ESS-DIVE Overview
2024 ESS CI Meeting
Shreyas Cholia & Charu Varadharajan

April 15, 2024

1

Presentation Overview

- ESS-DIVE feature updates

- What's next: Upcoming plans

- Community funds

- Discussion about ESS-DIVE's future plans

Key Takeaways: Learn about new ESS-DIVE features and discussion on our future plans/community funds

What is ESS-DIVE?

Data Repository to

preserve

expand **access** to

and improve **usability** of

ESS data

Findable Accessible Interoperable Reusable

data.ess-dive.lbl.gov

900+ public datasets

Use of ESS-DIVE is <u>mandatory</u> for ESS project data management plans

ESS-DIVE At-a-glance

Authorized users	Public datasets	Private datasets	Total file size	Number of Files	Contributing projects
383 (498%)	909 (144%)	266 (1673%)	2.82TB	16,762	90 (233%)
Public dataset updates	**Temporal coverage**	**Public Portals**	**File downloads**	**Data views**	**Support Requests**
102 (1600%)	1841 to 2024	15 (650%)	429.1K+ (504%)	753.8K+ (510%)	1333 (256%)

Growth statistics measured over the period from May 2020 - March 2024

ESS-DIVE Focus FY20-24

Community Adoption

Goal: Expand the range of ESS projects engaged and archiving with ESS-DIVE

Data/Metadata Standards

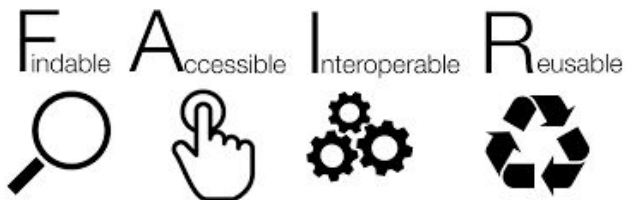**Goal:** Make data widely findable and usable

Scaling

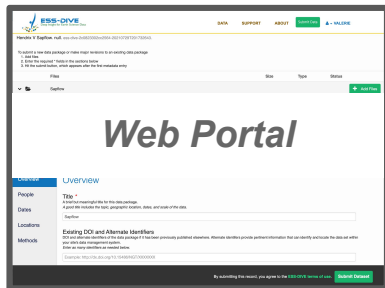**Goal:** Grow the repository capacity and resilience as demand grows

Automation

Goal: Apply automation techniques to ESS-DIVE processes to minimize human effort (human-in-the-loop).

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

ESS-DIVE Data Lifecycle

Upload

Web Portal

Programmatic

{ REST API }

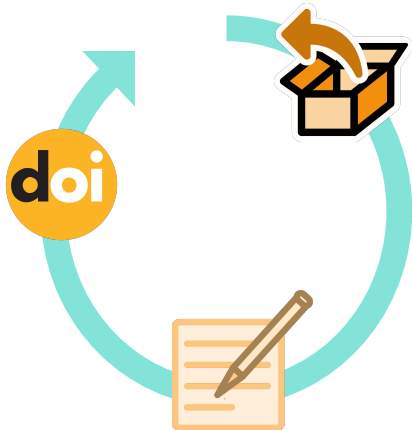## Publish

ESS-DIVE's *publication life cycle* allows users to iteratively curate, package, and update their data and metadata.
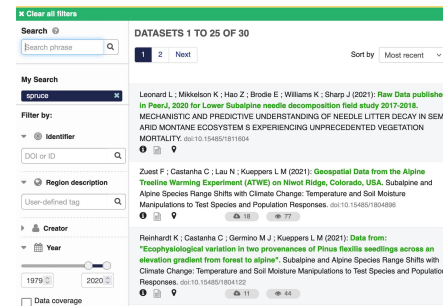
Preserve

Data Preservation Principles

- DOIs
- Quality metadata
- Change management rigor
- Redundancy
- Data auditing and reporting

Discover

DATASETS 1 TO 25 OF 30

OSTI.GOV

DataCite
FIND, ACCESS, AND REUSE DATA

DataONE

Google

Dataset Search

Dataset Submission and Automated Review

Web UI

NEW: Updated checks & reviews for reporting formats

```
provider = {
    "identifier": {
        "@type": "PropertyValue",
        "propertyID": "ess-dive",
        "value":
        "1e6d50d3-9532-43fb-a63f-bdcb4350bf0c"
        }
    }
```

