ESS-DIVE: Enabling Integration Across Diverse ESS Datasets

2024 Environmental Systems Sciences PI meeting

Joan Damerow[1], Shreyas Cholia[2], Deb Agarwal[2], Matthew Brooke[3], Madison Burrus[1], Hesham Elbashandy[2], Valerie Hendrix[2], Matthew B. Jones[3], Mario Melara[2], Rushiraj Nenuji[3], Fianna O'Brien[2], Dylan O'Ryan[1], Sarah Poon[2], Emily Robles[1], Shalki Shrivastava[2], Jing Tao[3], Karen Whitenack[2], Catherine Wong[2], Charuleka Varadharajan[1]

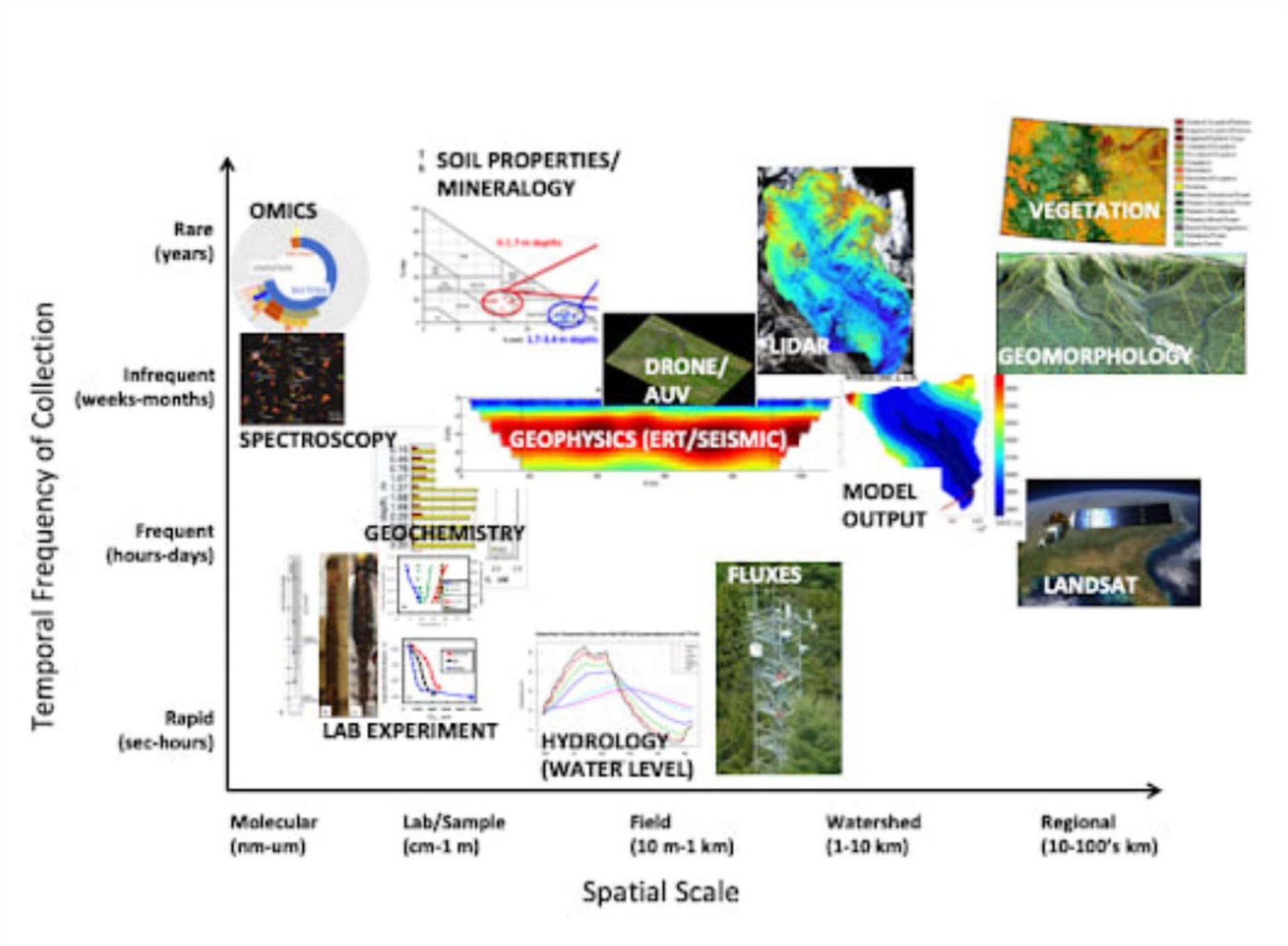
[1]Earth & Environmental Sciences Area, Lawrence Berkeley National Laboratory; [2]Scientific Data Division, Lawrence Berkeley National Laboratory; [3]National Center for Ecological Research and Synthesis

ESS Data Integration Challenges

The volume, complexity, and diversity of interdisciplinary data collected for ESS-sponsored research present unique data integration challenges.

