How to create your datasets and format data files for publication

Emily Robles
Rob Crystal-Ornelas

August 2021 Webinar

What is your familiarity with ESS-DIVE (select one)?

Published datasets previously

new registered contributor

contributor considering publishing data through ESS-DIVE

In relation to publishing data, which of the following terms are you familiar with?

Data package

Metadata

Dataset

Data repository

DOI (Digital Object Identifier)

FAIR

What kinds of files do you plan on submitting (select all that apply)

tabular data

images

maps

model components

PDF/supporting information

scripts and coding notebooks

other file types

What challenges do you face when creating datasets, or what challenges do you anticipate as a new user?

Presentation Overview

- **What** data to publish
- Considerations when delineating **multiple datasets** for a project
- How to **publish your dataset** on ESS-DIVE
- How to format and describe data files using the **file-level metadata (FLMD)** reporting format

Takeaways

- How to **organize** data files and package it with relevant metadata
- What happens after **requesting publication**
- How **CSV and file-level metadata** reporting formats can make your data more reusable

PUBLISHING DATASETS:
What to include

Important Terms

Dataset/Data Package: A dataset, also called a data package, contains data files and their relevant metadata. Public datasets can be viewed and downloaded from the ESS-DIVE main search portal.

Metadata: Accompanies data and provides users with enough information to interpret whether a dataset is useful to their purposes. Data contributors are required to include metadata with their data when submitting to ESS-DIVE.

DOI: A Digital Object Identifier is a persistent identifier that will link to your dataset's location on the internet. ESS-DIVE assigns a DOI when your data package is published and made available electronically.

Reasons to publish data

Abide by journal and funding requirements

Most journals are starting to **require** data associated with paper findings, figures, and tables to be **publicly available on a long-term data repository**

Include **DOI's**, such as those issued by ESS-DIVE, in the **Data Availability** section of a paper

Share your work with the community

Gain **publicity** from data publications, similarly to journal publications

Allow others to use your work for **future studies**

Promote FAIR data practices

Findable, Accessible, Interoperable, and Reusable

Data reporting formats and metadata requirements abide by these standards

DOI: Digital Object Identifier

Considerations to split up data packages

Author contributions

Based level of contributor effort for portions of data - affects author order

Data in a publication

All data (raw or processed) that went into a publication

Campaign / Time Period

Data from a field campaign or season that need to be viewed together

Data type

Particular data type from a project - e.g. continuously generated sensor data, sample data, data synthesis product

Steps to publish your data

Collect your data files

01

Collect and organize data files related to your findings, tables, and figures

Refer to **ESS-DIVE reporting formats**

Register as a contributor

02

Register to submit data ESS-DIVE

Ensure that your data **fits the requirements** of ESS-DIVE.

Publish with metadata

03

Publish data and metadata as a data package

Meet the funding/journal publishing time frame

BER: publish within 1 year of end of data collection, or at the time of publication

Components of a dataset (data package)

Data Files: csv, images, code, kml

Dataset

Data Files

Include data files, all data relevant to scientific findings and figures/tables of a paper, or all data necessary for replication of the science

- Measurements
- Model inputs, outputs, code, scripts
- Supporting information, maps, documentation

Metadata

Include information to allow other users to interpret whether your data package is useful to their purposes

- Purpose of the data
- Contents of the data files (results, variables, etc)
- Temporal and geographic information
- Instructions for use