{ REST API}
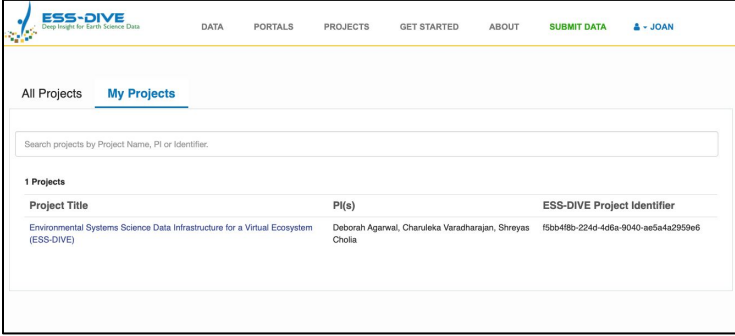
Project Data Management Tools

- **Project teams** for internal sharing, collaboration and curation

- **Bulk data uploads, updates** using API

- Project-centric data **search** and **portals**

- **COMING SOON!!** **Dataset tracking and reserving DOIs**

- **NEED INPUT!!** Project data **citations and metrics**

PIs will walk away with a To-Do List

Lunchtime Breakout: ESS-DIVE Tutorial for PIs and Data Managers
Wed., April 17th; 12:30-1:30pm EST

Project Teams

Teams are **groups of ESS-DIVE users** that have the same permission to access datasets

Teams provide an easier way to share with your project or group

8 projects have used this feature since 2023 launch

1. Share datasets with a team

2. Add people to team

New team members get immediate access to all project datasets!

Share Datasets

Sharing enables teams of scientists to build, curate, and publish datasets together

Data Contributor

Collaborator

PI / Project Manager

Click Share to give permissions to your data
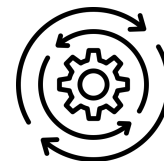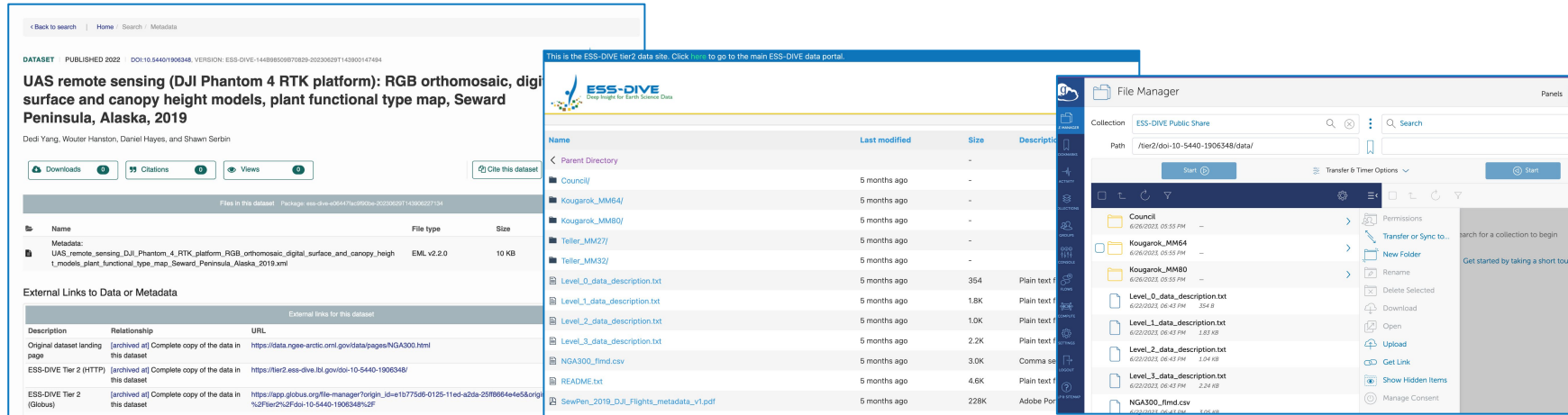
Large Data on Tier 2

Secondary, direct access storage with metadata on ESS-DIVE

- Large Data: 500GB - multiple TB range
- Browsable Folder Hierarchy
- Globus and HTTP Access
- Email ess-dive-support@lbl.gov if you need to store large data

7.3TB data published on Tier 2

10 Tier 2 datasets in past year

ESS-DIVE Formats and Guidelines for Submitting Data Developed in Partnership with the Community

A) Use high-level reporting formats that can apply across data types

- Dataset metadata
- File level metadata
- CSV files
- Model data
- Location metadata
- Samples - identifiers and metadata

B) Reporting formats for specific data types

- 16s amplicon sequencing
- Leaf-level gas exchange
- Continuous soil respiration
- Water & sediment chemistry
- Hydrologic monitoring
- UAS/Drone

C) Submit data to ESS-DIVE
Using the ESS-DIVE website or API

ESS-DIVE Community Space on GitHub

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

Hands-on Session: ESS-DIVE Data Reporting Formats
Mon., April 15th; 1:45-3:15pm EST

