



Data Package Quality Review

Creating and submitting high quality data packages to ESS-DIVE

Emily Robles, Rob Crystal-Ornelas



U.S. DEPARTMENT OF
ENERGY

Office of
Science

March 2021 Webinar

Presentation Overview

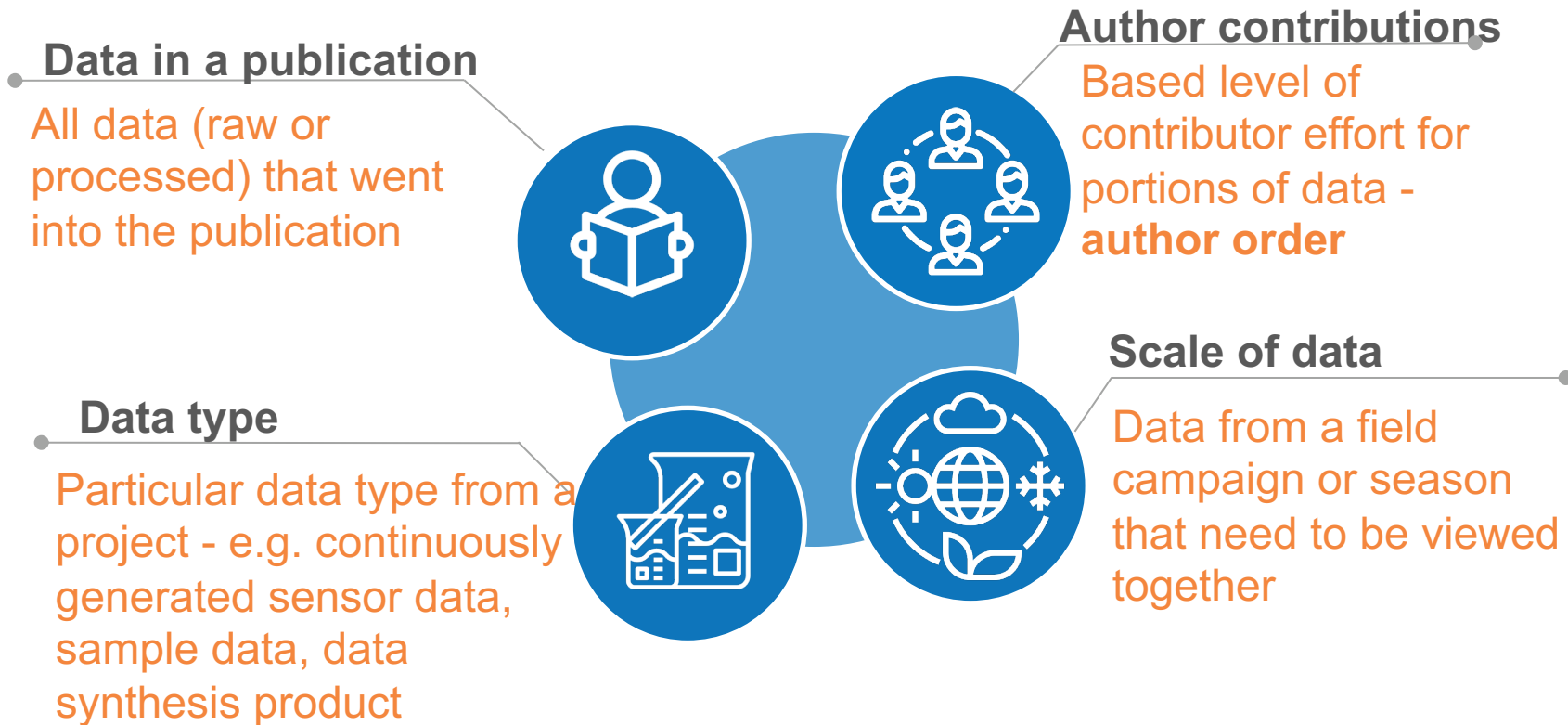
- Building a **data package**
- **Publishing high quality data** on ESS-DIVE
- Making your published data **reusable**

We are here to support you in submitting your data to ESS-DIVE!