Metadata: Data / info about data

Data File Types

01

ReadMe

Directory of files included with additional metadata

02

Tabular Data

Sample, analysis, observational data

03

File Level Metadata

Description of individual files within the package

04

Model Components

Model inputs, outputs, code, scripts

05

Maps

KML or KMZ files with geographic data

06

PDFs, Supporting Info

Instrument manuals, methods writeups, etc

Metadata components

01

Title

Descriptive overview of the data package

02

Abstract

Purpose, contents, location, instructions

03

Keywords

Variables, keywords not already included in title

04

Location

Latitude and longitude, location description

05

Methods

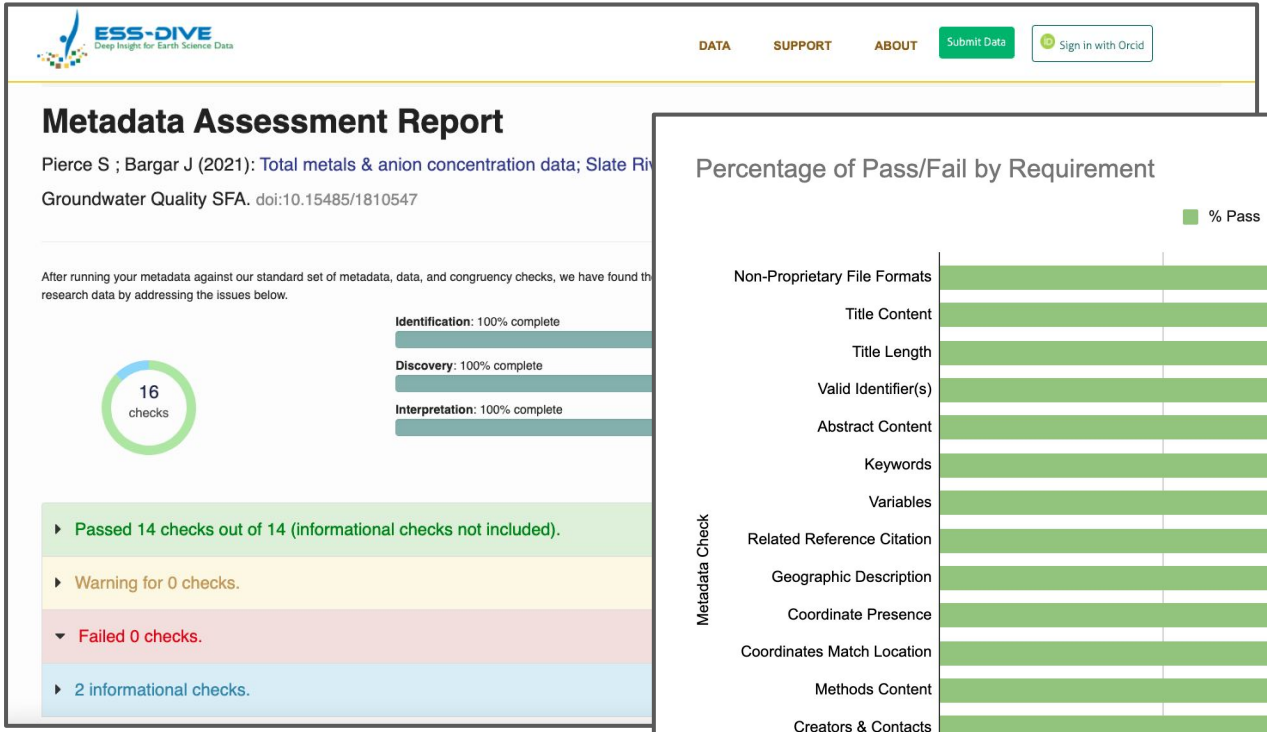
Data collection, processing, QA/QC, error