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

scientific **data**

Data Standards Enabling Deep Search & Integration

NEW: Fusion DB and Deep Dive API

- **Deep Dive API:** https://fusion.ess-dive.lbl.gov/
- FusionDB looks inside data files: Indexes & validates 100+ datasets with ESS-DIVE file-level metadata
- 37% of newly published datasets (since July 23) are **searchable via Deep Dive API** with 11,563 unique data fields.
- **Automated Parsing** and **Validation** checking for FLMD reporting format
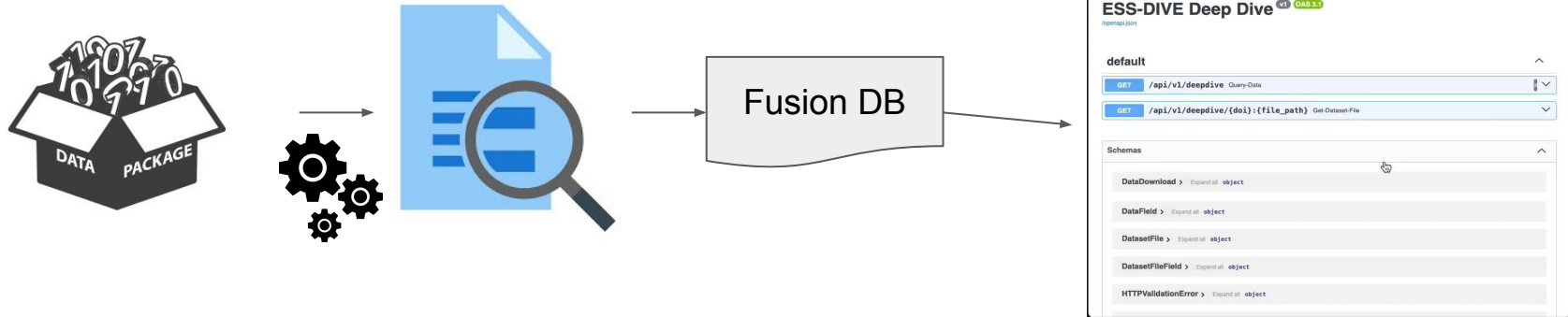
BASIN-3D: A Framework for Powering On-Demand Data Integration

v1 available: https://pypi.org/project/basin3d/ and https://github.com/BASIN-3D

BASIN-3D provides a common query mechanism to access multiple disparate data sources. Results are transformed into a common data model.

Single query to multiple data sources

BASIN-3D synthesized results

BASIN-3D Synthesizer

[Data Source 1]

[Data Source 2]

[Data Source N]

User-Defined Data Sources

<u>v1 Capabilities</u>

python package or django web framework

Synthesizes time series data from:
- **USGS Daily / Instantaneous Values**
- **EPA Water Quality eXchange (WQX)**
- ***ESS-DIVE Hydrologic Monitoring RF***

Output formats: pandas dataframe, hdf5

Harmonizes variable terms into common format

Documentation with examples:
   https://basin3d.readthedocs.io

Varadharajan et al. (2022). BASIN-3D: A brokering framework to integrate diverse environmental data, Computers & Geosciences, Volume 159, 105024, https://doi.org/10.1016/j.cageo.2021.105024. Source: Hubbard et al., 2017, Vadose Zone Journal & https://basin3d.readthedocs.io

Bringing Compute and Data Together

- Dedicated **storage and compute platform** for synthesis and analysis. (Ceph FS Storage 0.5 PB, Compute Nodes on LBL IT Lawrencium)

- *NEW!! Early Users*: Community **Jupyterhub** platform for users to directly operate on ESS-DIVE data.

Bringing Compute and Data Together

- Dedicated **storage and compute platform** for synthesis and analysis. (Ceph FS Storage 0.5 PB, Compute Nodes on LBL IT Lawrencium)

- *NEW!!* *Early Users*: Community **Jupyterhub** platform for users to directly operate on ESS-DIVE data.

Linking Related Sample Data Using Identifiers

[ESS-DIVE Sample Data Portal](#) with IGSNs: *26 datasets*
- Project Use Cases Linking BER Data: **7 datasets**

ESIP Physical Samples Curation Cluster
- Scientific Author Guide and Paper: publishing sample-based research

Exploring Data Harmonizer tool (used by NMDC) for validating ESS-DIVE sample metadata

4 STEPS TO PUBLISH OPEN EARTH SCIENCE SAMPLES

1. Describe samples with rich metadata, ideally using a standardized community template.
2. Assign or use identifiers (such as IGSNs) for samples
3. Publish and cite datasets with sample identifiers
4. Reference samples in your papers using consistent formatting

