for the sequencing of a soil metagenome. *Nat. Rev. Microbiol.* **7**, 252–252 (2009).

143. CUAHSI-HIS. *Master Controlled Vocabulary Registry for ODM 1.1* http://his.cuahsi.org/mastercvreg/cv11.aspx (2008).

144. Blodgett, D., Lucido, J. & Kreft, J. Progress on water data integration and distribution: a summary of select US Geological Survey data systems. *J. hydroinformatics* **18**, 226–237 (2016).
145. NWIS Inventory. https://waterdata.usgs.gov/nwis/inventory (2022).
146. NEON. https://www.neonscience.org/data-samples/data-management/data-formats-conventions (2020).
147. WQP User Guide. https://www.waterqualitydata.us/portal_userguide/ (2022).
148. Garayburu-Caruso, V. A. *et al.* FTICR, NPOC, TN, and moisture of variably inundated sediment across 48 north American rivers. *ESS-DIVE* https://www.doi.org/10.15485/1834208 (2021).
149. Alves, R. J. E. *et al.* Kinetic and temperature sensitivity properties of soil exoenzymes through the soil profile down to one-meter depth at a temperate coniferous forest (Blodgett, CA). *ESS-DIVE* https://www.doi.org/10.15485/1830417 (2021).
150. Rogers, A., Ely, K. & Serbin, S. *Leaf Photosynthetic Parameters: Quantum Yield, Convexity, Respiration, Gross CO2 Assimilation Rate and Raw Gas Exchange Data, Utqiagvik (Barrow), Alaska, 2016*. *NGEE Arctic Data Search* https://www.osti.gov/biblio/1482338 (2021).
151. Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A. & Stegen, J. C. Perceived costs and benefits of ICON science and foundational documents associated with “Integrated, Coordinated, Open, and Networked (ICON) science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles". (2022).
152. Roebuck, A. *et al.* FTICR-MS, Sensor, and Environmental Data from 5 Streams Impacted by the 2020 Holiday Farm Fire Associated with: “Spatiotemporal controls on the delivery of dissolved organic matter to streams following a wildfire.” *ESS-DIVE* https://www.doi.org/10.15485/1869708 (2022).

153. Allison, S. & Martiny, J. B. H. Fungal and bacterial growth variation due to drought and nitrogen addition experimental treatments. Loma Ridge Experimental Project. 2010-2012. *ESS-DIVE* https://www.doi.org/10.15485/1828589 (2021).
154. Dove, N., Torn, M., Hart, S. & Tas, N. Chemistry data from soils and soil incubation experiments from the whole-soil warming experiment at Blodgett Forest, CA, 2018, from: "Metabolic capabilities mute positive response to direct and indirect impacts of warming throughout the soil profile." *ESS-DIVE* https://www.doi.org/10.15485/1866269 (2022).
155. ESS-DIVE. *ESS-DIVE* https://ess-dive.lbl.gov/ (2022).

Tables and Figures

Table 1

A list of all 11 ESS-DIVE (meta)data reporting formats and citations for associated publications. Each row includes a description of the reporting format, a list of existing data standards as well as other resources (i.e., data repositories, data systems, datasets, data files from DOE projects, instruments) consulted during development of each reporting format.

Reporting format	Description	Standards reviewed	Other resources reviewed
Dataset metadata[25]	The dataset metadata reporting format is in use as part of ESS-DIVE's data submission process. It ensures that data submitted to ESS-DIVE has a minimum set of metadata (e.g., title, abstract, author contact information) to enable data reuse.	● Ecological metadata language[52] ● JavaScript Object Notation for Linked Data[61] ● DataCite 4.1[62] ● Federal Geographic Data Committee[57]	● Office of Science and Technical Information Announcement Notice 241.6[43]
File-level metadata[26,51]	Metadata about the data files included as part of a larger dataset. Information about the data files and the spatiotemporal context for each data file. Includes script that can extract file-level metadata from data files that adhere to this format.	● Ecological Metadata Language[52] ● ASCII File Format Guidelines for Earth Science Data[63] ● National Geospatial Data Assets Metadata Guidelines[64] ● Common format for CSV files[65] ● USGS Data Dictionaries[66]	● Environmental Data Initiative: Five phases of data publishing[67] ● Centre for Environmental Data Analysis[68] ● United States Geological Survey: How "clean" should Excel files be to be considered machine readable[69] ● National Ecological Observatory Network file naming conventions[70] ● Consortium of Universities for the Advancement of Hydrologic Science ODM specifications v1.1.1[71] ● StreamPulse[72]

