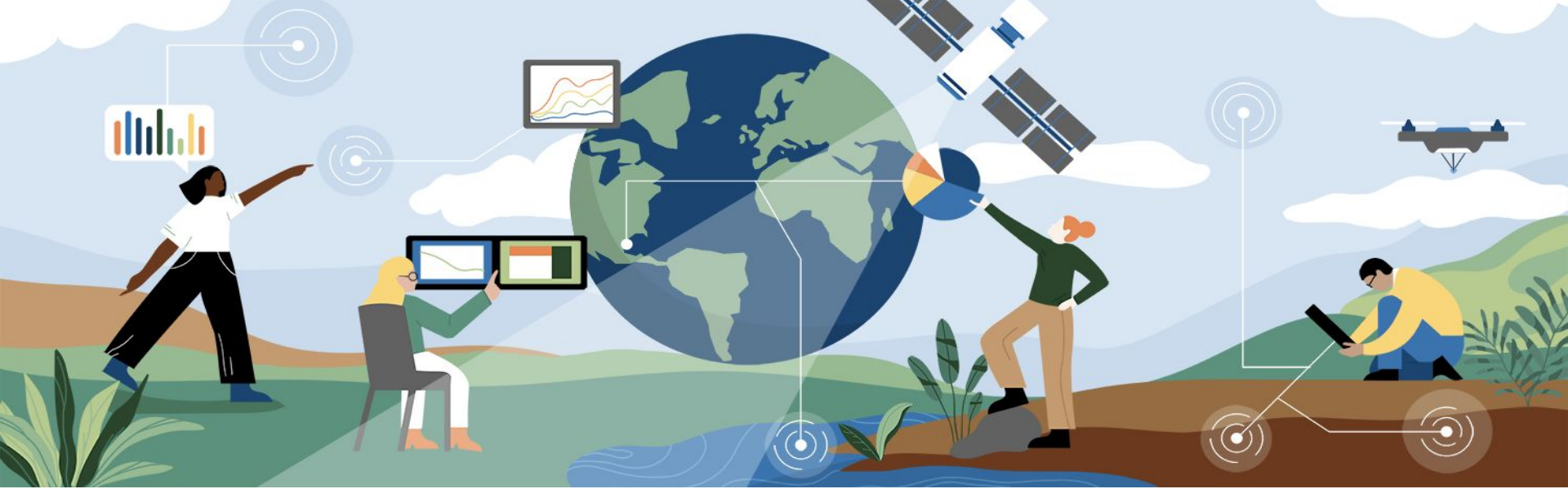


ESS-DIVE File-level Metadata Reporting Format

Status and version 2.0 changes

Emily Robles, *Senior Research Associate*

What is your familiarity with the ESS-DIVE Reporting Formats?

No familiarity

0%

Have heard of them, but have not used them

0%

Plan to use at least one in the future

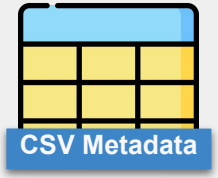
0%

Have used one or more in datasets

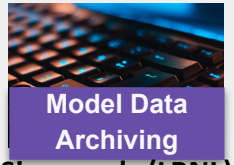
0%

ESS-DIVE Reporting Formats

Data Files



Velliquette, Heinz, Devarakonda (ORNL)

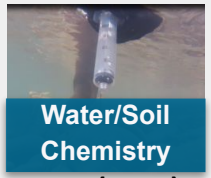


Simmonds (LBNL)

Bond-Lamberty, Pennington (PNNL)

Rogers, Ely (BNL)

Goldman (PNNL)



Boye (SLAC)

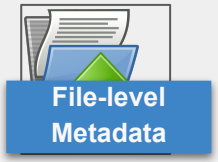
Weisenhorn (ANL)

Serbin, Ely (BNL)

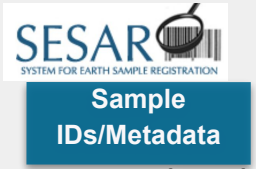
Metadata

Agarwal, Hendrix (LBNL)

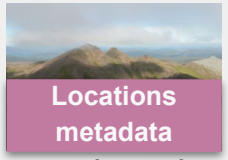
Damerow (LBNL)



Velliquette, Heinz, Devarakonda (ORNL)



Damerow (LBNL)