Flyer: https://doi.org/10.6084/m9.figshare.24291148.v2
Author Guide: https://doi.org/10.6084/m9.figshare.24669057.v1

	Sample Name	Collector/Chief Scientist	Collection date	Collection time	Collection method description	Sample processing	Field program/cruise	Latitude	Longitude
1	EC1_K001_WATER_40ML_FILT	Donnie Day	12/14/2021					41.6228	-83.2362
2	EC1_K001_WATER_15ML_FILT	Donnie Day	12/14/2021					41.6228	-83.2362
3	EC1_K001_WATER_1L_UNFILT	Donnie Day	12/14/2021					41.6228	-83.2362
4	EC1_K001_WATER_125ML_UNFILT	Donnie Day	12/14/2021					41.6228	-83.2362
5	EC1_K001_WATER_FILTER	Donnie Day	12/14/2021					41.6228	-83.2362
6	EC1_K001_SEDIMENT_JAR	Donnie Day	12/14/2021					41.6228	-83.2362
7	EC1_K001_SEDIMENT_BAG	Donnie Day	12/14/2021					41.6228	-83.2362
8	EC1_K001_WETLAND_JAR	Donnie Day	12/14/2021					41.62182	-83.23883
9	EC1_K001_WETLAND_BAG	Donnie Day	12/14/2021					41.62182	-83.23883
10	EC1_K001_WETLAND_RING	Donnie Day	12/14/2021					41.62182	-83.23883
11	EC1_K001_UPLAND_JAR	Donnie Day	12/14/2021					41.61511	-83.22979

2023 Annual Community Data Workshop

Open Data Workshop Agenda
November 15-16, 2023

Wed, Nov 15, 2023
All times listed in Pacific Time zone

Time	Session
09:00AM - 10:00AM	**Opening Session** • Welcome and intros • BER remarks on open data • ESS-DIVE features update • ESS-DIVE roadmap and vision • Q&A
10:00AM - 11:00AM	**Project Data Management and Open Science**
11:00AM - 11:30AM	**Break**
11:30AM - 12:30PM	**Community Led Discussion:** Project Data Publishing Decisions to Support Open Environmental System Science
12:30PM - 02:00PM	**Connecting Related Data for your Project**

Thurs, Nov 16, 2023
Attendees to choose their own breakout sessions in zoom

Time	Session
09:00AM - 11:00AM	**ESS-DIVE How-to Tutorials*** • Publishing data on ESS-DIVE • Managing Project Data on ESS-DIVE • Finding & Accessing ESS-DIVE Data
11:00AM - 11:30AM	**Break**
11:30PM - 1:30PM	**Advanced Search and Data Integration Enabled by ESS-DIVE Reporting Formats** • Enabling Data Validation and Advanced Search • BASIN-3D Data Synthesis Demonstration • Curating High-Quality Datasets for your Project
1:30PM - 02:00PM	**Workshop Wrap-Up and Future Directions**

Community Presenters
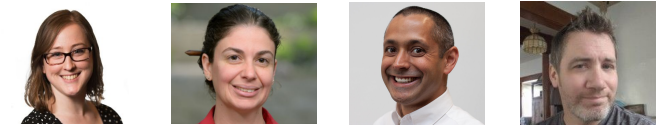
ESS-DIVE Community Workshop

Virtual, Fall 2024 – Dates TBD

18

What's Next?

Tracking Dataset Status

All Available Statuses

Streamlining Publication Process

We will host a webinar with more information when these features are released

Improving DOI Management Transfer

Make it easier for ESS-DIVE to manage your OSTI DOIs

- Mirror your existing datasets on ESS-DIVE

- Work with ESS-DIVE team to transfer ownership of OSTI records

Current work: Collaborating with OSTI to streamline ownership transfers & automate record updates for transferred DOIs

Existing DOI records will be **automatically updated** with latest changes to associated ESS-DIVE datasets

Ask us about updating your DOI landing pages!

Project Archive
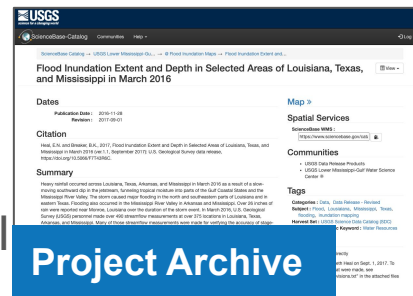
*doi:10.**21952**/1545781*

Mirrored on ESS-DIVE
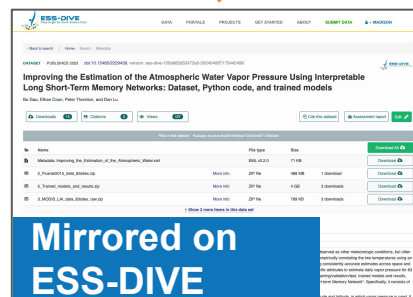
Update OSTI record on ESS-DIVE

Project Data Management Improvements

PIs and data managers default access for their project data

PI

ESS-wide project data reporting metrics
Need your input!