			● Ameriflux[39]
Comma Separated Value files[27]	Recommendations for organizing tabular data stored in Comma Separated Values. Recommends optimal file structure, naming structure, and field structure.	● Common format for CSV files[65] ● ASCII format for Network Interchange[73] ● International System of Units[74] ● World Geodetic System 1984[75] ● ORNL DAAC CSV Standards[76] ● ORNL DAAC Data Quality Checklist[77] ● USGS Data Templates[78] ● National Archives: Tables of File Formats[79] ● ASCII File Format Guidelines for Earth Science Data[63] ● Date and Time on the Internet: Timestamps[80]	*Datasets downloaded from repositories/databases:* ● Ameriflux[81,82] ● Environmental data initiative[83,84] ● Mercury SFA[85,86] ● NEON data product[87] ● Next-Generation Ecosystem Experiments - Arctic[88,89] ● Next Generation Ecosystem Experiments - Tropics[90,91] ● Spruce and Peatland Responses Under Changing Environments[92,93] ● United States Geological Survey[94] ● WHONDRS[95,96] *Other data resources:* ● National Ecological Observatory Network file naming conventions[70] ● Environmental Data Initiative: Five phases of data publishing[67] ● Centre for Environmental Data Analysis[68]
Sample ID[28,44] and metadata	Enables tracking and reuse of environmental samples by providing guidance for registering physical samples for persistent identifiers. Enables sample linking across multiple collaborators, labs, and data systems online. Includes sample identifier metadata, descriptions about the sample and location.	● SESAR IGSN [97] ● IGSN Schema[98] ● DataCite[62] ● Dublin Core[99] ● Darwin Core[8] ● Minimum information about any sequence (MIxS)[55] ● Observations and measurements[100]	● Joint Genome Institute Genome Online Database and sample template [101] ● Environmental Protection Agency Water Quality Exchange[14] ● United States Geological Survey National Digital Catalog[102] ● National Geological and Geophysical Data Preservation Program Geologic Collections

			Management System[103] ● National Ecological Observatory Network Biorepository Data Portal[104]
Model archiving guidelines[29,49]	Helps modelers determine which components of their model data to archive and how to bundle data files for publication based on authorship criteria and repository storage capacities. Recommendations include archiving model inputs, testing or validation data, and any code or scripts associated with data preprocessing or modeling runs.	● National Science Foundation Earthcube modeling rubric[105]	● Earth System Grid Federation[106] ● Dryad[107] ● Zenodo[108] ● NASA's Earth Observing System Data and Information System[109] ● National Center for Atmospheric Research Data Archive[110] ● Earth Observatory Lab data archive[111] ● ESS-DIVE[4] ● National Science Foundation Arctic Data Center[112]
Locations metadata[30]	Minimum set of metadata to report about locations where field samples or observations are collected. While the sampling features described in the reporting format are primarily point locations, the format can optionally be expanded to capture metadata about plots and regions	● CF Conventions[56] ● DarwinCore[8] ● Content Standard for Digital Geospatial Metadata[57] ● Content Standard for Digital Geospatial Metadata: Biological Data Profile[113] ● Open Geospatial Consortium observations and measurements[58]	● Watershed Function Scientific Focus Area sensor metadata template[41] ● Watershed function SFA locations template[42] ● Ameriflux Biological, Ancillary, Disturbance, and Metadata[15] ● iSamples[46] ● Framework for Reporting Data and Metadata for Earth Science[114] ● United States Geological Survey National Water Dashboard[115] ● United States Department of Agriculture Snowpack Telemetry[116]
Amplicon abundance	Microbial abundance profiles and sequences obtained from genomic	● Minimum information about a marker gene sequence	● Earth Microbiome Project Metadata Guide[117]