BUILDING A DATA PACKAGE:

First steps and organization

Deciding what to include in a data package



Helpful resources

Data file organization

- [NCEAS webinar for tidy-data practices](#)

ESS-DIVE submission and publication process

- [Video Tutorials](#)
- [Previous webinar](#)
- [Help documentation](#)

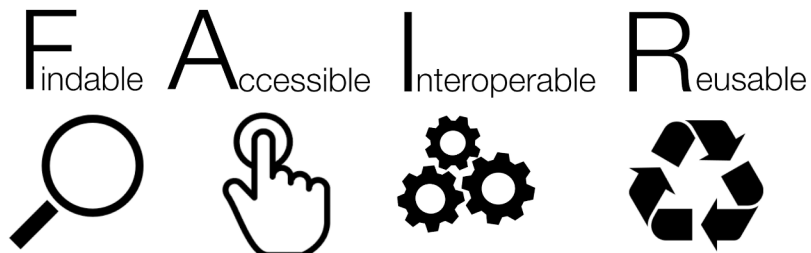
PUBLISHING QUALITY METADATA:

Creating and publishing high-quality metadata
to describe your data

Descriptive metadata make data more FAIR



- Datasets are valuable research contributions
- More journals and funders are implementing public data archival (PDA) policies
- The FAIR principles have been accepted across scientific domains as guidelines for maximizing the value and longevity of data
- Complete metadata can increase the FAIRness of your data



COMMENT · 04 JUNE 2019 · CORRECTION 05 JUNE 2019

Make scientific data FAIR

All disciplines should follow the geosciences and demand best practice for publishing and sharing data, argue Shelley Stall and colleagues.

ESS-DIVE tools to create your metadata

- Package level metadata guide: <https://docs.ess-dive.lbl.gov/data-and-metadata-upload/package-level-metadata>
 - Highlights specific criteria we use in standard review
- [Offline metadata template](#) can be used to easily collaborate with team members

Package Level Metadata Guide

Use this page to review guidelines and expectations for each metadata field. The JSON-LD Field Names refer to vocabulary used for submission through the API.

Our [Offline Metadata Template](#) can be used to prepare your data package metadata prior to submission. We recommend using the template to collaborate with your team members in Google Docs, then copying and pasting the completed fields into the ESS-DIVE Data Package Submission form when you are ready to create your data package.

- **Overview**
 - Title
 - Existing DOI & Alternate Identifiers
 - Abstract
 - Keywords
 - Data Variables
 - Publication Date
 - Usage Rights
 - Project
 - Funding Organization
 - DOE Contracts
 - Related References
- **People**
 - Contact

ESS-DIVE Offline Metadata Template For Data Package Web Form Submissions		Required to Save Draft Package	Required to Publish
Web Form Metadata Fields	Data Package Metadata		
OVERVIEW			
Data Package Title A brief title between 7-20 words long which contains relevant information such as the topic, geographic location, dates, and scale of data.	Example: Raw sapflow and soil moisture data from Jan 2016-Apr 2016 in Manaus, Brazil		
Existing DOI(s) and Alternate Identifier(s) If this data package has been previously published elsewhere, enter the DOI or alternate identifier. Identifiers are used to locate the dataset within your project's data management system and can provide pertinent	Example: http://dx.doi.org/10.15486/NGT/XXXXXX		

Metadata review process

A two-part review system with **automated** and **manual** components streamlines metadata curation.

Automated Review

- Run as soon as data package is submitted
- Provides instant feedback in the form of a Metadata Assessment Report
- Cover a wide range of metadata checks, but do not include content-related checks

Manual Review

- After data package publication is requested
- Content-related metadata checks
- An ESS-DIVE team member carries out each review
- Detailed feedback responses are sent to data contributors for revisions

Using the Metadata Assessment Report

Metadata Assessment Report

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.