***Use** and **synthesis** of these data are challenging.*

*How can we make these data **reusable** and **interoperable** over the long-term?*



Complexity of Data in ESS
Varadharajan et al., 2022, Computers & Geosciences

ESS-DIVE Approach

ESS-DIVE enables ESS projects to follow *FAIR* (Findable, Accessible, Interoperable, Reusable) principles and address data integration challenges by developing *community data standards* and *technologies* that build on their adoption.

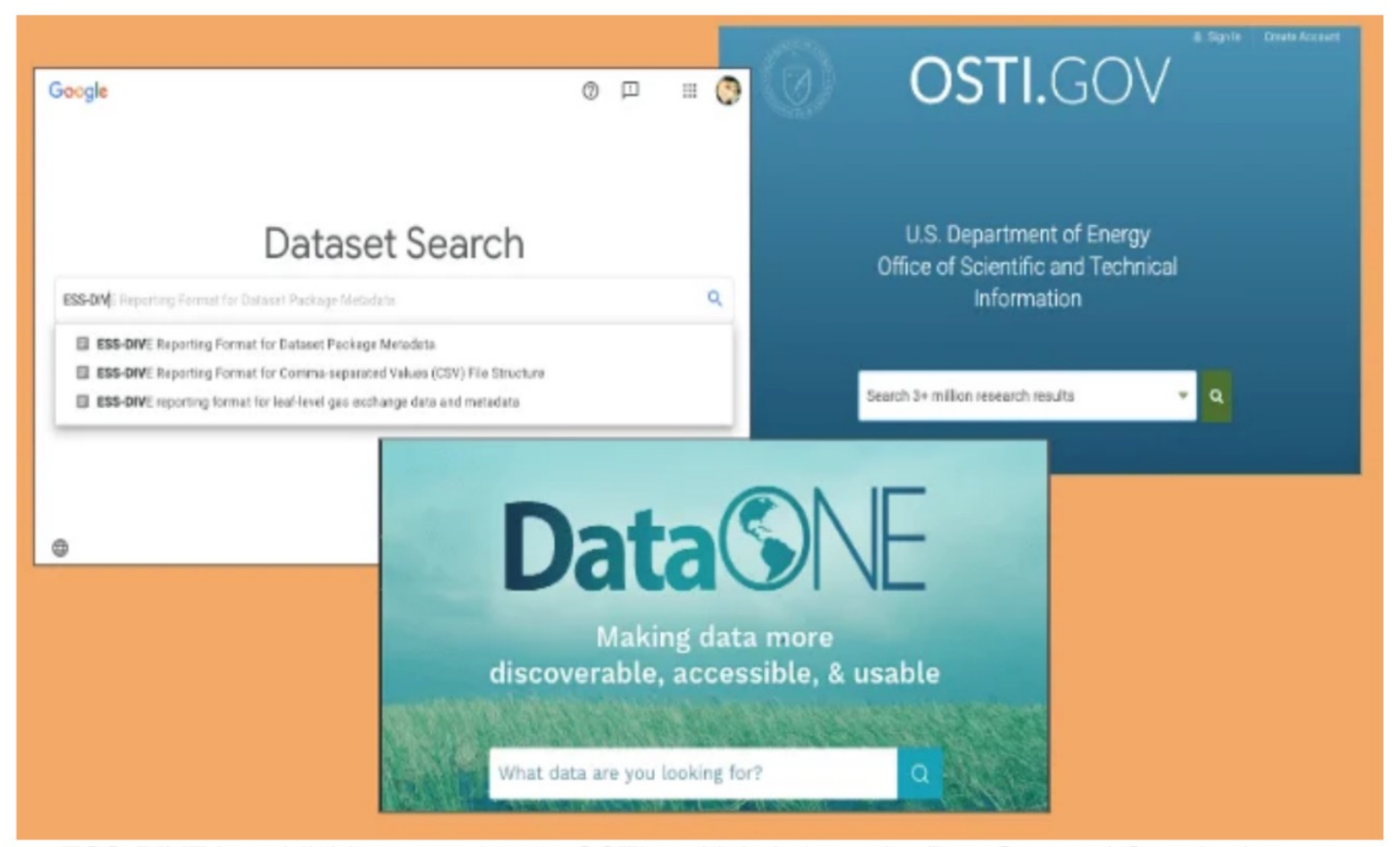
ESS-DIVE is making *ESS data reusable and interoperable*

