ESS-DIVE and Urban Integrated Field Laboratories Data

data.ess-dive.lbl.gov
Presenters

Joan Damerow
Project Scientist

Emily Robles
Data Management Research Associate
Presentation Overview

- Overview of ESS-DIVE and our data;
- ESS-DIVE project data management features;
- Resources to organize and curate high-quality datasets;
  - Dataset organization
  - Overview of reporting formats and templates
- Discussion on Urban IFL Data Management Needs

Key Takeaway: Learn how you can use ESS-DIVE features and reporting formats for your project data management and publication
What Urban IFL project or ESS project do you work on?
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
Overview of ESS-DIVE
What is ESS-DIVE?

Data Repository to preserve expand access to and improve usability of Environmental Systems Science data

data.ess-dive.lbl.gov

Findable Accessible Interoperable Reusable

Watershed Sciences Terrestrial Ecology Coastal Systems

729 public datasets
Complex data in Earth and Environmental Science

- The data are **diverse** and interdisciplinary
- Work with Urban-IFL projects to understand data types and support publication

Varadharajan et al. (2021), Comp. & Geosci. [https://doi.org/10.1016/j.cageo.2021.105024]
How to upload and publish data on ESS-DIVE

Register w/ ORCID.
Organize **data packages**.

**Define metadata** for data package.

Submit to save privately.
When ready click **Publish**.

Upload **files** for data package.

Available for **public search/download**.

ESS-DIVE **review**. **DOI** assigned.

ESS-DIVE **review**.

**ess-dive.lbl.gov/archive/**
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
Web Interface Submission and Automated Review

Overview

Title *
A brief but meaningful title for this data package. A good title includes the topic, geographic location, dates, and scale of the data.
Example: SapFlow and Soil Moisture Raw sensor data, Jan 2016-Apr 2016, BR-M4: Manaus

Existing DOI and Alternate Identifiers
DOI and alternate identifiers of the data package if it has previously been published elsewhere. Enter as many identifiers as needed below.
Example: http://dx.doi.org/10.15486/NGT/X000000

Abstract *
A concise description of the purpose and content of this data package. A good abstract would provide users with adequate information to determine if the data are useful for their needs.
Example: Raw output from the data logger connected to 9 sapflow and 5 soil moisture sensors where the sensors were installed, and other installation/maintenance details. No data processing.

People

Dates

Locations

Methods

Creators

The main researchers involved in producing the data. These include authors, owners, originators and principal investigators. One or more creators is required. If none are entered, you will be set as the creator of this document. List creator names in this text box.

Creator #1
First name
Last name
Organizational name

Metadata Assessment Report


After running your metadata against our standard set of metadata, data, and conformance checks, we have found the following:

Assessment suite: ESS-DIVE Conformance

- Findable: 100% complete
- Accessible: 100% complete
- Reusable: 100% complete
- Interoperable: 100% complete

- Passed 19 checks out of 19 (informational checks not included).
- Warning for 0 checks.
- Failed 0 checks.

https://ess-dive.lbl.gov/webinars/
Downloads, Citing and Reusing Data

Goldman A E ; Chu R K ; Danczak R E ; Deley R A ; Fansler S ; Garayburu-Caruso V A ; Graham E B ; McCall M L ; Ren H ; Renteria L ; Resch C T ; Tafari M ; Tolic N ; Torgeson J M ; Toyota J G ; Wells J ; Wrighton K C ; Stegen J C ; WHONDRS Consortium T (2020): WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Sediment FTICR-MS, NPOC, and Aerobic Respiration. River Corridor and Watershed Biogeochemistry SFA Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS). doi:10.15485/1729719

Files in this dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>File type</th>
<th>Size</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHONDRS_Summer_2019_Sampling_Campaign_Global_River_Corridor_Sediment_FTICR_MS_NPOC_and_Aerobic_Respiration.xml</td>
<td>EML v2.1.1</td>
<td>64 KB</td>
<td>Download</td>
</tr>
<tr>
<td>WHONDRS_S19S_SitePhotos.zip</td>
<td>More info</td>
<td>ZIP file</td>
<td>2 GB</td>
</tr>
<tr>
<td>WHONDRS_S19S_Sediment_v2.zip</td>
<td>More info</td>
<td>ZIP file</td>
<td>68 MB</td>
</tr>
</tbody>
</table>