table metadata[31]	sequencing of environmental samples	(MIMARKS) for sequencing metadata.[55] ● No data standards existed for amplicon abundance table bioinformatic processing metadata.	● National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) Metadata and Submission Overview[118]
Leaf-level gas exchange[32,47]	Guidelines for providing methods metadata and instrument metadata for experiments collecting leaf-level gas exchange data. Defined variable names and units for data, and specifies required variables for various measurement types.	No data standards existed	● TRY database[119] ● Biofuel Ecophysiological Traits and Yields[120] ● Botanical Information and Ecology Network[121] *Instruments* ● ADC iFL ● ADC LCi T ● ADC proT ● CID CI-340 ● Licor 6400XT ● Licor 6800 ● PP Systems CIRAS-2 ● PP Systems CIRAS-3 ● PP Systems TARGAS-1 ● Walz GFS-3000
Soil respiration[33,48]	Brings together decentralized and individual-driven field-measured data. The chamber-level metadata template facilitates reporting information about experimental conditions including variables being measured, chamber size and design, and instrumentation. The other template is for data entry of CO2 measurements from experimental chambers.	● Integrated Carbon Observation System[122]	● AmeriFlux[38] ● FLUXNET[123] ● Global N20 Database[124] ● International Soil Radiocarbon Database[125] ● Soil Incubator Database[126] ● Soil Respiration Database[127] ● Trace Gas Network[128] ● Hibbard et al. 2005[129]

Sample-based water and soil chemistry[34]	Recommendations for consistently formatting data and metadata obtained from laboratory chemical analysis of field samples. Reporting chemical concentrations and metadata for data collection methods.	● Observations Data Model 2 (ODM2)[130] ● Minimum information about any sequence[55] ● Water Quality Exchange[14] ● National Ecological Observatory Network Protocols & Standardized Methods[131] ● EarthChem[132]	● SLAC Floodplain Hydro-Biogeochemistry SFA[133] ● LBNL Watershed Function SFA[134] ● PNNL River Corridor Hydrobiogeochemistry SFA[135] ● ANL Wetlands SFA [136] ● LLNL Biogeochemistry at Interfaces SFA [137] ● ORNL Mercury SFA [138] ● WHONDRS[139] ● AmeriFlux [140] ● NEON[141] ● Terragenome[142]
Water level and sonde-based hydrologic monitoring[35]	Hydrologic sensor-based data and corresponding deployment metadata. Includes a terminology file that recommends terms for commonly used hydrological variable names that had significant variation across prior standards and data resources	● Observations Data Model 1 used in the Environmental Data Initiative (EDI) and Consortium of Universities for the Advancement of Hydrologic Science - Hydrologic Information System[143] ● Observations Data Model 2[130]	● United States Geological Survey National Water Information Systems[144,145] ● National Ecological Observatory Network Data Formats and Conventions[146] ● Water Quality Portal[147]

Table 2

Examples of datasets published on ESS-DIVE utilizing at least one of the 11 ESS-DIVE (meta)data reporting formats. Each row includes the dataset title, citation, and the reporting format(s) used in the dataset.