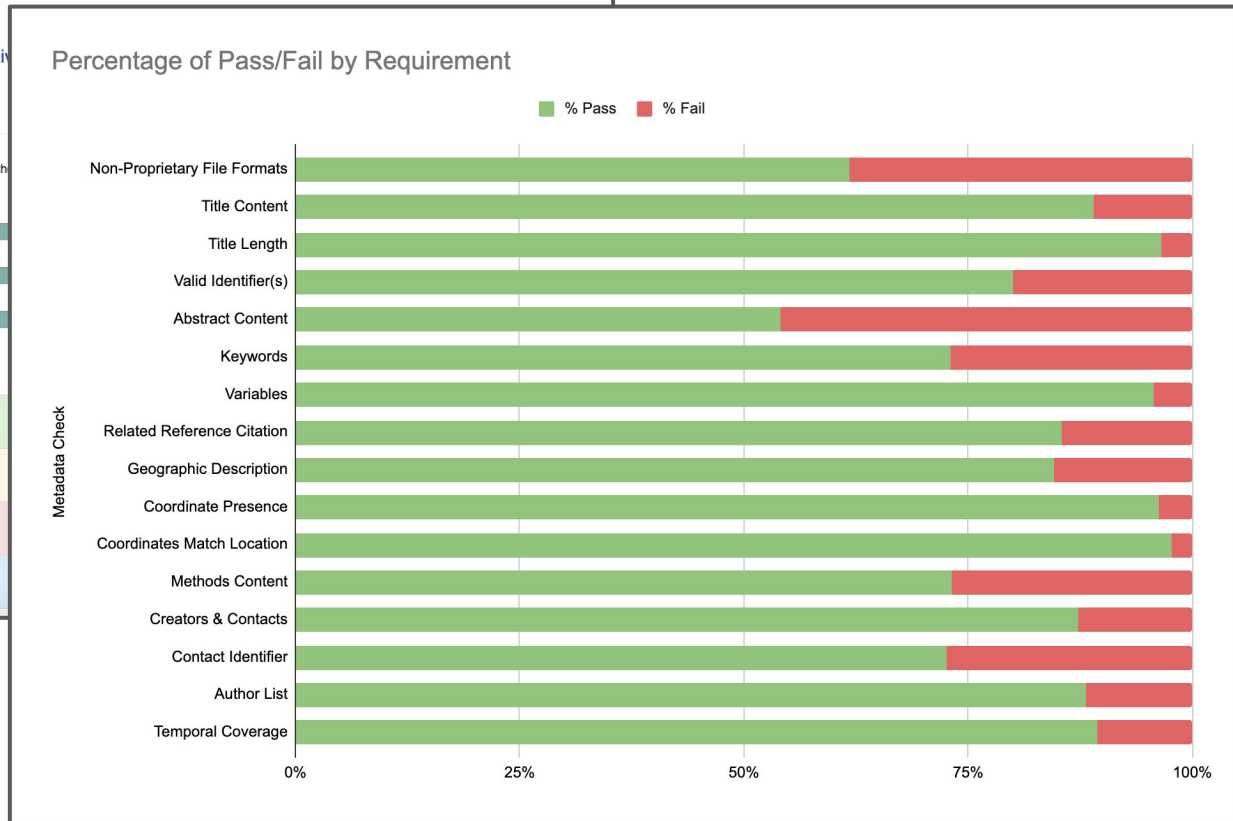
06

Authors

In the order of contributions

Metadata review



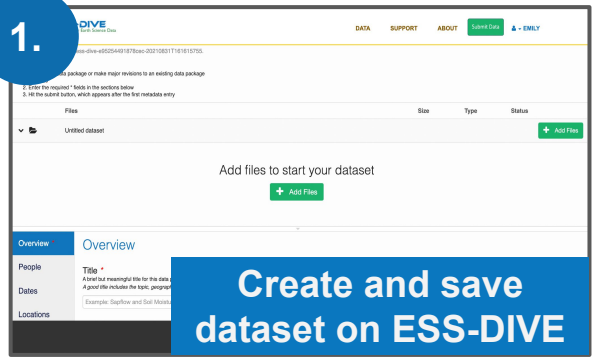


PUBLISHING DATASETS:
Getting started

Publication process

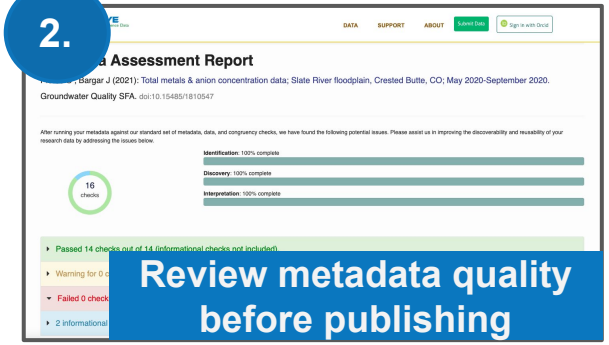
ESS-DIVE

1.