- publishing dataset metadata in multiple formats for broad access
- creating and encouraging adoption of community data standards
- encouraging use of sample data identifiers
- linking datasets to other recognized data providers

Addressing *challenges of use and synthesis* via technologies that enable advanced data discovery and synthesis of ESS-DIVE datasets that follow community standards.

Cross-listing Datasets

ESS-DIVE dataset metadata are *searchable* and *accessible* beyond the primary ESS-DIVE search page by publishing in multiple formats (e.g. JSON-LD, EML) to Google Dataset Search and DOE's Office of Science and Technical Information (OSTI)


Google
Dataset Search
ESS-DIVE Reporting Format for Dataset Package Metadata
ESS-DIVE Reporting Format for Dataset Package Metadata
ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure
ESS-DIVE reporting format for leaf-level gas exchange data and metadata
OSTI.GOV
U.S. Department of Energy Office of Scientific and Technical Information
Search 3+ million research results
DataONE
Making data more discoverable, accessible, & usable
What data are you looking for?

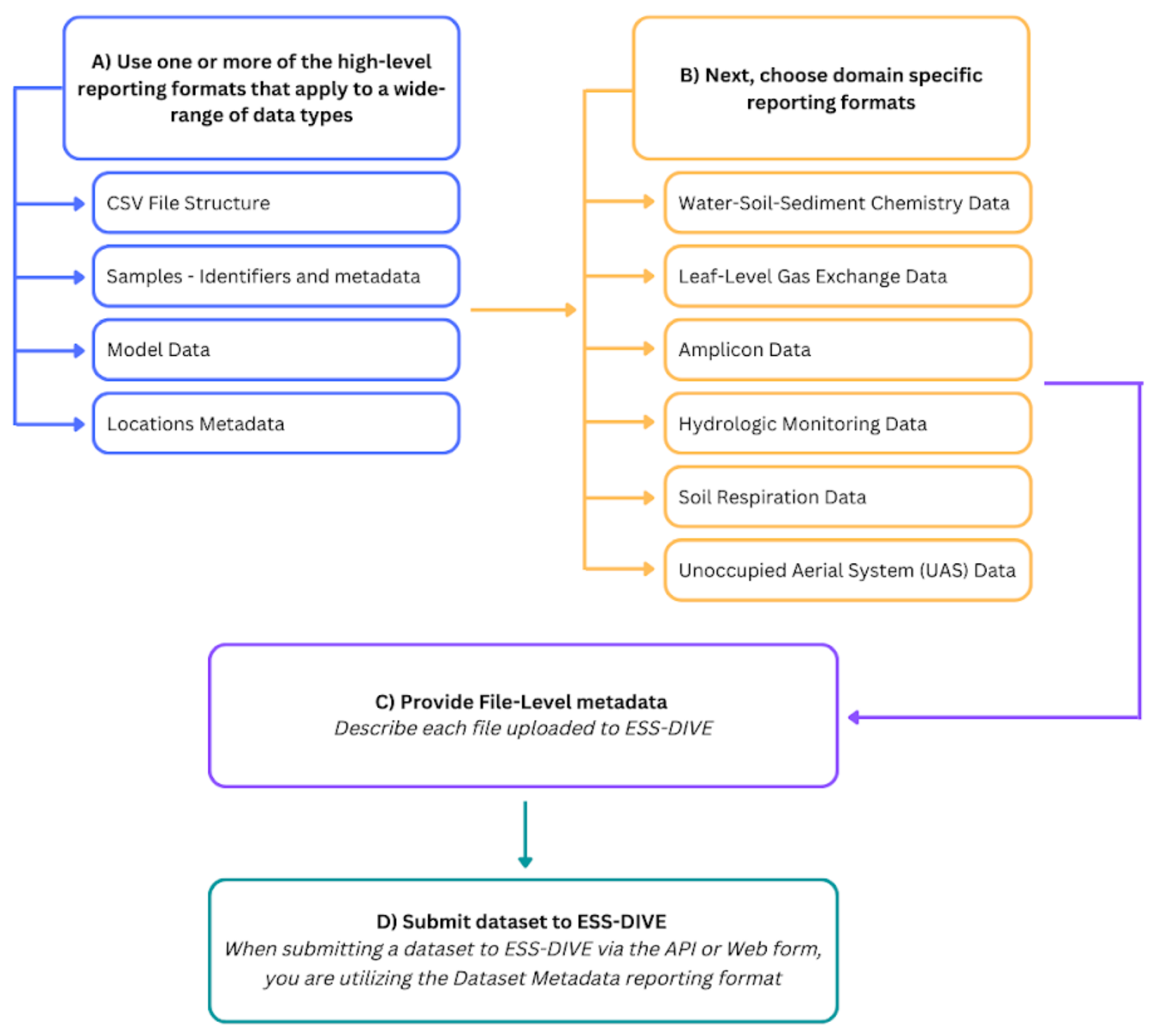
ESS-DIVE is publishing metadata to OSTI and it is harvest by DataOne and Google dataset search, which allows data to gain a much broader footprint.