Identification: 89% complete



Discovery: 67% complete



Interpretation: 100% complete



▶ Passed 11 checks out of 14 (informational checks not included).

▶ Warning for 0 checks.

▶ Failed 3 checks. Please correct these issues.

▶ 2 informational checks.

Reviewing and resolving failed checks

- Assessment reports can be reviewed after submission
- Failed checks must be resolved before publication

▶ Passed 11 checks out of 14 (informational checks not included).

▶ Warning for 0 checks.

▼ Failed 3 checks. Please correct these issues.

✘	No ORCID is provided for the contact person.	?	identification	REQUIRED	FAILURE
✘	The abstract is only 1 word(s) long, and 100 or more words are required.	?	discovery	REQUIRED	FAILURE
✘	Start and end dates describing the temporal coverage of this dataset are not present.	?	discovery	REQUIRED	FAILURE

▶ 2 informational checks.

Manual review

- Focus on content checks
 - Metadata fields may pass the automated check but fail the manual
 - Important to review our requirements
- Failed checks must be resolved before approval for publication
- Feedback generated from the manual review are sent over email

ESS-DIVE Manual Package Metadata Review

* Required

Keywords and Data Variables

There are at least 3 keywords, which must differ from words in the title and be relevant and appropriate for the dataset. Generic entries that will not improve the findability of the dataset should not be counted as a part of the 3 keyword minimum.

Keywords *

☐ PASS

☒ FAIL

☐ N/A

[Back](#) [Next](#)

Resolving other revision requests

Detailed feedback will be listed with instructions on how to resolve requirements that are not met

Hello,

We have finished reviewing your data package and request the following improvements to your metadata before publication:

1. Consider adding a full Geographic Description to improve the findability of your dataset, including coordinates so that we can geolocate your data package and it will show up in geographic searches. In addition to providing standard coordinates here, you have the option to include a KML file if the data is better represented by a shape.
2. Consider adding a methods description, which is essential for interpretation and reusability of your data. A complete methods section will also improve findability of your data, as all text entered into methods will also be searchable for those using keyword searches. However, if desired you can provide a citation for any methods used that were published previously.
3. Please include email addresses for each creator. Authors with insufficient information may be difficult for users to locate.

Note that once the data package is published, it will be publicly available for search and download. The publication action cannot be undone, however, you will still be able to edit or retire the data package. As the dataset contact, you are responsible for obtaining consent for publication from the data package authors and notifying them about its status. Please respond to this message once you are finished making these revisions so that we can continue the approval process.

Thanks,

The ESS-DIVE Support Team

Common metadata improvements

- Title
- Abstract
- Keywords and Variables
- Methods

****Think about what others need to find and understand your data***

Create a descriptive title

Unclear, not meaningful

- Unexplained acronyms
- Project-specific vocabulary

Good

- 7-20 words
- Information on topic, geographic location, dates, and scale of data
- Necessary acronyms are explained in the data package abstract

Create a descriptive abstract

Unclear, not meaningful

- < 100 words
- Incomplete sentences, grammatical errors
- Undefined acronyms
- Project specific terminology that could be unclear to users

Good

- **At least 100 words**
- Clear and concise description of content and purpose of the data
- Outlines research question
- Accessible for readers outside of your project

Abstract example



Purpose

Abstract

Sample
Collection

Analyses

Contents

This dataset is from a global survey of surface water metabolites to provide understanding of the character of organic carbon that may be delivered to subsurface sediments via hydrologic exchange. To implement the global survey, free stream sampling kits were provided to interested researchers throughout the world. Samples were collected with minimal constraints in terms of location, but following strict protocols, and shipped to the Environmental Molecular Sciences Laboratory (EMSL) for metabolomic analysis via Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS). In addition, basic geochemistry analyses (e.g., dissolved organic carbon concentration, cations, anions) were conducted, standardized photos of each field system were taken, surface water hydrographs were collated from existing instrumentation, and extensive metadata were captured. All data types are provided in a standard format. In addition, the data package contains an R function that will launch a GUI that can be used to easily search, compile, and download data. The data are free to be used for any purpose, such as for manuscripts, presentations, and grant proposals. Please use the data package's DOI to cite the data package. Note that individual hydrographs have separate DOIs, which are provided in the associated hydrograph files. These hydrograph-specific DOIs should also be cited when using those data. We ask that you email us at WHONDRS@pnnl.gov to let us know that you're using the data and acknowledge WHONDRS and the U.S. Department of Energy's Subsurface Biogeochemical Research program—which generously provides funding to WHONDRS—in your documents, presentations, etc. There is no obligation to include WHONDRS members as co-authors.