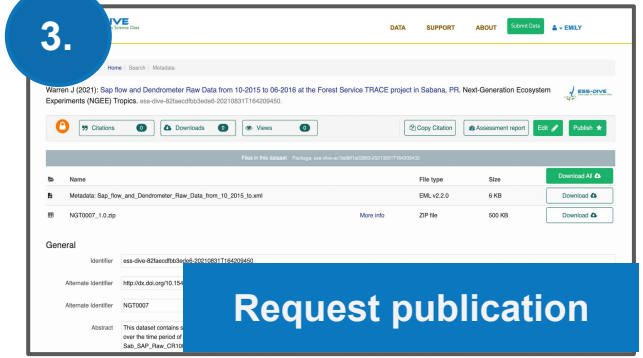
Create and save dataset on ESS-DIVE

2.



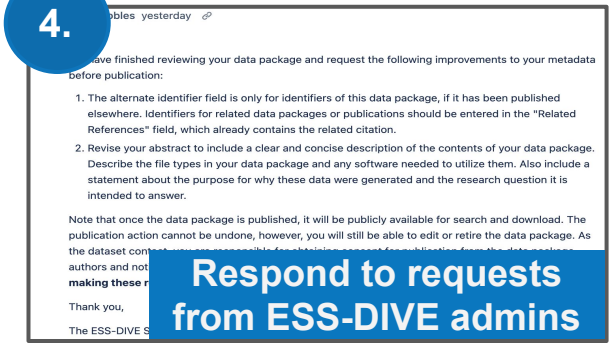
Review metadata quality before publishing

3.



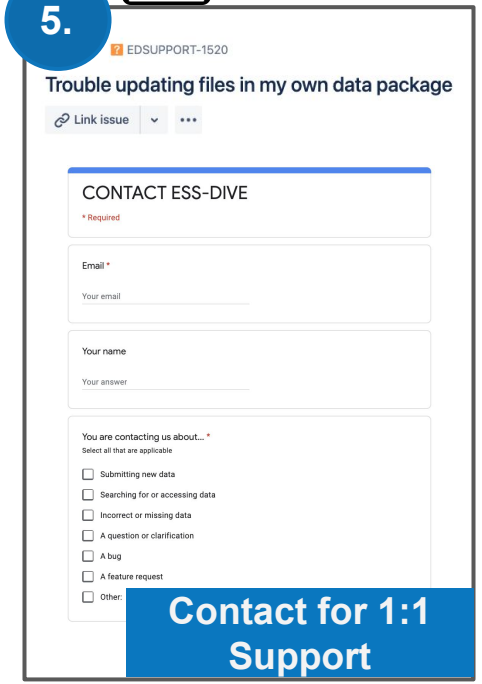
Request publication

4.



Respond to requests from ESS-DIVE admins

5.



Contact for 1:1 Support

Icons made by https://www.freepik.com

Multiple ways to create and edit data packages

- **Web Upload Form**
  - Manually enter data package metadata on the ESS-DIVE user interface (UI), one data package at a time. Drag and drop or select files from your file manager to upload data.
  - Interacting with the web form when editing a data package on ESS-DIVE
- **Package Service API**
  - A programmatic method for creating AND editing data package metadata on ESS-DIVE. Upload files and create metadata using JSON-LD.

Tools and resources for dataset creation

Help Documentation

Refer to **ESS-DIVE's website and Gitbooks** for detailed information on data package requirements

ess-dive.lbl.gov/
docs.ess-dive.lbl.gov/

Offline Metadata Guide

Collaborate on metadata with co-authors before working on ESS-DIVE

docs.ess-dive.lbl.gov/

Sandbox Testing Server

Practice uploading datasets to our sandbox test server, which does not permanently save data

data-sandbox.ess-dive.lbl.gov/

Support Email Service

Feel free to **contact the ESS-DIVE support team** through email for any questions

ess-dive.lbl.gov/contact/
ess-dive-support@lbl.gov

Using the CSV & file-level metadata reporting formats

What type of data do you typically collect/work with?

Are you familiar with data standardization practices?