Hierarchical Data Views

Feature to be released late 2024

Downloads	0	Citations	0	Views	0			Cite this dataset	Assessment report

Files in this dataset   Package: resource_map_urn:uuid:edb8385f-54df-49f9-982d-b23e7df27e74

Files and Folders	Size	Type	Metrics	Actions
▾ test 7				⬇
⊞ level1.png	1.20 MiB	image/png		⬇
⊞ Metadata: test.xml	2 KiB	EML v2.2.0		⬇
▾ 📁 photos				
⊞ level2_1.png	1.66 MiB	image/png		⬇
⊞ level2_2.png	1.93 MiB	image/png		⬇

All designs subject to change

BER Data Systems and Project Use Cases

Collaborative paper

1. Scientific use cases to link and exchange (meta)data
2. Identify community practices and infrastructure needs

Approach:

- 8 BER Data Systems, 10+ project use cases
- Connect related field to lab to model (meta)data
- Samples, 'omics, sensors, remote sensing, model outputs, processed data, papers

Breakout: Strategic Data Management to Advance Environmental Systems Science
Tues., April 16th; 11:15-12:45pm EST
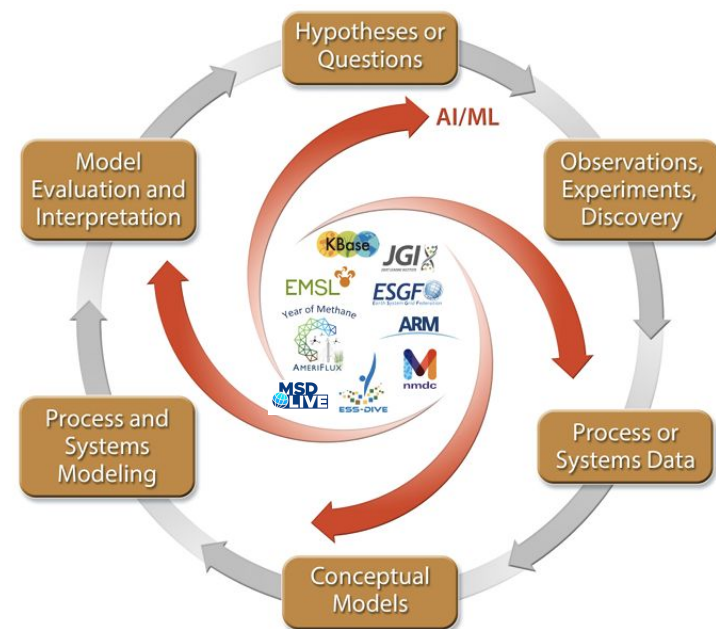
Connecting BER Data for ModEx
Figure from AI4CH4 workshop report (US DOE 2023)
https://ess.science.energy.gov/ai4ch4/

Expanding Cross-Repository Metadata Integration

External Repositories

Project Data Archive

University Repository

Exploring:
- Cross-linked data
- Enhancing search across repositories
- Indexing metadata and connecting to other repos

External Links to Data or Metadata

External links for this dataset		
Description	Relationship	URL
Soil thickness estimation v1.0 (archived at Zenodo)	[archived at] Complete copy of the data in this dataset	http://doi.org/10.5281/zenodo.4445383

BASIN-3D Enhancements

AmeriFlux **FLUXNET data product** with gap-filled, QA/QC'd data for 196 sites across the Americas (subset of figure)
- C, H_2O, energy fluxes
- Meteorological data
- NPP, RECO

ARM Atmospheric & meteorological measurements co-located with ESS sites/data (e.g. SAIL at East River Watershed, Colorado, Ameriflux/ARM sites)

Data availability
- ⊙ data:20+ year (24)
- ⊙ data:15–19 year (30)
- ⊙ data:10–14 year (55)
- ⊙ data:5–9 year (177)
- ⊙ data:1–4 year (168)
- ✳ registered (165)

Sites by IGBP
- Evergreen broadleaf forest (12)
- Decidous broadleaf forest (59)
- Mixed forest (16)
- Evergreen needleleaf forest (100)
- Decidous needleleaf forest (1)
- Woody savanna (9)
- Savanna (8)
- Closed shrubland (13)
- Open shrubland (39)
- Grassland (85)
- Cropland (127)
- Cropland–vegetation mosaics (10)
- Wetland (107)
- Water (10)
- Snow/ice (1)
- Barren land (7)
- Urban (15)