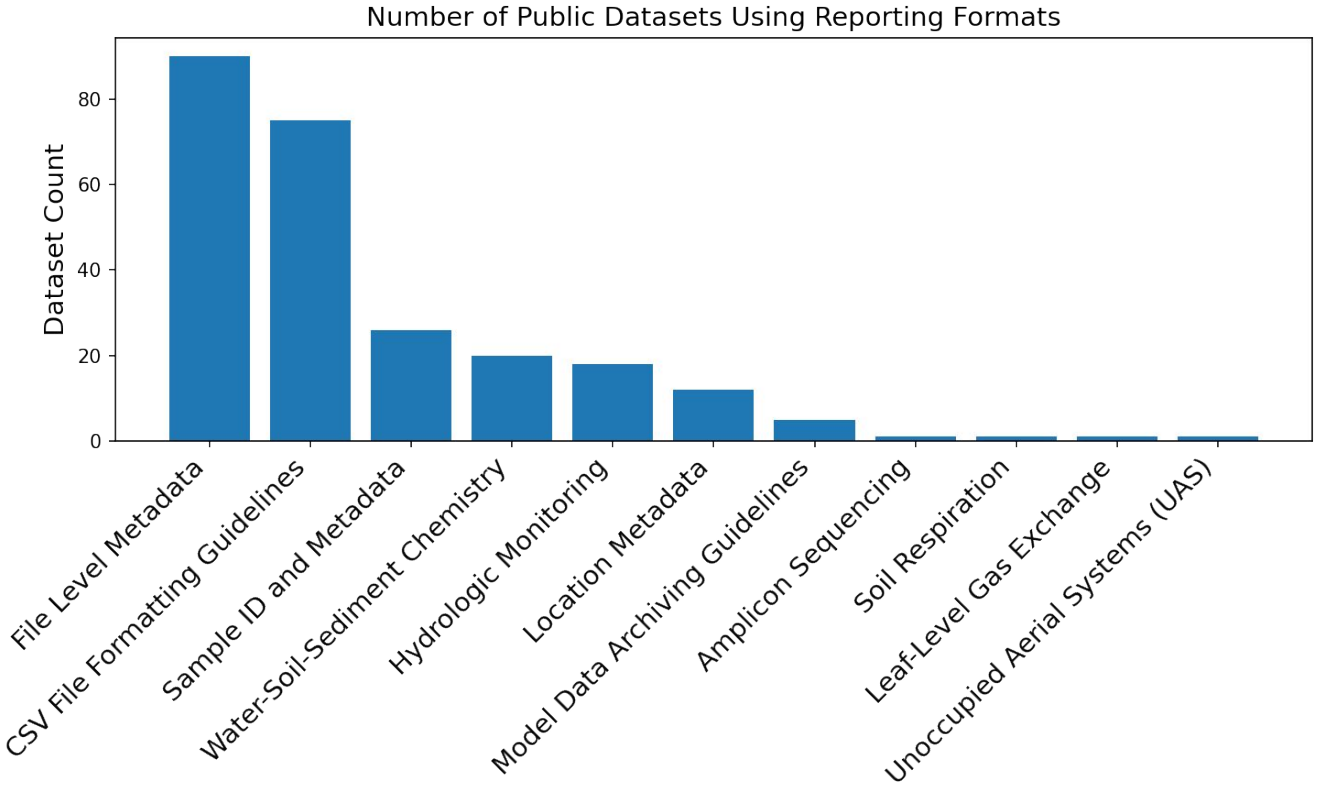


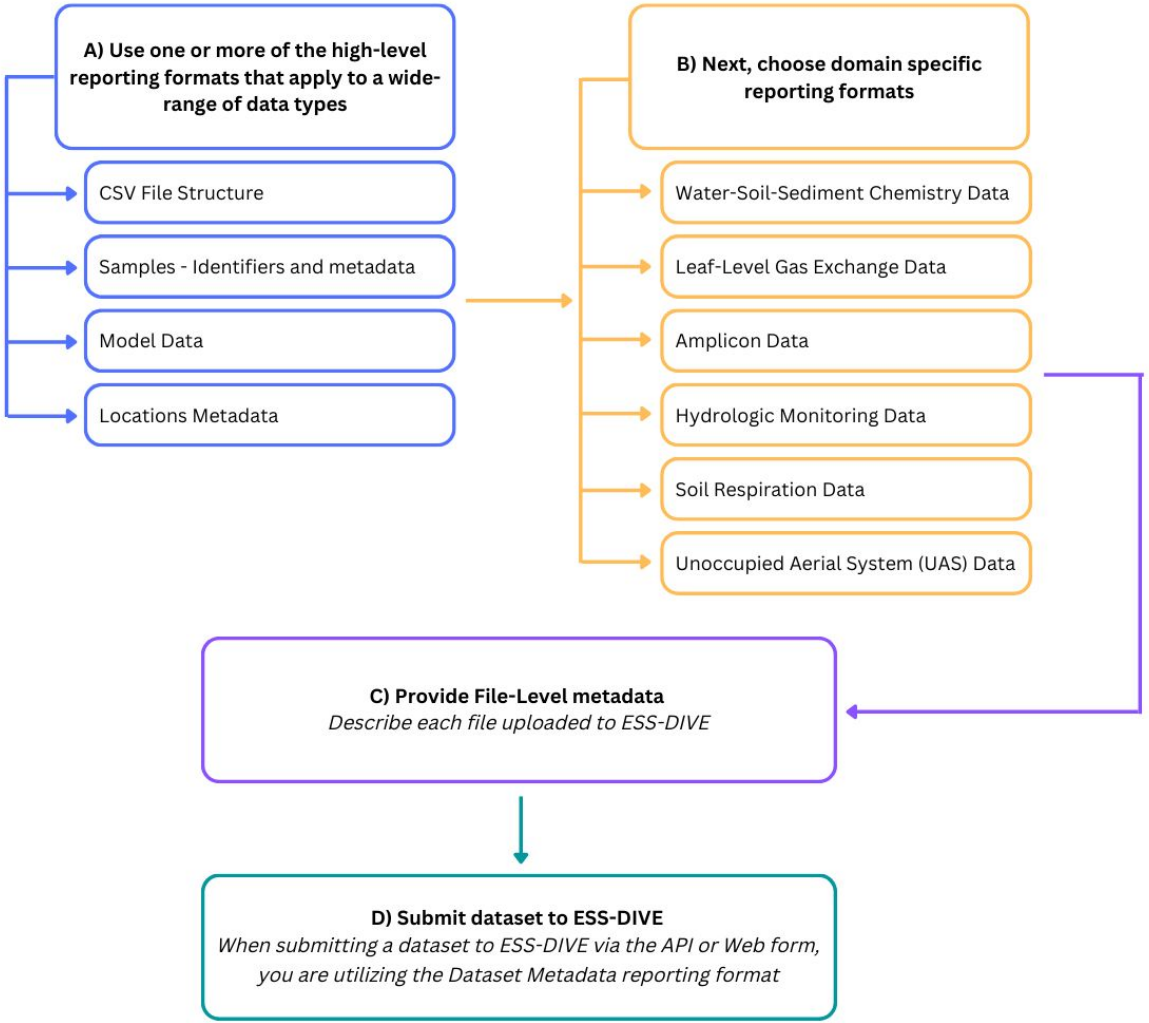
Crystal-Ornelas (LBNL)