Yes, very familiar

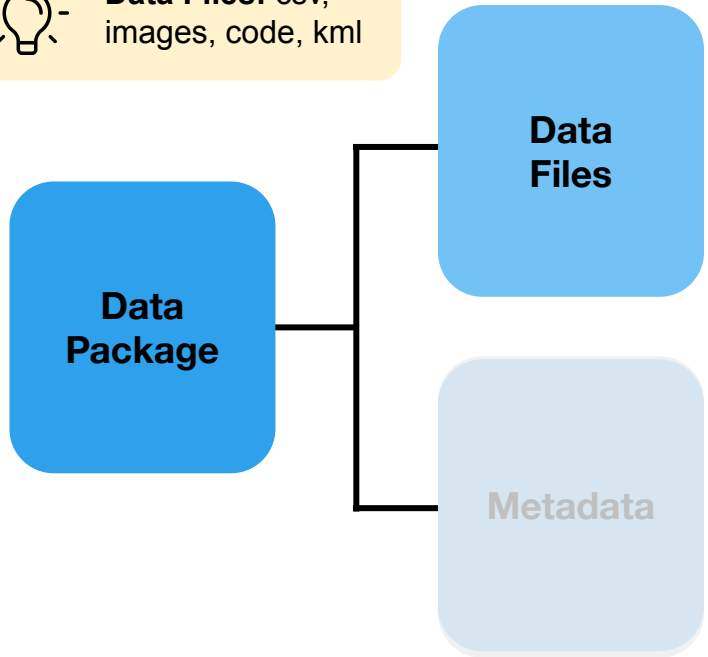
Somewhat

Not at all

When you collect data (e.g., in notebooks, spreadsheets, etc.) how do you decide how you will organize the data?

Focusing in on your **data files**

Data Files: csv, images, code, kml



*Focusing on ways of formatting your data files so that others can **find** and **reuse** the data files <u>within</u> your data package.*

Enable finding and reusing your data files

ESS-DIVE

CSV Reporting Format

Guidelines for formatting your tabular data

(column/row headers, temporal data, missing values)

Terri Velliquette

Jessica Welch

Michael Crow

Ranjeet Devarakonda

Susan Heinz

OAK RIDGE National Laboratory

The CSV Reporting Format

What is the format?

- The CSV is non-proprietary format for tabular data
- Archives tabular data in its simplest form
- Defines structure and some content

Why use the format?

- Specifies common format for file organization and elements within your CSV files (e.g., missing values) which make CSVs easier to read
- Reduces inconsistencies across datasets (e.g., 2021-26-04 vs. 4/26/2021)

The CSV Reporting Format

File structure

- Character set
- Delimiter
- Data Matrix
- Column or Row names

Naming Structure

- File Name
- Column or Row Names
- Units

Field Structure

- Consistent Values
- Missing Value Codes
- Temporal Data
- Temporal Data Range
- Spatial Data

https://ess-dive.gitbook.io/csv-file-structure-reporting-format/

Well-formatted Data File (viewed in Excel prior to CSV save)

thaw_and_water_depth_201007.csv

	A	B	C	D	E	F	G
1	area	plot_type	Latitude	Longitude	date	thaw_depth	water_table_depth
2	N/A	N/A	Decimal degrees	Decimal degrees	yyyy-mm-dd	cm	cm
3	Site 6	CLC1	71.29573	-156.66473	2010-07-07	35	0
4	Site 6	CLC2	71.29571	-156.66469	2010-07-07	38	1
5	Site 6	CLC3	71.2957	-156.66467	2010-07-07	35	0.5
6	Site 6	CLC5	71.28615	-156.59787	2010-07-07	-9999	-9999
7	Site 6	CLC6	71.28615	-156.59787	2010-07-07	-9999	-9999
8	Site 6	CLC7	71.28615	-156.59787	2010-07-07	-9999	-9999
9	N/A	DS1	71.29775	-156.66404	2010-07-10	67	-67
10	N/A	DS2	71.29774	-156.66397	2010-07-10	31	-31
11	N/A	DS3	71.29776	-156.66394	2010-07-10	43	-43
12	Beaver Road Mile 17	FC1	71.29461	-156.68819	2010-07-22	23	-5
13	Beaver Road Mile 17	FC2	71.29461	-156.68819	2010-07-22	27	-6.5
14	Beaver Road Mile 17	FC3	71.29461	-156.68819	2010-07-22	29	-7

Questions about the CSV format?