257 words

Relevant keywords and variables

Unclear, not meaningful

- Already included in title
- Repeated terms in variables and keywords
- Spelling errors
- Undefined acronyms

Good

- 3 keywords or variables, controlled list where possible
- Differ from the title
- Clarifies data/variable types

Keywords example

Keywords *

Keywords that should be associated with this data package to enable thematic searches.

Search for a keyword from the list or write in your own. Tab or click enter to add to the list below with one keyword per line. The list contains [GCMD](#) keywords.

Use autocomplete feature to pick from the existing keywords.

Earth|

EARTH SCIENCE > AGRICULTURE > AGRICULTURAL AQUATIC SCIENCES: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > AGRICULTURAL CHEMICALS: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > AGRICULTURAL ENGINEERING: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > AGRICULTURAL PLANT SCIENCE: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > ANIMAL COMMODITIES: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > ANIMAL SCIENCE: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > FEED PRODUCTS: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > FOOD SCIENCE: CATEGORICAL:GCMD

Keywords example



Keyword
EARTH SCIENCE > BIOSPHERE > ECOSYSTEMS
EARTH SCIENCE > BIOSPHERE > VEGETATION

Keyword
alpine tundra
belowground plant production
functional traits

Include descriptive methods

Unclear, not meaningful

- No methods provided
- Link to paper methods, which often focus on statistical analyses and exclude details on data production

Good

- At least 7 words in length
- Focus on all aspects of **data production**
- Can include details on: experimental design, laboratory and/or field collection methods, source data for synthesis studies, data processing and QA/QC procedures
- Thorough enough for your work to be reproduced

Methods example



Methods & Sampling

Methods

Step 1

Description

The methods below are reproduced from those given in this data package's Data User's Guide, which is adapted from the publication. The full publication can be found at <https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/1365-2745.12750>.

Citation: Conlisk, E., Castanha, C., Germino, M.J., Veblen, T.T., Smith, J.M. and Kueppers, L.M. (2017), Declines in low-elevation subalpine tree populations outpace growth in high-elevation populations with warming. *J Ecol*, 105: 1347-1357. <https://doi.org/10.1111/1365-2745.12750>

Seed production and dispersal

A variety of studies show ample, but highly variable, seed production for Engelmann spruce, with most un-germinated seeds not surviving in the field to the following year (see "Conlisk_JofEcology_SI_01262017" from this archive). We assumed that the largest Engelmann spruce individuals could produce 1462 viable seeds per year. Limber pine produces fewer, larger, better-provisioned seeds that are also highly desirable to seed predators. We assumed that the largest limber pine individuals could produce 479 viable seeds per year. For both species, seed production increased linearly (based on Stromberg & Patten 1993) with stage, starting with one seed produced when a tree was, on average, 45 years of age.

We assumed dispersal between the tree line patch and either the alpine or forest patch (and no dispersal between the alpine and forest patches) of roughly 0.05% of Engelmann spruce seeds and 0.5% of limber pine seeds. Engelmann spruce dispersal was based on Alexander (1987) who reported exponential decline of dispersing seed with distance. We could not find a study documenting limber pine dispersal with distance. However, long-distance dispersal has been reported for *Strobus* pines dispersed by corvids and small mammals (see "Conlisk_JofEcology_SI_01262017" from this archive). Thus, we assumed greater overall dispersal for limber pine.

