How to create your datasets and format data files for publication

Emily Robles
Rob Crystal-Ornelas
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.
Considerations to split up data packages

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

Campaign / Time Period
Data from a field campaign or season that need to be viewed together

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

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

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

Refer to ESS-DIVE reporting formats

Register as a contributor

Register to submit data ESS-DIVE

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

Publish with metadata

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**: 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
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
After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following issues with your research data by addressing the issues below:

- **Identified**: 100% complete
- **Discovery**: 100% complete
- **Interpretation**: 100% complete

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

Warning for 0 checks.

Failed 0 checks.

2 informational checks.
PUBLISHING DATASETS:
Getting started
Publication process

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

**Data Package**

**Data Files**

**Metadata**

Focusing on ways of formatting your data files so that others can **find** and **reuse** the data files within your data package.
Enable finding and reusing your data files

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

<table>
<thead>
<tr>
<th>File structure</th>
<th>Naming Structure</th>
<th>Field Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>● Character set</td>
<td>● File Name</td>
<td>● Consistent Values</td>
</tr>
<tr>
<td>● Delimiter</td>
<td>● Column or Row Names</td>
<td>● Missing Value Codes</td>
</tr>
<tr>
<td>● Data Matrix</td>
<td>● Units</td>
<td>● Temporal Data</td>
</tr>
<tr>
<td>● Column or Row names</td>
<td></td>
<td>● Temporal Data Range</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Spatial Data</td>
</tr>
</tbody>
</table>

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

thaw_and_water_depth_201007.csv

<table>
<thead>
<tr>
<th></th>
<th>area</th>
<th>plot_type</th>
<th>Latitude</th>
<th>Longitude</th>
<th>date</th>
<th>thaw_depth</th>
<th>water_table_depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>N/A</td>
<td>N/A</td>
<td>Decimal</td>
<td>Decimal</td>
<td>yyyy-mm-dd</td>
<td>cm</td>
<td>cm</td>
</tr>
<tr>
<td>3</td>
<td>Site 6</td>
<td>CLC1</td>
<td>71.29573</td>
<td>-156.66473</td>
<td>2010-07-07</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>Site 6</td>
<td>CLC2</td>
<td>71.29571</td>
<td>-156.66469</td>
<td>2010-07-07</td>
<td>38</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Site 6</td>
<td>CLC3</td>
<td>71.2957</td>
<td>-156.66467</td>
<td>2010-07-07</td>
<td>35</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>Site 6</td>
<td>CLC5</td>
<td>71.28615</td>
<td>-156.59787</td>
<td>2010-07-07</td>
<td>-9999</td>
<td>-9999</td>
</tr>
<tr>
<td>7</td>
<td>Site 6</td>
<td>CLC6</td>
<td>71.28615</td>
<td>-156.59787</td>
<td>2010-07-07</td>
<td>-9999</td>
<td>-9999</td>
</tr>
<tr>
<td>8</td>
<td>Site 6</td>
<td>CLC7</td>
<td>71.28615</td>
<td>-156.59787</td>
<td>2010-07-07</td>
<td>-9999</td>
<td>-9999</td>
</tr>
<tr>
<td>9</td>
<td>N/A</td>
<td>DS1</td>
<td>71.29775</td>
<td>-156.66404</td>
<td>2010-07-10</td>
<td>67</td>
<td>-67</td>
</tr>
<tr>
<td>10</td>
<td>N/A</td>
<td>DS2</td>
<td>71.29774</td>
<td>-156.66397</td>
<td>2010-07-10</td>
<td>31</td>
<td>-31</td>
</tr>
<tr>
<td>11</td>
<td>N/A</td>
<td>DS3</td>
<td>71.29776</td>
<td>-156.66394</td>
<td>2010-07-10</td>
<td>43</td>
<td>-43</td>
</tr>
<tr>
<td>12</td>
<td>Beaver Road Mile 17</td>
<td>FC1</td>
<td>71.29461</td>
<td>-156.68819</td>
<td>2010-07-22</td>
<td>23</td>
<td>-5</td>
</tr>
<tr>
<td>13</td>
<td>Beaver Road Mile 17</td>
<td>FC2</td>
<td>71.29461</td>
<td>-156.68819</td>
<td>2010-07-22</td>
<td>27</td>
<td>-6.5</td>
</tr>
<tr>
<td>14</td>
<td>Beaver Road Mile 17</td>
<td>FC3</td>
<td>71.29461</td>
<td>-156.68819</td>
<td>2010-07-22</td>
<td>29</td>
<td>-7</td>
</tr>
</tbody>
</table>
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**
Creating a Data Dictionary

Enter them into your data dictionary

<table>
<thead>
<tr>
<th>Column_Name</th>
<th>Unit</th>
<th>Definition</th>
<th>Column_Long_Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>area</td>
<td>N/A</td>
<td>Name of the intensive field site within the project. Possible values: Area A, Area B, Area C</td>
<td>Field site name</td>
</tr>
<tr>
<td>date</td>
<td>yyyy-mm-dd</td>
<td>Date samples were collected in the field.</td>
<td>N/A</td>
</tr>
<tr>
<td>sampleID</td>
<td>N/A</td>
<td>Samples were collected in the field. Bags marked with sequential ID numbers.</td>
<td>Unique sample identifier</td>
</tr>
<tr>
<td>Latitude</td>
<td>decimal degrees</td>
<td>Latitude provided in WGS84</td>
<td>Latitude</td>
</tr>
<tr>
<td>Longitude</td>
<td>decimal degrees</td>
<td>Longitude provided in WGS84</td>
<td>Longitude</td>
</tr>
<tr>
<td>sample_volume_collected</td>
<td>mg/L</td>
<td>The volume of the sample collected.</td>
<td></td>
</tr>
</tbody>
</table>

- 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

(formattting 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

<table>
<thead>
<tr>
<th>File_Name</th>
<th>File_Description</th>
<th>Standard</th>
<th>UTC_offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>soil_samples_* .csv</td>
<td>15 soil samples taken in the summer of 2019 using small hand trowel and soil probe.</td>
<td>csv v1.0</td>
<td>- 5 hours</td>
</tr>
<tr>
<td>SoilPoreWaterHillslope2019.csv</td>
<td>50 soil pore water samples taken from the hillslope at the site over a one year period.</td>
<td>EPA</td>
<td>- 5 hours</td>
</tr>
</tbody>
</table>

- **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!

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