For more information: https://ess-dive.gitbook.io/csv-file-structure-reporting-format/

GitHub repository: https://github.com/ess-dive-community/essdive-csv-structure

Enable finding and reusing your data files

CSV Reporting Format

Guidelines for formatting your tabular data

(formatting column/row headers, temporal data, missing values)

Data Dictionary

A list of column headers you use in your datasets

(Definition, units, data type)

The Data Dictionary

What is a data dictionary?

- A spreadsheet where you list & define all the terms in your column header (e.g., variable names, units)

Why use a data dictionary?

- Researchers can have information about the variables in your data files
- Search interfaces can help users find the data they are looking for

Creating a Data Dictionary

Take the Column names and Units from your **CSV Data File**

	A	B	C	D	E	F	G	H	I	J
1	area	date	sampleID	Latitude	Longitude	sample_volume_collected	sample_type	Fl	Cl	Ca
2	N/A	yyyy-mm-dd	N/A	decimal degrees	decimal degrees	mg/L	N/A	mg/L	mg/L	mg/L
3	Area A	2018-07-22	A-1	71.2728	-156.7817	0.8	ISCO	0.0005	0.0013	0.0223
4	Area B	2018-07-23	B-1	71.5888	-157.7817	0.68	surface_grab	0.0001	0.002	0.0345
5	Area C	2018-07-24	C-1	71.9671	-157.0002	-9999	surface_grab	-9999	0.0006	0.0356

Creating a Data Dictionary

Enter them into your data dictionary

	A	B	C	D
1	**Column_Name**	**Unit**	**Definition**	**Column_Long_Name**
2	area	N/A	Name of the intensive field site within the project. Possible values: Area A, Area B, Area C	Field site name
3	date	yyyy-mm-dd	Date samples were collected in the field.	N/A
4	sampleID	N/A	Samples were collected in the field. Bags marked with sequential ID numbers.	Unique sample identifier
5	Latitude	decimal degrees	Latitude provided in WGS84	Latitude
6	Longitude	decimal degrees	Longitude provided in WGS84	Longitude
7	sample_volume_collected	mg/L	The volume of the sample collected.	The volume of the sample collected.

- We have **templates**: https://ess-dive.gitbook.io/file-level-metadata-reporting-format/csv_dd
- **Reuse** data dictionary when your datasets have same headers

Enable finding and reusing your data files

CSV Reporting Format

Guidelines for formatting your tabular data

(formatting column/row headers, temporal data, missing values)

Data Dictionary

A list of column headers you use

(Definition, units, data type)

File-level metadata

A list of all files that appear in your data package

(file description, date, latitude, longitude)

File-level metadata

What are file-level metadata?

- Granular information at the data file level (e.g., file name & description, start and end dates)

Why provide file-level metadata?

- Data users will have general understanding of info contained within a file
- FLMD can enable automatic parsing of data files so that users can eventually search & locate files across data collections

File-level metadata example

	A	B	C	D
1	File_Name	File_Description	Standard	UTC_offset
2	soil_samples_*.csv	15 soil samples taken in the summer of 2019 using small hand trowel and soil probe.	csv v1.0	- 5 hours
3	SoilPoreWaterHillslope2019.csv	50 soil pore water samples taken from the hillslope at the site over a one year period.	EPA	- 5 hours

- FLMD **template**: https://ess-dive.gitbook.io/file-level-metadata-reporting-format/
- Can use wildcard * to indicate when FLMD applies to multiple files

Summary

- **Organizing** data files and packaging them with relevant metadata
- **Data publication** process
- Reporting formats can make your data easier to **find** and **reuse**

Upcoming webinars

Sept 28 - Project portals and new portals discovery tool

October 26 - Standardizing data using ESS-DIVE reporting formats

November 30 - Dataset permission management

Thanks!

@ESS-DIVE

Join ESS-DIVE's Community Mailing List!

http://bit.ly/essdiveMailingList

Contact us at ess-dive-support@lbl.gov

Clarifying terminology: Data standards and reporting formats

- Data Standards - Decades of development, accredited by governing org.

Darwin Core

ISO

- Reporting Formats - Community-driven still enable data harmonization and synthesis