General

Identifier: ess-dive-67e6bbdd9f83ad60-20210430T045323206818

Abstract

The WHONDRS Summer 2019 Sampling (S19S) Campaign collected samples in 97 globally distributed river corridor systems between July and September 2019. Surface streamed sediments were collected at three locations within each site (upstream, midstream, and downstream). Surface water was collected at the downstream site. This dataset includes a portion of the data types produced from the sediment samples and does not include any results from the surface water. S19S surface water data can be found at https://data.ess-dive.lbi.gov/view/doi/10.15485/1603775. Future datasets from this study will include geochemical, hydrologic, and microbial data from the surface water and sediment. The S19S campaign was designed with the science community to ask questions associated with links among core/transient metabolomes, microbial metabolism, biogeochemical function, and physical properties of watershed and river corridor systems. This dataset contains (1) high resolution characterization of dissolved organic matter from sediment via 12 Tesla Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS) through the Environmental Molecular Sciences Laboratory (EMSL); https://www.pnnl.gov/environmental-molecular-sciences-laboratory); (2) non-purgeable organic carbon (NPOC); (3) respiration rates calculated from laboratory incubations; (4) site photos; and (5) metadata. The field metadata file (WHONDRS_S19S_metadata_v3.csv) is the same as the field metadata included in the S19S surface water data packages but has had extra columns added and is listed as version 1. The
Supporting and Engaging the ESS Community

WEBSITE, TWITTER, COMMUNITY EMAIL LIST

ESS-DIVE Bootcamp
April Webinar
April 27, 2020 10:00-11:00

WEBINARS

DOCUMENTATION

ESS-DIVE Docs
Welcome
Accessing & Searching Data
New User Registration
Data & Metadata Upload
Portals
Frequently Asked Questions
Contact us
07
ESS-DIVE Data Portal
Quick Start Guide

TUTORIALS/ANNUAL WORKSHOP

HELP DESK

WEBINARS

Launching an Accessible Archive of Environmental Data
A new digital archive enables community use of terrestrial and subsurface ecosystem data sets.

COMMUNITY GITHUB

PUBLICATIONS

1:1 MEETINGS/SITE VISITS

WEBINARS
Overview of Data Management Features
Overview of Data Management Features

- **Dataset Sharing**: Collaborate
- **Data Portals**: Create a collection of datasets
- **External Linking**: Link to other repos and data systems
- **Project Data Management**: Additional project features
What is dataset sharing?

Review and Edit datasets with registered team members.
Dataset Sharing Permission Types

**Viewer**
- view private datasets

**Editor**
- view private datasets
- edit metadata
- add/remove files
- publish datasets

**Manager**
- view private datasets
- edit metadata
- add/remove files
- publish datasets
- share datasets with others
Sharing datasets - PIs and data managers

**Access and report on project publication progress**

- **View public and private datasets** generated by the project
  - Verify datasets are being published on time
  - Track how many datasets are on ESS-DIVE
  - Ability to review dataset content before publication

- **Manage datasets** from your project that were originally created or edited by team members
What is a data portal?

A data portal is a collection of any ESS-DIVE datasets. Data portals allow you to easily highlight and share related datasets.
What’s inside a Data Portal?

Collection of Datasets

Custom Descriptive Pages

Metrics: Summary of Holdings, View/Download/Citation Metrics…
Why use data portals?

- **Increase Findability** of project datasets
- Elevate public project identity in ESS-DIVE
- Collect **cumulative metrics** across all project datasets
- Link to project data portals in annual reports and paper publications
Your project data portal will be listed on ESS-DIVE’s portal page. Anyone from the public can access this page and locate your data portal.
Portal Sharing

* When granting a registered team member access to edit your portal, the same permission types as dataset sharing are available: view, edit, and manage.
What is External Linking?

Data **stored on other repositories** can be connected to ESS-DIVE metadata **without needing to upload** it directly to the dataset.
What does external linking look like?