Crystal-Ornelas, R. et al. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. Sci Data 9, 700 (2022). https://doi.org/10.1038/s41597-022-01606-w

Status of Reporting Format Datasets

90 datasets using reporting formats are publicly available





ESS-DIVE

File-level metadata can be used to describe files in any dataset, regardless of data type

The Role of File-level Metadata

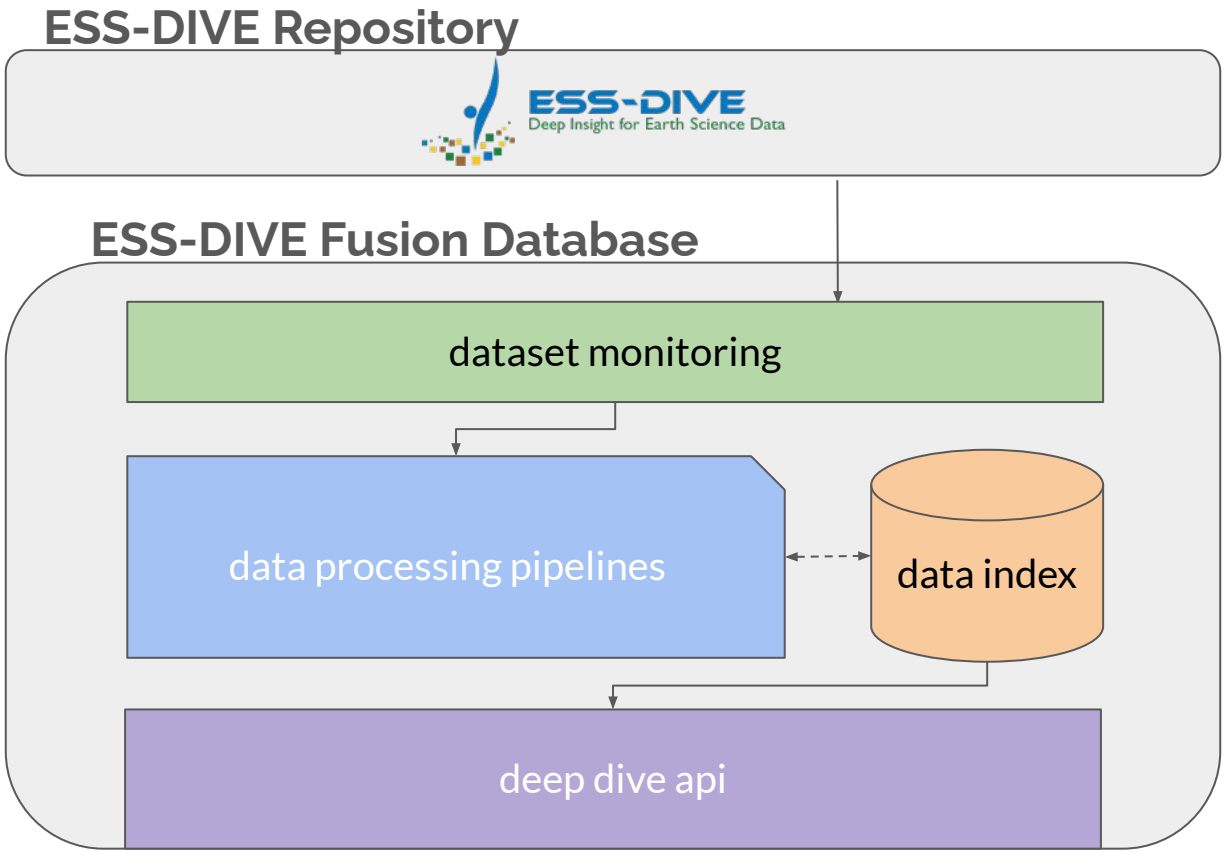
File-level metadata provides the information needed to understand, parse, and extract data from files. It consists of two files:

1. File-level Metadata File (FLMD)
    - Each row contains information about a **file** within the dataset

2. Data Dictionary File (dd)
    - Each row contains the **header row/column information** of **individual files**

The ESS-DIVE Fusion DB uses the FLMD reporting format to parse CSV files and enable advanced search capabilities

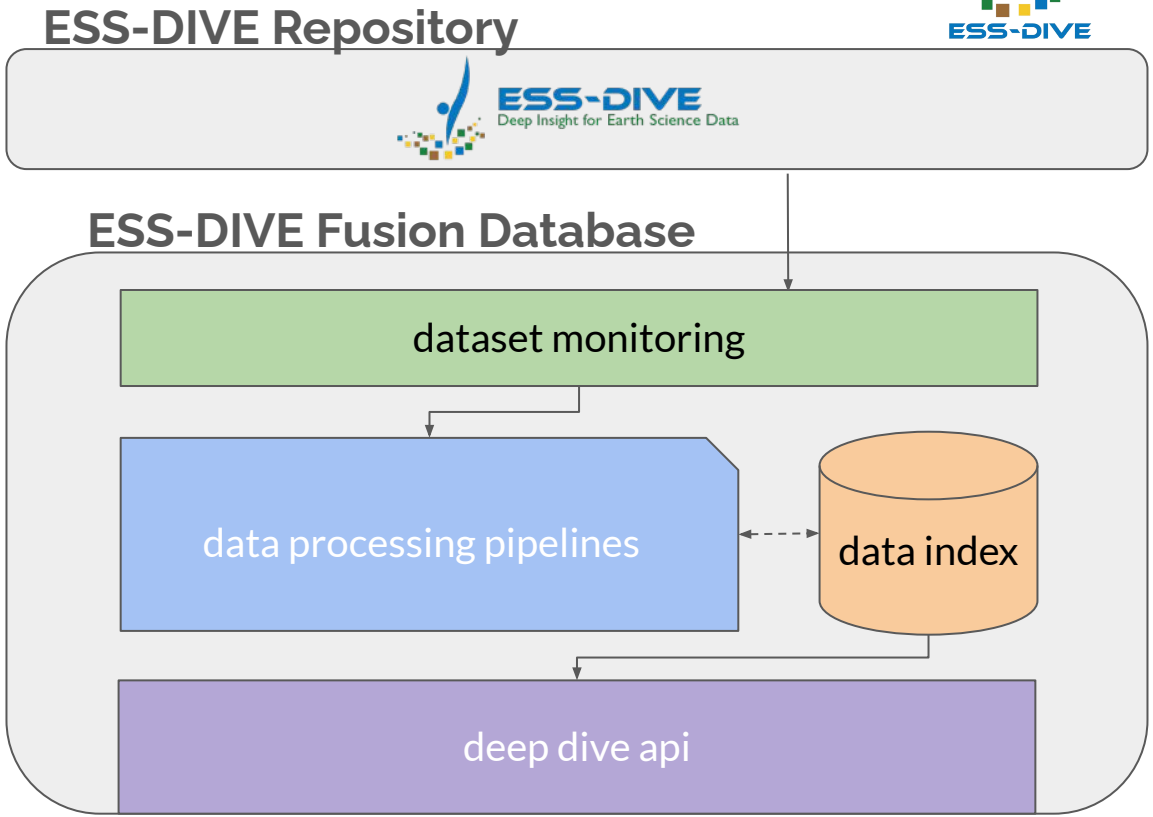
ESS-DIVE Fusion DB



ESS-DIVE Fusion DB

Dataset monitoring watches for

- datasets in publication review that use reporting formats
- newly published data ready for deep dive api