Making Data Usable & Interoperable

ESS-DIVE is enabling data is reported within Environmental System Science (ESS) to make data more reusable and interoperable.

Data Format Standards

ESS-DIVE works with the scientific community to *co-develop* data and metadata ***standards*** and ***reporting formats*** (https://ess-dive.lbl.gov/data-reporting-formats/)



A) Use one or more of the high-level reporting formats that apply to a wide-range of data types

- CSV File Structure
- Samples - Identifiers and metadata
- Model Data
- Locations Metadata

B) Next, choose domain specific reporting formats

- Water-Soil-Sediment Chemistry Data
- Leaf-Level Gas Exchange Data
- Amplicon Data
- Hydrologic Monitoring Data
- Soil Respiration Data
- Unoccupied Aerial System (UAS) Data

C) Provide File-Level metadata
Describe each file uploaded to ESS-DIVE

D) Submit dataset to ESS-DIVE
When submitting a dataset to ESS-DIVE via the API or Web form, you are utilizing the Dataset Metadata reporting format

12 reporting formats are available for standardizing data and metadata (Crystal-Ornelas et al., 2022). The vision is that these reporting formats:

- will make data in ESS-DIVE *more useful* across communities
- allow scientists to *work across* datasets
- in the longer term, these formats will be more *broadly applied*.