Sensitivity analyses

To evaluate the effect of alternate parameter choices, particularly for unobserved parameters, we conducted sensitivity analyses that considered lower dispersal (10% of original value), lower seed production (80% of original value), reduced sapling survival (98% of original value) and reduced adult survival (98% of original value). Sensitivity tests found that population growth rates were most sensitive to small changes in adult survival, but that differences among climate scenarios (e.g. warmed, watered) were robust to model parameterization (see "Conlisk_JofEcology_SI_01262017" from this archive).

Conlisk E ; Castanha C ; Germino M J ; Veblen T T ; Smith J M ; Kueppers L M (2017): Data from: "Declines in low-elevation subalpine tree populations outpace growth in high-elevation populations with warming". Subalpine and Alpine Species Range Shifts with Climate Change: Temperature and Soil Moisture Manipulations to Test Species and Population Responses. doi:10.15485/1730950

Data reporting formats and *your data*

Complex data in Earth and Environmental Science



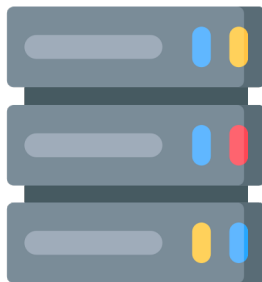
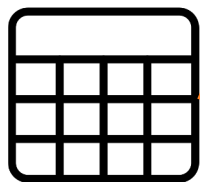
- The data are **diverse**
- How can we make them reusable over the long-term?



Photo credit: LBNL

ESS-DIVE is DOE's central location for long-term data preservation

Data stored in repositories should be Findable Accessible Interoperable and Reusable



60% of publicly archived data face reusability issues



Challenge

How can increase the % of reusable data in long-term archives?



Solution

Make sure the data are **well-described** and **consistently formatted**

“Of the many potential applications for a particular dataset, there is often only time to explore a small subset”

PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Principles for data analysis workflows

Sara Stoudt^{1,2}, Valeri N. Vásquez^{1,3}, Ciera C. Martinez^{1,4*}

Similar types of data can be difficult to reuse if they lack consistent formatting



idNumber	material	temperature
3928	soil	23.2
3234	groundwater	9.02



sampleNum	substance	temp
8765	dirt	21.1
2312	ground liquid	7.0

Small changes can make data more reusable



idNumber	material	temperature
3928	soil	23.2
3234	groundwater	9.02



idNumber	material	temperature
8765	soil	21.1
2312	groundwater	7.0

Clarifying terminology: Data standards and reporting formats

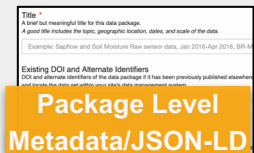


- **Data Standards** - Decades of development, accredited by governing org.
- **Reporting Formats** - Community-driven still enable data harmonization and synthesis

Darwin Core



ESS-DIVE Community-led reporting formats for many types of data



Agarwal, Hendrix (LBNL)

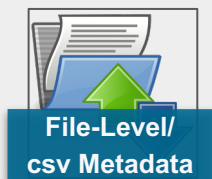


Damerow (LBNL)



Damerow (LBNL)

- Review of **existing formats**
- **Rounds of feedback** with researchers and projects
- All documentation on our **GitHub Community Space**



Velliquette, Heinz,
Devarakonda (ORNL)



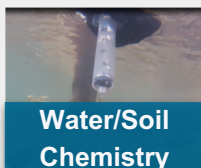
Bond-Lamberty,
Pennington (PNNL)



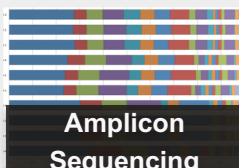
Rogers, Ely (BNL)



Goldman (PNNL)



Boye (SLAC)



Weisenhorn (ANL)