Version 2 of BASIN-3D will:

- Synthesize time series data from: Ameriflux FLUXNET, ARM, USGS Daily & Instantaneous Values, EPA Water Quality eXchange (WQX), ESS-DIVE Hydrologic Monitoring RF
- Authenticate users for data source
- Centrally handle data provided in files
- Customize missing value support
- Have citations added to metadata

Chu et al 2023 SciData

Photo: US-ARM site, Sebastien Biraud, https://ameriflux.lbl.gov

Exploring Use of Large Language Models

Advanced search:
- Dataset recommendations based on plain language queries
- Summarization, extraction of key themes from published datasets
- Question/answering via chat

Data Integration:
- Variable harmonization, semantic matching
- Automated review of reporting format terms to augment and clean metadata

Courtesy Chris Mungall (LBNL)

Community Funds Discussion

Community Funds

- **$1M funds** available in FY24
- Submit **white paper** proposals by May 15 (max 2 pages)
  - Email with details and template will be sent later this week
- Preference given to **Priority Topics**
  - **New reporting formats**
  - **New versions of existing reporting formats:** Improve machine readability and compatibility with FusionDB, BASIN-3D etc., model data archiving for large outputs/ML
  - **Community data curators**: Provide guidance to ESS community on best practices for submitting data to ESS-DIVE & help with adoption of reporting formats
  - **Data integration**: Tools to integrate ESS-DIVE and other BER data
  - **Community data products:** Products using ESS-DIVE & BER data for broad scientific use
- Can propose other ideas that have clear value for the ESS community

Adoption of ESS-DIVE Reporting Formats

A) Use high-level reporting formats that can apply across data types

- Dataset metadata
- File level metadata
- CSV files
- Model data
- Location metadata
- Samples - identifiers and metadata

B) Reporting formats for specific data types

- 16s amplicon sequencing
- Leaf-level gas exchange
- Continuous soil respiration
- Water & sediment chemistry
- Hydrologic monitoring
- UAS/Drone

C) Submit data to ESS-DIVE
Using the ESS-DIVE website or API

Published Datasets Using ESS-DIVE Reporting Format

Reporting Format	Count
CSV File Formatting Guidelines	80
File Level Metadata	78
Sample ID and Metadata	26
Water-Soil-Sediment Chemistry	20
Hydrologic Monitoring	18
Location Metadata	12
Model Data Archiving Guidelines	6
Leaf-Level Gas Exchange	2
Amplicon Sequencing	1
Soil Respiration	1
Unoccupied Aerial Systems (UAS)	1

96 published datasets use one or more reporting formats

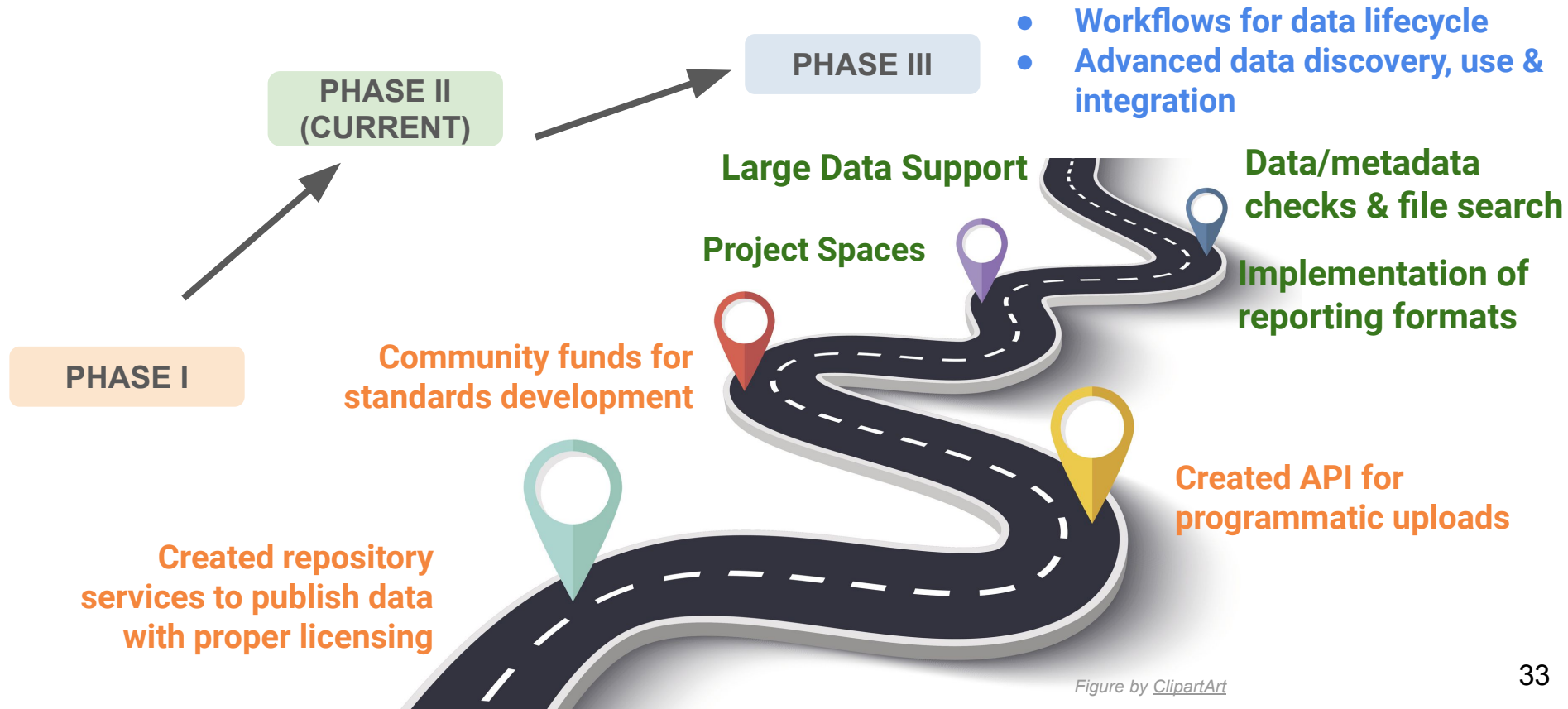*Figure source: Crystal-Ornelas, R. et al. 2022. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. Nature Scientific Data. https://doi.org/10.1038/s41597-022-01606-w*