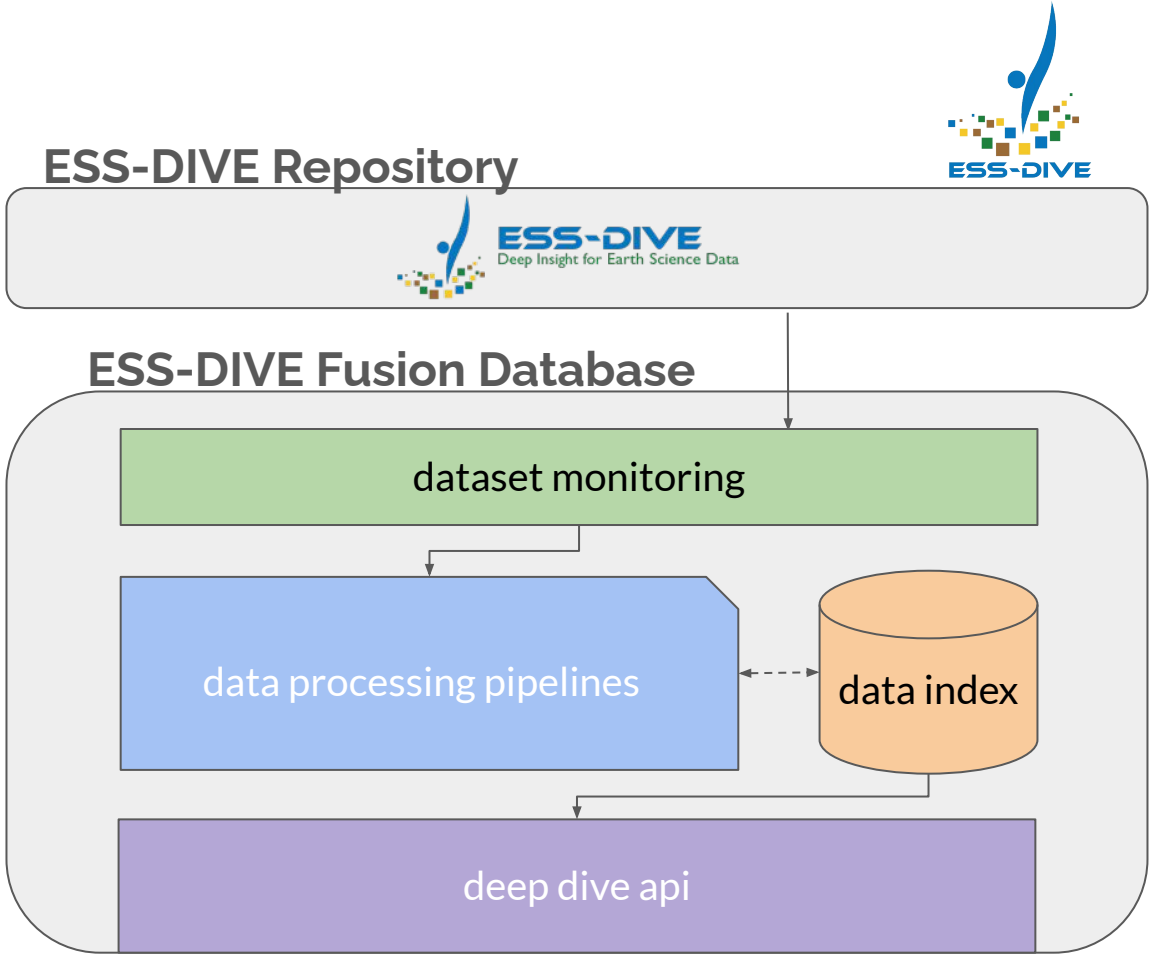
ESS-DIVE Fusion DB

Dataset monitoring watches for

- datasets in publication review that adopt reporting formats
- newly published data ready for deep dive api

Data processing pipelines

- ***validate*** reporting formats
- ***parse*** dataset files
- ***report*** feedback for publication review*
- ***prepare*** data for deep dive api



** Used internally by ESS-DIVE Support*

ESS-DIVE Fusion DB

Dataset monitoring watches for

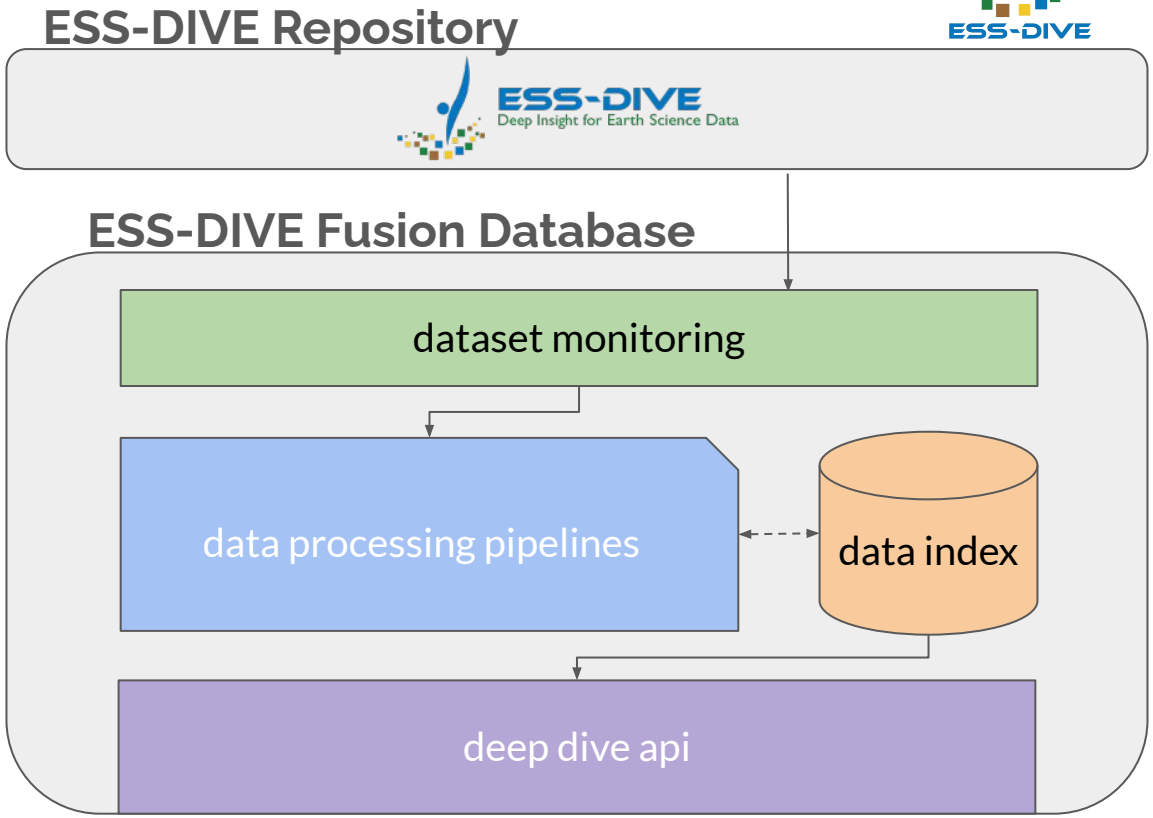
- datasets in publication review that adopt reporting formats
- newly published data ready for deep dive api