Using the data reporting formats



Visit our **Community Space on GitHub**: <https://github.com/ess-dive-community>

ESS-DIVE Community Space

A workspace for code, tools, reporting formats and other products related to the ESS-DIVE repository. Please message ess-dive-support@lbl.gov to contribute


essdive-water-soil-sed-chem

IN DEVELOPMENT. A reporting format for water-soil-sediment chemistry data 

water soil sediment reporting-format

 CC-BY-4.0  0  0  0  0 Updated 23 days ago

essdive-csv-structure

READY TO USE. Reporting format for CSV files submitted to the ESS-DIVE repository 

reporting-format

 CC-BY-4.0  1  0  4  0 Updated 27 days ago


The CSV reporting format **on GitHub**: <https://github.com/ess-dive-community/essdive-csv-structure>












ess-dive-community / essdive-csv-structure Unwatch

[Code](#) [Issues 4](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#)

[master](#) [1 branch](#) [0 tags](#) [Go to file](#) [Add file](#) [Code](#)

 **robcristalornelas** Change 'standard' to 'reporting format' ✓ b7e2674 on Feb 26 🕒 26 commits

 .github/ISSUE_TEMPLATE	Create other.md	3 months ago
 images	updates to detailed guide	3 months ago
 LICENSE.md	Create LICENSE.md	3 months ago
 README.md	Update README.md	2 months ago
 SUMMARY.md	Update SUMMARY.md	3 months ago
 csv_crosswalk.md	Update csv_crosswalk.md	3 months ago
 csv_detailed_guide.md	update bullet points	3 months ago
 csv_instructions.md	Update csv_instructions.md	3 months ago
 csv_quick_guide.md	Change 'standard' to 'reporting format'	last month

An example: Using the CSV reporting format



Be sure to read through supporting documents before following reporting format recommendations

Recommendations for **Column names**

- Descriptive as possible
- Only use letters, numbers, hyphens, and underscores

	A	B	C
1	Site Name	pH	Daily Temp
2	Missing	7.3	23.2
3	AB-123	7.1	-94720



	A	B	C
1	Site_Name	pH	Daily_Temp
2	N/A	7.3	23.2
3	AB-123	7.1	-9999

An example: Using the CSV reporting format



Be sure to read through supporting documents before following reporting format recommendations