scientific **data**

31

Discussion on ESS-DIVE's Future Plans

Community Engagement Drives ESS-DIVE Services and Feature Development

PHASE III

- Workflows for data lifecycle
- Advanced data discovery, use & integration

PHASE II (CURRENT)

PHASE I

Large Data Support

Project Spaces

Data/metadata checks & file search

Implementation of reporting formats

Community funds for standards development

Created API for programmatic uploads

Created repository services to publish data with proper licensing

Figure by ClipartArt

33

Vision: Easy Data Submission & Broad Data Availability

EESA20-027
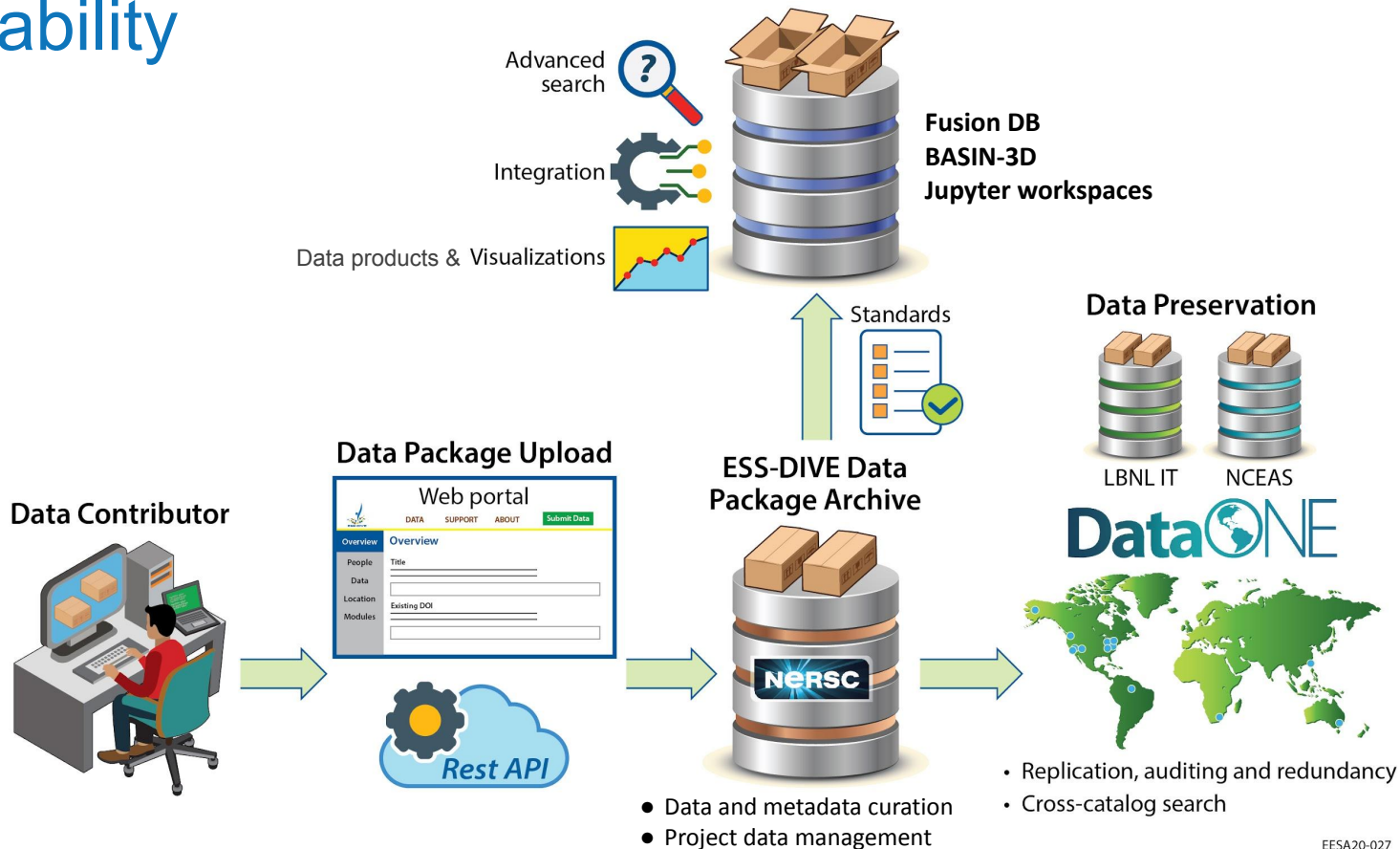
We need your input!