Data processing pipelines

- ***validate*** reporting formats
- ***parse*** dataset files
- ***report*** feedback for publication review*
- ***prepare*** data for deep dive api

Data index - stores results of all data processing pipelines.

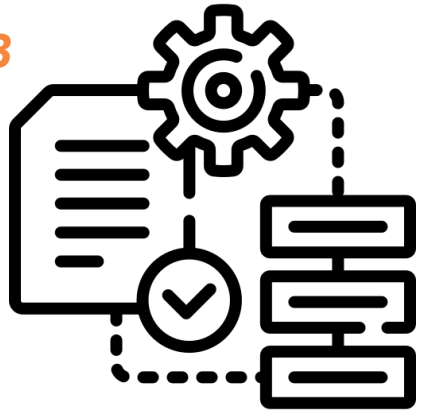
Deep dive API - public API access for use in discovery of indexed data



** Used internally by ESS-DIVE Support*

Status of Reporting Format Datasets

20 datasets are being successfully parsed by the Fusion DB



Common Errors

- Incorrect naming of required FLMD and DD fields
- Parsing issues in CSVs: variables with spaces or special characters, UTF-8 errors
- Incorrect data orientation defined in FLMD

Errors are communicated to the data contributor during the publication review process

Publication Review with Reporting Formats

Request Publication

01

Review your dataset's **Automated Assessment Report.**

Request publication.

Receive **confirmation email** from ESS-DIVE team.

Review and Revise

02

Receive revision request email from ESS-DIVE

Revise based on recommendations and **respond once finished**.

If no revisions are needed, your dataset will be published.

Dataset Published

03

Receive final confirmation message with your dataset DOI and citation.

Dataset is now publicly available on data.ess-dive.lbl.gov

Revise at any time.

Publication Review with Reporting Formats

File-level review is a new component of the publication workflow when a dataset is using reporting formats

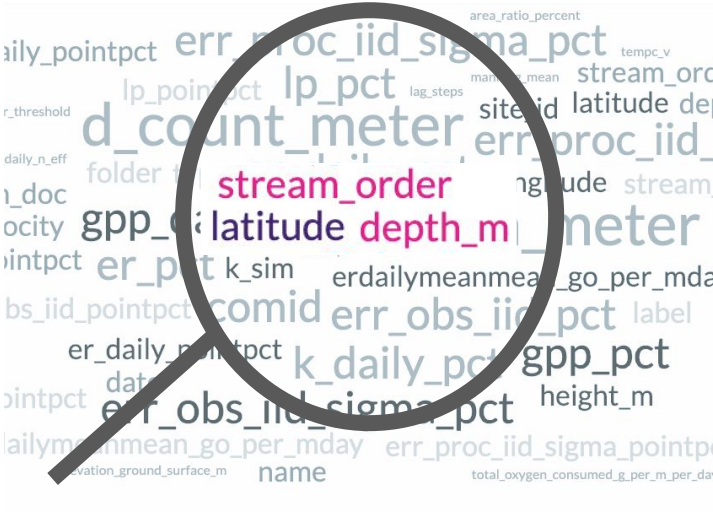
- Review occurs at the time of publication request, ***along with standard dataset review***
- Validation results from Fusion DB inform feedback
- File-level feedback provided with all other revision requests if relevant

Published Datasets Enable Enhanced Search

The Fusion DB enables enhanced search through the **Deep Dive API**

- Separate from ESS-DIVE main search
- Searches data **within dataset files**

Published datasets that employ reporting formats are instrumental to enabling advanced search