Dataset Title	**Reporting Format(s) Used**
FTICR, NPOC, TN, and Moisture of Variably Inundated Sediment across 48 North American Rivers[148]	Sample-based water and soil chemistry, Sample ID and metadata, Comma Separated Value files, and File-level metadata Reporting Formats
Kinetic and temperature sensitivity properties of soil exoenzymes through the soil profile down to one-meter depth at a temperate coniferous forest (Blodgett, CA)[149]	Sample ID and metadata, Comma Separated Value files, and File-level metadata Reporting Formats
Leaf Photosynthetic Parameters: Quantum Yield, Convexity, Respiration, Gross CO2 Assimilation Rate and Raw Gas Exchange Data, Utqiagvik (Barrow), Alaska, 2016[150]	Leaf-level gas exchange Reporting Format
Perceived Costs and Benefits of ICON Science and Foundational Documents associated with “Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles[151]	File-level metadata Reporting Format
FTICR-MS, Sensor, and Environmental Data from 5 Streams Impacted by the 2020 Holiday Farm Fire Associated with: "Spatiotemporal controls on the delivery of dissolved organic matter to streams following a wildfire[152]	Hydrologic Monitoring, Comma Separated Value files, and File-level metadata Reporting Formats
Fungal and bacterial growth variation due to drought and nitrogen addition experimental treatments.[153]	File-level metadata and Comma Separated Value files Reporting Formats
Chemistry data from soils and soil incubation experiments from the whole-soil warming experiment at Blodgett Forest, CA, 2018, from: “Metabolic capabilities mute positive response to direct and indirect impacts of warming throughout the soil profile”[154]	File-level metadata and Comma Separated Value files Reporting Formats

Table 3

Evaluation of how ESS-DIVE's reporting formats achieve the Findable, Accessible, Interoperable, and Reusable data principles [1,21] in the context of the ESS-DIVE data repository (some of the FAIR Principles are not applicable at the reporting format level and implementation occurs through the data repository supported infrastructure / metadata). All datasets published in the ESS-DIVE repository adhere to the Dataset Metadata reporting format, which at least partially satisfies many FAIR principles, particularly Findability and Accessibility. For some FAIR principles (e.g., F4) submitting data to ESS-DIVE (regardless of which domain-specific format is used) will automatically ensure that the principle is met. When reporting formats do not fully achieve a FAIR principle, we list the reporting format and progress made toward fully machine-actionable and FAIR (meta)data.

FAIR Principle	Number of reporting formats achieving FAIR principle when combined with repository features	Reporting formats not fully achieving FAIR principle	Progress toward achieving FAIR principle
F1. (meta)data are assigned a globally unique and persistent identifier.	11	All achieve FAIR principle on ESS-DIVE or other data repositories+	All datasets uploaded to ESS-DIVE are associated with a persistent DOI. Data and metadata associated with samples and associated metadata are assigned an IGSN.
F2. data are described with rich metadata	11	All achieve FAIR principle	All reporting formats provide guidance and many include templates for including descriptive metadata about entire datasets, individual data files, samples, and specific data types. Metadata requirements were defined by scientists who provide and potentially search for the data.
F3. (meta)data clearly and explicitly include the identifier of the data it describes.	9	CSV reporting format*, model data archiving guidelines*	Data files are assigned an identifier and explicitly linked via ESS-DIVE's use of the Ecological Metadata Language[52]
F4. (meta)data are registered or indexed in a searchable resource.	1	CSV reporting format, model data archiving guidelines, soil respiration, leaf-level gas exchange, sonde-based hydrologic monitoring, sample-based water and soil chemistry, amplicon abundance table metadata, file-level	Metadata provided about each dataset submitted to ESS-DIVE are indexed and searchable as part of the ESS-DIVE repository. Individual reporting formats are not currently at a stage where the data contained within files are automatically indexed and searchable except for dataset metadata. This is a target for future work.