As a community resource how can ESS-DIVE serve your project data management needs?

- Enhancements to current features or new features (including data reporting formats)
- Pain points with current data management workflows or capabilities
- New areas: data integration, data products, Jupyter workspaces
- What kind of support & engagement would you need from ESS-DIVE staff?
- Has our current community engagement worked well and should we make changes?
- Anything else you would like to tell us about?

⭐ **Informal Office Hours / Help** – Mon., April 15th; 5:30-6:00pm EST

Email us at ess-dive-support@lbl.gov to submit your feedback!

ESS-DIVE at the CI/PI meetings

ESS-DIVE Updates / Planning
Discussion
Mon., April 15th; 9:30-10:30am EST

Hands-on Session: ESS-DIVE Data
Reporting Formats
Mon., April 15th; 1:45-3:15pm EST

Hands-on Session: Accessing HPC
resources with Jupyter notebooks
Mon., April 15th; 1:45-3:15pm EST

Round Robin: Model data archiving and
code management
Mon., April 15th; 3:30-5:00pm EST

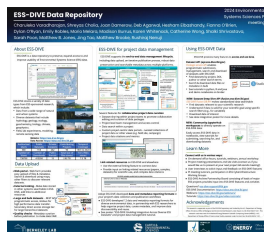⭐ **Informal Office Hours / Help** – Tue., April 16th; 6:30pm EST

Strategic Data Management to Advance
Environmental Systems Science
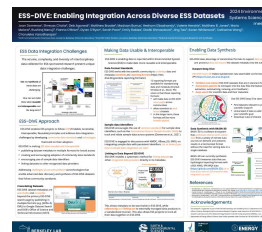Tues., April 16th; 11:15-12:45pm EST

Inclusive and Equitable Research: Best
Practices and Lessons Learned
Tues., April 16th; 5:30-6:30pm EST

Lunchtime Breakout: ESS-DIVE Tutorial
for PIs and Data Managers
Wed., April 17th; 12:30-1:30pm EST

ESS-DIVE Data Repository Overview
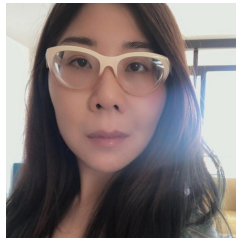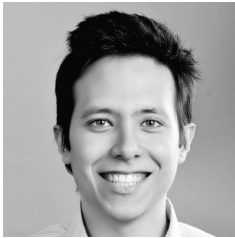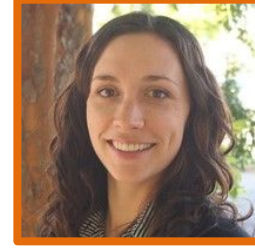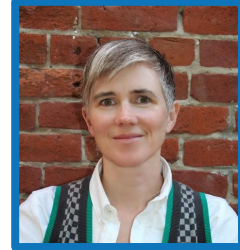Poster Session II: Tues., April 16th, 4:00-5:30pm EST; Poster #12

ESS-DIVE: Enabling Integration Across Diverse ESS Datasets
Poster Session II: Tues., April 16th, 4:00-5:30pm EST; Poster #18

Questions?

Connect with the ESS-DIVE Team!

ESS-DIVE Leadership

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