External Links to Data or Metadata

<table>
<thead>
<tr>
<th>Description</th>
<th>Relationship</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>doi:10.6073/pasta/388c70359c8d0cd79dc3875d6510071d</td>
<td>[same as] Original publication of this dataset (where the data+metadata can be found)</td>
<td><a href="https://doi.org/10.6073/pasta/388c70359c8d0cd79dc3875d6510071d">https://doi.org/10.6073/pasta/388c70359c8d0cd79dc3875d6510071d</a></td>
</tr>
</tbody>
</table>
Upcoming: Project Spaces to Enable Project Data and User Management, Team Collaboration

Desired Features from User Interviews

- Creating project data **collections**
- Project **administration** of users and **curation** of data packages
- **Internal sharing within project teams** to promote collaboration

*Expected Early 2023!*
It’s important to keep project information updated on ESS-DIVE

- **Project information may change over time.** The default project information provided to ESS-DIVE is when the funding was awarded.

- Don’t let this affect your data! Inconsistent *project titles* can lead to incorrect data citations and make it difficult to manage project data.

- **We can help!** PIs or project managers can send a request to ess-dive-support@lbl.gov to update project information. Look out for further announcements from ESS-DIVE asking projects to update information.
ORGANIZING YOUR DATA
Best practices, considerations, & requirements
Components of a dataset

Dataset

Data Files: csv, images, code, kml

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
Choosing data to publish

Data related to paper publications
- Data that are used to create figures/tables and data that support the conclusions of paper and validate results

Data collected as part of a project
- Meet BER and other funding requirements by making data publicly available in ESS-DIVE

Publishing *scientifically useful data*
- Raw data, processed data, code, model inputs/outputs, or products built off of data

Considerations to divide data into datasets

**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

**Size of Data**
Could impact upload method, usability/accessibility

**Data Type & Use**
Particular data type from a project - e.g. continuously generated sensor data, sample data, data synthesis product (raw data vs processed)

https://dataoneorg.github.io/Education/bestpractices/decide-what-data
Reporting formats enable discovery and reuse

ESS-DIVE Reporting Formats ensure that

- Researchers can **find** data more easily in ESS-DIVE
- Data can be **reused** in synthesis research
- **Integrated search** that facilitates data interoperability

*Benefits for data contributors, collaborators and project members, and data users.*
Choose high-level reporting formats, *then* domain-specific where applicable.

Package-level metadata is **required** for all ESS-DIVE datasets.
File-level metadata (FLMD)

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 each file
- FLMD can enable automatic parsing of data files so that users can eventually search & locate files across data collections

FLMD documentation: https://ess-dive.gitbook.io/file-level-metadata-reporting-format/
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>

Can use wildcard * to indicate entry that applies to multiple files

**FLMD template:** https://ess-dive.gitbook.io/file-level-metadata-reporting-format/
CSV Reporting Format

What is the CSV reporting format?

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

Why use the CSV reporting format?

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

CSV documentation: https://ess-dive.gitbook.io/csv-file-structure-reporting-format/
Documentation and Community Space

GitHub: https://github.com/ess-dive-community

Website version: https://docs.ess-dive.lbl.gov/contributing-data/data-reporting-formats
How to Connect With Our Team

To get help:
Webinars: https://ess-dive.lbl.gov/webinars/
Guide to using ESS-DIVE: docs.ess-dive.lbl.gov
Support email: ess-dive-support@lbl.gov

To stay updated:
ess-dive-community@lbl.gov
@essdive

Acknowledgements
Advisory Groups: ESS-DIVE Archive Partnership Board, ESS Cyberinfrastructure Working Groups
Funding: EESSD Data Management
Questions?
Urban IFL Data Management Needs
What data types are you planning to collect?
Approximately how many datasets do you anticipate publishing across the project in a year?
What challenges do you face when creating datasets, or what challenges do you anticipate as a new user?
Urban IFL Data Management Needs

1. Are you interested in using existing standards or formats for your data (e.g. any of the ESS-DIVE reporting formats)? If yes, which ones?
2. Considering the data type responses, is there need to develop standard formats for any specific data type, particularly relevant across the Urban IFL projects?
3. Who manages and publishes data on your project (e.g. how many team members, and what is their role on the project)?
4. What tools are you planning to use for data management (e.g. any of the ESS-DIVE features mentioned in the presentation, or others)?
5. What challenges have you encountered or anticipate in managing data for your project, team or individual research?