Interactive API at fusion.ess-dive.lbl.gov

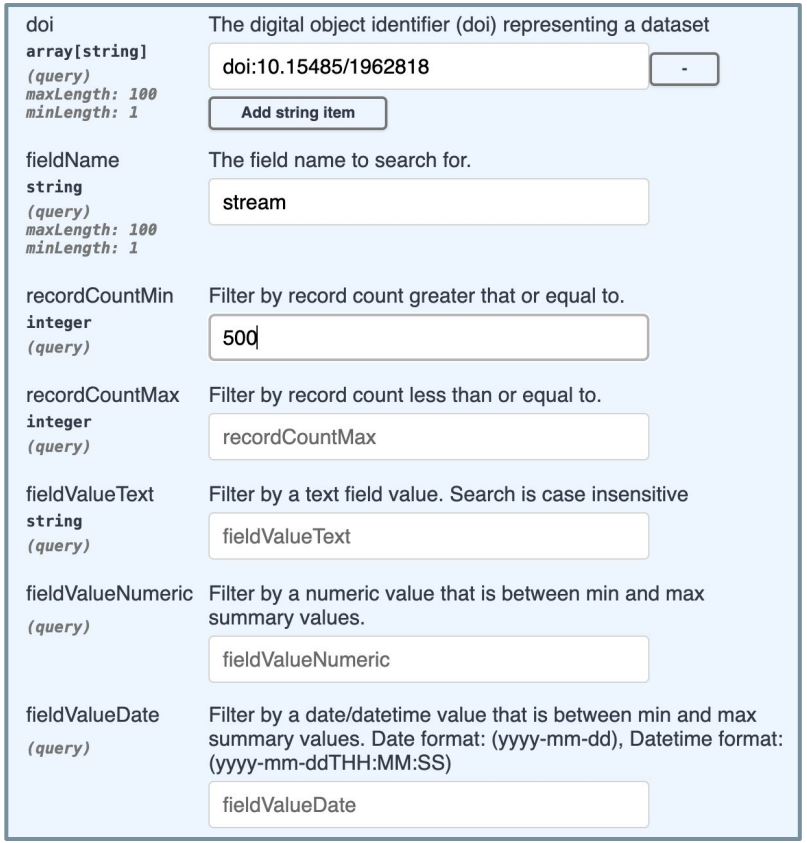
Deep Dive Search

Search Parameters

- DOIs
- Field Name
- Record count
- Field Value - text, numeric, date(time)

Example

- fieldName “site” & fieldValueText “S22RR”
  - Returns data from two datasets
    doi:10.15485/1987520
    doi:10.15485/1999774



Questions?

File-level Metadata Version 2.0

Goal of v2 Revisions

- Make the file-level metadata reporting format ***easier to follow***
- ***Reduce clutter*** from optional fields and redundancy
- Provide ***clarification in documentation*** where needed
- Improve ***harmonization*** across reporting formats
- Increase number of datasets ***successfully parsing***

V2.0 changes have been heavily informed by community feedback from early adopters, in addition to Fusion DB validation

FLMD v1 Requirements and Recommendations

Required

File Name
- Name of associated file

File Description
- Brief description of file that distinguishes it from other files
- Information about data type

Recommended

Standard
- State if any data or metadata standard was applied to the data file (including reporting formats)

UTC Offset
- Report the Local Standard Time offset or time zone

Optional

- File Version
- Contact
- Date Start / Date End
- Coordinates
- Latitude / Longitude
- Missing Value Codes
- Data Orientation
- Notes

FLMD v1.1 Requirements and Recommendations

Required

File Name

- Name of associated file

File Description

- Brief description of file that distinguishes it from other files
- Information about data type

Recommended

Standard

- State if any data or metadata standard was applied to the data file (including reporting formats)

Optional

- **Header rows**
- **Column or row name position**
- File Version
- Data Orientation
- Notes
- ***Missing Value Codes****

**** moving to DD file***

NEW FLMD Header Rows and Position

Optional FLMD fields allow for handling of additional header rows/columns before ***and*** after the column or row names

Information before headers					
Additional information					
One more row of information					
leaf	**date**	**measurement_time**	**conductance**	**temperature**	**leaf_sensor**
1	2017-06-23	7.0	318.8	37.3	48.5
2	2017-06-23	7.0	277.1	35.9	62.8
3	2017-06-23	7.0	267.3	36.1	62.9
4	2017-06-23	7.0	200.5	36.2	68.0

leaf	**date**	**measurement_time**	**conductance**	**temperature**	**leaf_sensor**
Additional information before data					
Another row of information before data					
1	2017-06-23	7.0	318.8	37.3	48.5
2	2017-06-23	7.0	277.1	35.9	62.8
3	2017-06-23	7.0	267.3	36.1	62.9
4	2017-06-23	7.0	200.5	36.2	68.0