		metadata, sample ID and metadata, locations metadata	
A1. (meta)data are retrievable by their identifier using a standardized communications protocol.	9	CSV reporting format*, model data archiving guidelines*	When data are archived in the ESS-DIVE repository, metadata are available over HTTP.
A1.1. the protocol is open, free, and universally implementable.	11	All achieve FAIR principle on ESS-DIVE or other data repositories+	Users can access (meta)data on ESS-DIVE over HTTP.
A1.2. the protocol allows for an authentication and authorization procedure, where necessary.	11	All achieve FAIR principle on ESS-DIVE+	Data that adheres to reporting formats is stored in ESS-DIVE repository which allows for authentication and authorization, so individual reporting formats do not need to have their own protocols.
A2. (meta)data are accessible, even when the data are no longer available	11	All achieve FAIR principle on ESS-DIVE+	Data are stored using persistent identifiers and in non-proprietary formats. Also, ESS-DIVE stores metadata independent of the datasets and the metadata are replicated across the DataONE network.
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	1	CSV reporting format, model data archiving guidelines, soil respiration, leaf-level gas exchange, sonde-based hydrologic monitoring, sample-based water and soil chemistry, amplicon abundance table metadata, file-level metadata, sample ID and metadata, locations metadata	All reporting formats made good strides toward machine readable knowledge representation. However, based on community feedback most prioritized human-readable explanations of variables via data dictionaries and term lists rather than fully machine readable knowledge representation. Having flexibility in term usage was considered important to lower the barrier for data contributors to adopt the reporting formats.
I2. (meta)data use vocabularies that follow FAIR principles.	4	CSV reporting format, model data archiving guidelines, soil respiration, leaf-level gas exchange, amplicon abundance table metadata, file-level metadata, locations metadata	Based on feedback from our community engagement, we made pragmatic choices that involved leaving out ontologies that made adoption of the formats burdensome in lieu of terms that scientists preferred to use. For samples, we chose a small subset of formal ontology terms to describe environmental context relevant for the ESS community. Scientists

			are encouraged, but not required to use ontologies when applicable.
I3. (meta)data include qualified references to other (meta)data.	11	All achieve FAIR principle	All data submitted to ESS-DIVE can be externally linked to related resources in the repository.
R1. (meta)data are richly described with a plurality of accurate and relevant attributes.	11	All achieve FAIR principle	The dataset metadata and file-level metadata reporting formats ensure that entire datasets and individual files are described with rich metadata. Moreover, reporting formats also provide space for metadata provision for specific data types (e.g., chamber metadata for the soil respiration reporting format).
R1.1. (meta)data are released with a clear and accessible data usage license.	11	All achieve FAIR principle	All (meta)data submitted to ESS-DIVE are published with either CC BY 0 or CC BY 4.0 licenses.
R1.2. (meta)data are associated with detailed provenance^.	11	All achieve FAIR principle	All reporting formats make progress at tracking information about how and/or when data were collected via elements that collect detailed information about locations and timestamps when data were collected. Some reporting formats (e.g., sonde-based hydrologic monitoring, sample-based water and soil chemistry) include templates for methods-specific metadata. ESS-DIVE allows linking datasets to associated external (meta)data using a set of formal machine-readable relationship fields/terms, and is working towards extending to a variety of related resources involved in ESS work.
R1.3. (meta)data meet domain-relevant community standards	11	All achieve FAIR principle	The community reporting format effort is directly achieving this principle by involving research communities in creating (meta)data standards (i.e., reporting formats).

An asterisk (*) indicates when the FAIR principle did not apply to the reporting format because it was an instruction-based guideline rather than templates and tools.
The caret (^) indicates that our interpretation of provenance in this case is that all reporting formats allow users to track components of provenance fundamentals including timestamps, location information, and instruments used, among other elements.

A plus (+) indicates that the FAIR principle is met when combined with specific features offered by the ESS-DIVE repository as specified in the last column of the table. The reporting format can achieve the FAIR principle on other repositories that offer the same features (e.g., assignment of persistent identifiers).

Box 1

Guidelines for research communities to self-organize and create, document, and share (meta)data reporting formats when formats do not already exist or do not fit scientists' needs.

1. **Research existing (meta)data standards and other data resources** across agencies and organizations both within the US and internationally.

2. Create a **(meta)data crosswalk** (Supplementary Files 1-10) to define how other standards and data resources translate to the proposed reporting format.

3. Work with the scientific community to **iteratively develop and obtain feedback** (see Figure 2) on (meta)data reporting format.

4. Develop **documentation** (instructions, templates, variables, descriptions, units, metadata) to support the format. Consider appropriate **file formats** for any templates.

5. Archive finalized version of the reporting format in a **long-term data repository** as well as a version control platform (e.g., GitHub[37]).