Recommendations for **Missing values**

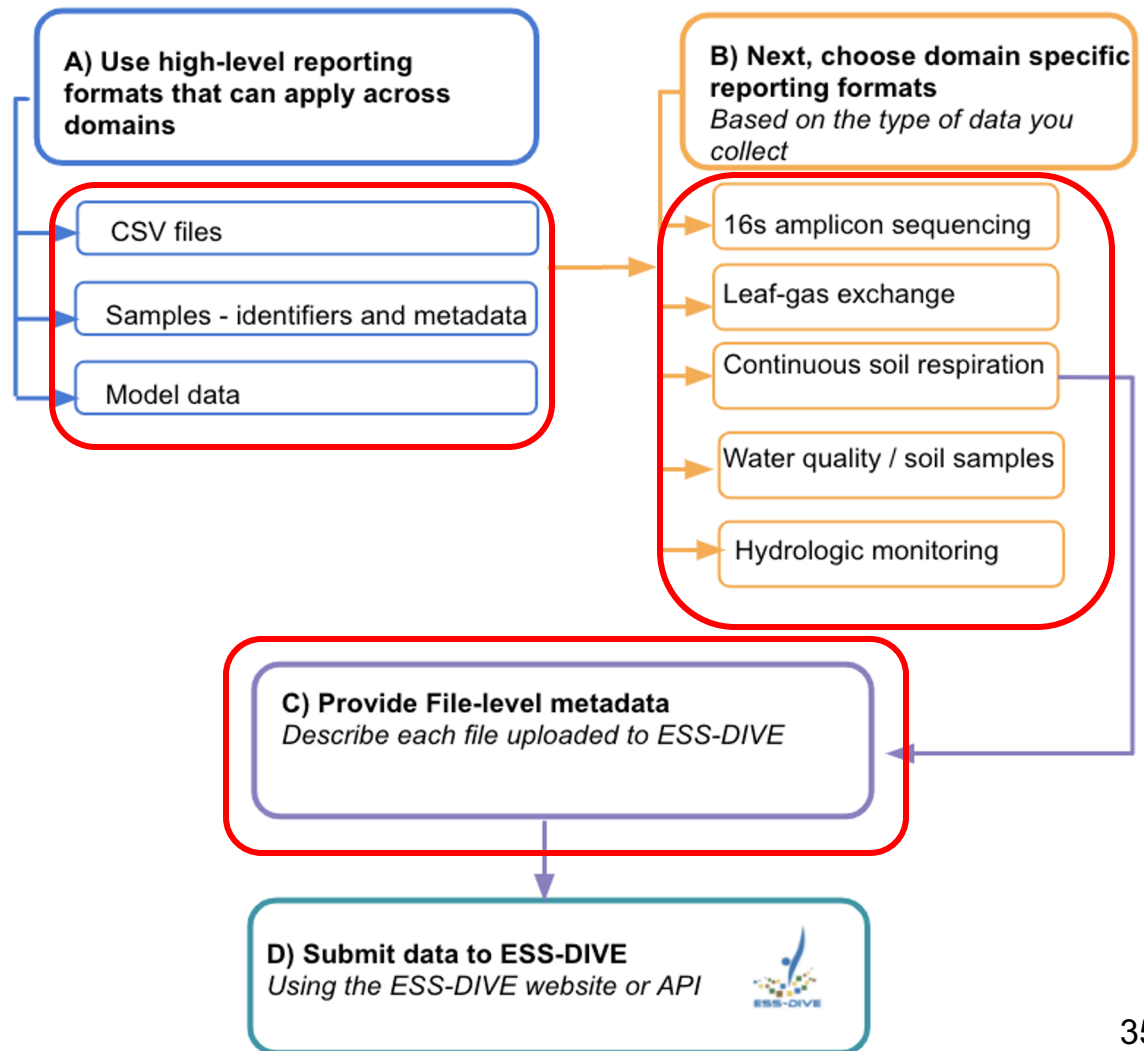
- No empty cells
- “-9999” for missing numeric, “N/A” for missing text

	A	B	C
1	Site Name	pH	Daily Temp
2	Missing	7.3	23.2
3	AB-123	7.1	-94720



	A	B	C
1	Site_Name	pH	Daily_Temp
2	N/A	7.3	23.2
3	AB-123	7.1	-9999

Choose the reporting formats the fit your data



We are here to help with reporting formats

- Reporting formats can enable reuse and synthesis
- Visit our Community Space on GitHub
- Send questions to: ess-dive-support@lbl.gov

Thank you!



References

1. ESS-DIVE's Community Space on GitHub: <https://github.com/ess-dive-community>
2. Stoudt, S., Vásquez, V. N., & Martinez, C. C. (2021). Principles for data analysis workflows. *PLoS Computational Biology*, 17(3), e1008770

ESS-DIVE Glossary



- **DataONE** - The Data Observation Network for Earth (DataONE) is a distributed framework and sustainable cyberinfrastructure that provides open and secure access to Earth observational data. ESS-DIVE is a DataONE member.
- **DOE** - The U.S. Department of Energy (DOE) is a Cabinet-level department of the United States whose mission is to ensure America's security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions.

ESS-DIVE Glossary (*cont.*)

- **DOI** - A Digital Object Identifier (DOI) is a unique alphanumeric string assigned by a registration agency (e.g., The Office of Scientific and Technical Information (OSTI)) to identify content and provide a persistent link to its location on the internet. ESS-DIVE assigns a DOI when your data package is published and made available electronically.

ESS-DIVE Glossary (*cont.*)

- **ESS** - Environmental Systems Science (ESS) is a U.S. Department of Energy Office of Science program under the Biological and Environmental Research Program seeking to advance a robust predictive understanding of terrestrial surface and subsurface ecosystems.
- **ESS-DIVE** - Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) is a U.S. Department of Energy archive for earth and environmental science data, models and software generated from research on terrestrial and subsurface environments.