NEW FLMD Header Rows and Position

header_rows

- Used when rows after variable/header names and before data
- Provide the number of header rows that occur after the column or row names in a file and before the data begins

leaf	date	measurement_time	conductance	temperature	leaf_sensor
Additional information before data					
Another row of information before data					
1	2017-06-23	7.0	318.8	37.3	48.5
2	2017-06-23	7.0	277.1	35.9	62.8
3	2017-06-23	7.0	267.3	36.1	62.9
4	2017-06-23	7.0	200.5	36.2	68.0

NEW FLMD Header Rows and Position

column_or_row_name_position

- Used when rows before variable/header names
- Provide the row or column number that contains the header names. If not included it will be assumed that header names are in row 1 (horizontal orientation) or column 1 (vertical orientation)

Information before headers					
Additional information					
One more row of information					
leaf	**date**	**measurement_time**	**conductance**	**temperature**	**leaf_sensor**
1	2017-06-23	7.0	318.8	37.3	48.5
2	2017-06-23	7.0	277.1	35.9	62.8
3	2017-06-23	7.0	267.3	36.1	62.9
4	2017-06-23	7.0	200.5	36.2	68.0

FLMD Standard

Each Reporting Format has a ***package level keyword*** and a term to be used in the ***FLMD standard field***

- New set terms for the ESS-DIVE reporting formats to be used in the FLMD standard field

Keyword	FLMD Standard
ESS-DIVE Amplicon Sequencing Reporting Format	ESS-DIVE Amplicon v1
ESS-DIVE Location Metadata Reporting Format	ESS-DIVE Location v1
ESS-DIVE Hydrologic Monitoring Reporting Format	ESS-DIVE Hydrologic Monitoring v1
ESS-DIVE Sample ID and Metadata Reporting Format	ESS-DIVE Sample v1
ESS-DIVE CSV File Formatting Guidelines Reporting Format	ESS-DIVE CSV v1

Variable Naming Convention: Snake Case

Fusion DB reported validation errors when variable capitalization was inconsistent with RF standard

FLMD v1.0

File_Name	File_name	file_Name	file_name

FLMD v1.1 - snake case strongly recommended

file_name	File_Name	file_Name	File_name

A Note on Data Orientation

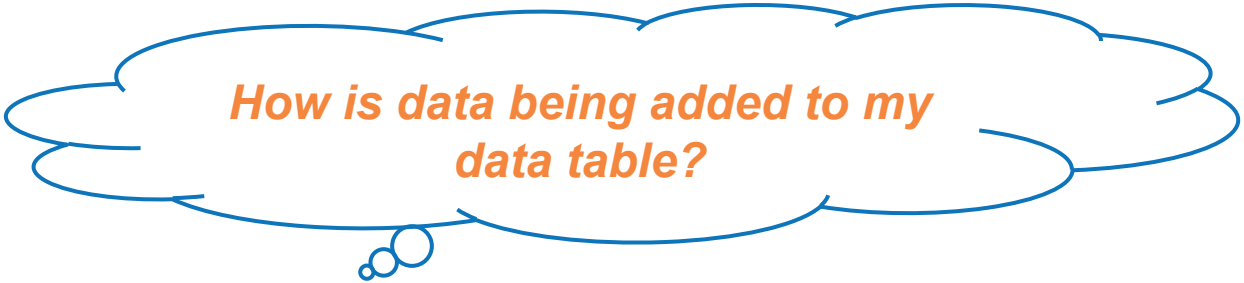
Two options for noting data orientation for CSV files within the file-level metadata:

1. Horizontally with Names at the top of each column

OR

2. Vertically with Names at the start of each row.

A Note on Data Orientation



New row = Horizontal orientation

area	plot_type	Latitude	Longitude	year	CH4_flux
Site 6	CLC1	71.29573	-156.66473	2010-07-07	91.8
Site 6	CLC2	71.29571	-156.66469	2010-07-07	54.3

New column = Vertical Orientation

area	Site 6	Site 6	Site 6
plot_type	CLC1	CLC2	CLC3
Latitude	71.29573	71.29571	71.2957
Longitude	-156.66473	-156.66469	-156.66467
year	2010-07-07	2010-07-07	2010-07-07
CH4_flux	91.8	54.3	63.9

Timeline for Finalizing FLMD v2.0

Feedback until April 5 → **Revise documentation** → **Finalize and publish**

Leading hands-on tutorial session during CI Working Group meeting!

Discussion

- What challenges have you come across when using reporting formats? (FLMD or other)
- Are there tools you feel could address these challenges? Do you use any tools currently?
- How do you feel about the version 2.0 changes - are there additional revisions you would like to see?

Connect With ESS-DIVE

To get help:

ess-dive.lbl.gov

ess-dive-support@lbl.gov

docs.ess-dive.lbl.gov

To stay updated:

ess-dive-community@lbl.gov

@essdive

https://bit.ly/essdiveMailingList

Acknowledgements

Advisory Groups: ESS-DIVE Archive Partnership Board, ESS Cyberinfrastructure Working Groups

Funding: EESSD Data Management