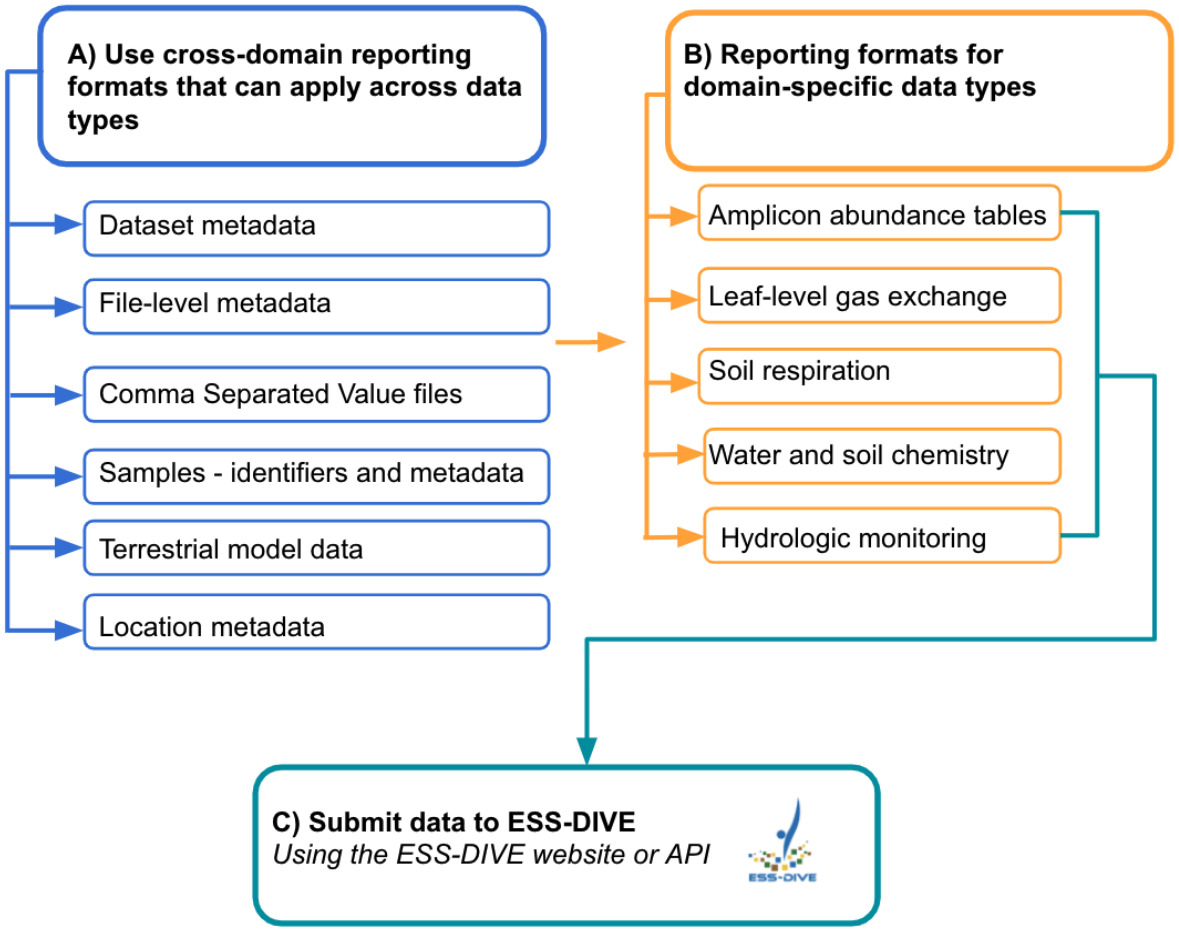

Figure 1

Workflow to help determine which (meta)data reporting formats apply to datasets. The set of 11 ESS-DIVE (meta)data formats are either (A) cross-domain guidelines that can be applied to many data types or (B) are data type-specific. For those archiving data with ESS-DIVE, researchers can upload data through the ESS-DIVE web user interface[155] or programmatically through an API.

Figure 2

Each of the 11 ESS-DIVE (meta)data reporting formats were developed in cross-functional teams that often involved domain scientists, software engineers, and informatics specialists.