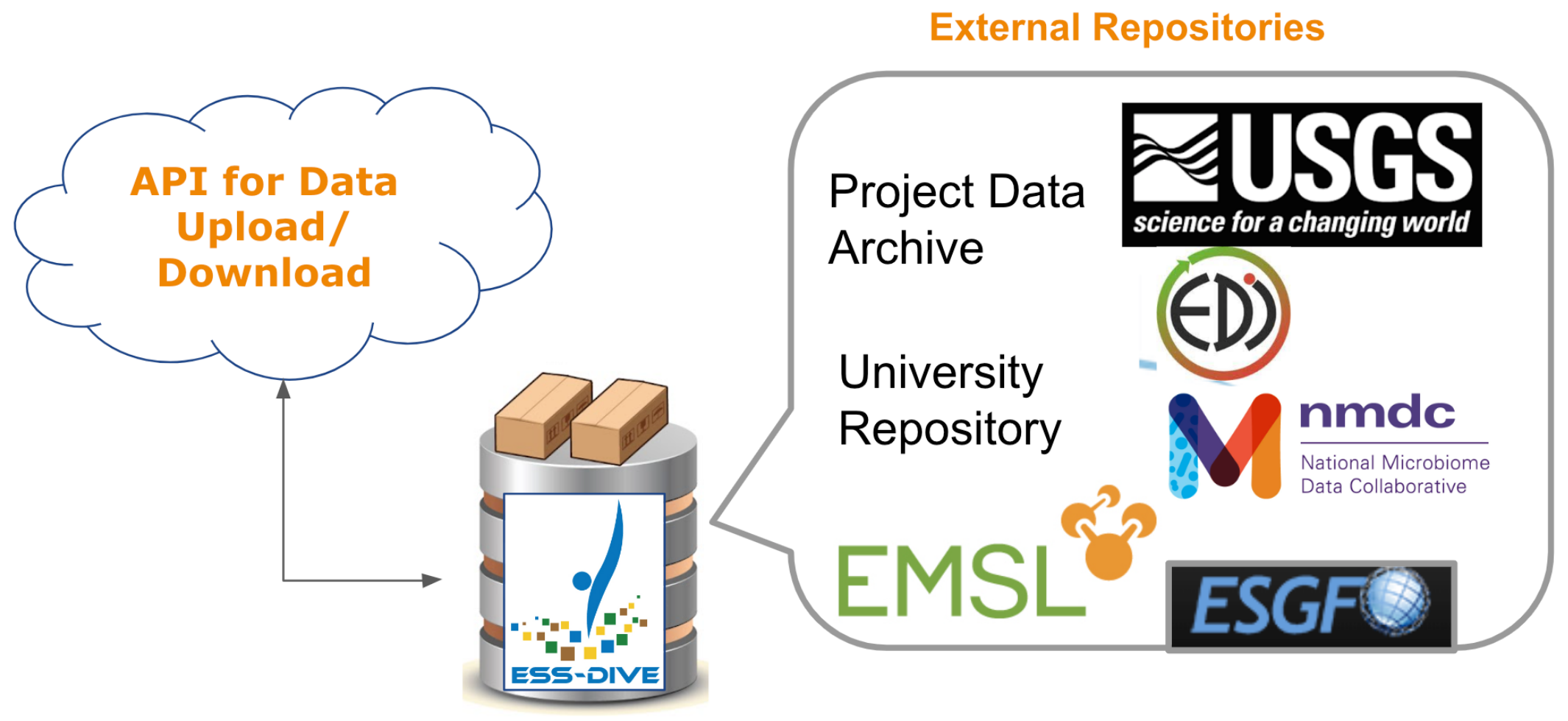
Sample data Identifiers

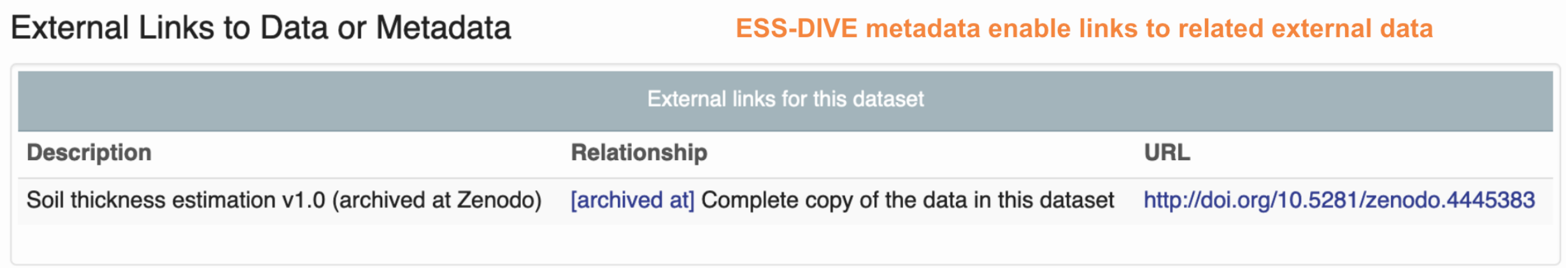
ESS-DIVE encourages the use of *common standards* for sample data identifiers, such as the *International Generic Sample Number (IGSN)* to track and relate sample data across systems (Damerow et al., 2021).

ESS-DIVE is engaged in discussions with NMDC, KBase, JGI, EMSL on integrating sample data with persistent identifiers. *Let us know if you have a sample data integration use case!*

Linking to Data Beyond ESS-DIVE

ESS-DIVE enables a systematic method for *linking datasets* to other *recognized data providers* directly in its metadata.


External Repositories
API for Data Upload/ Download
Project Data Archive
University Repository
USGS science for a changing world
EDI
nmdc National Microbiome Data Collaborative
EMSL
ESGF
ESS-DIVE



External Links to Data or Metadata

ESS-DIVE metadata enable links to related external data

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

This allows metadata to be searchable in ESS-DIVE, while ***referencing** and **linking out*** to externally managed data products in a standardized manner. This also allows ESS projects to track all their data together on ESS-DIVE.

Enabling Data Synthesis

ESS-DIVE takes advantage of standardized formats to support *data synthesis* and provide a *deep dive beyond* the dataset metadata into the data.

Dataset Deep Dive

ESS-DIVE Fusion DB makes standardized data searchable via the Deep Dive API (https://fusion.ess-dive.lbl.gov/).

- ***Validates and indexes*** ESS-DIVE datasets that are in standard formats
- ***Automation pipeline*** to introspect into the data files themselves (e.g. extraction, summarizing, indexing, error feedback)
- ***Deep search*** for scientific data and their metadata.

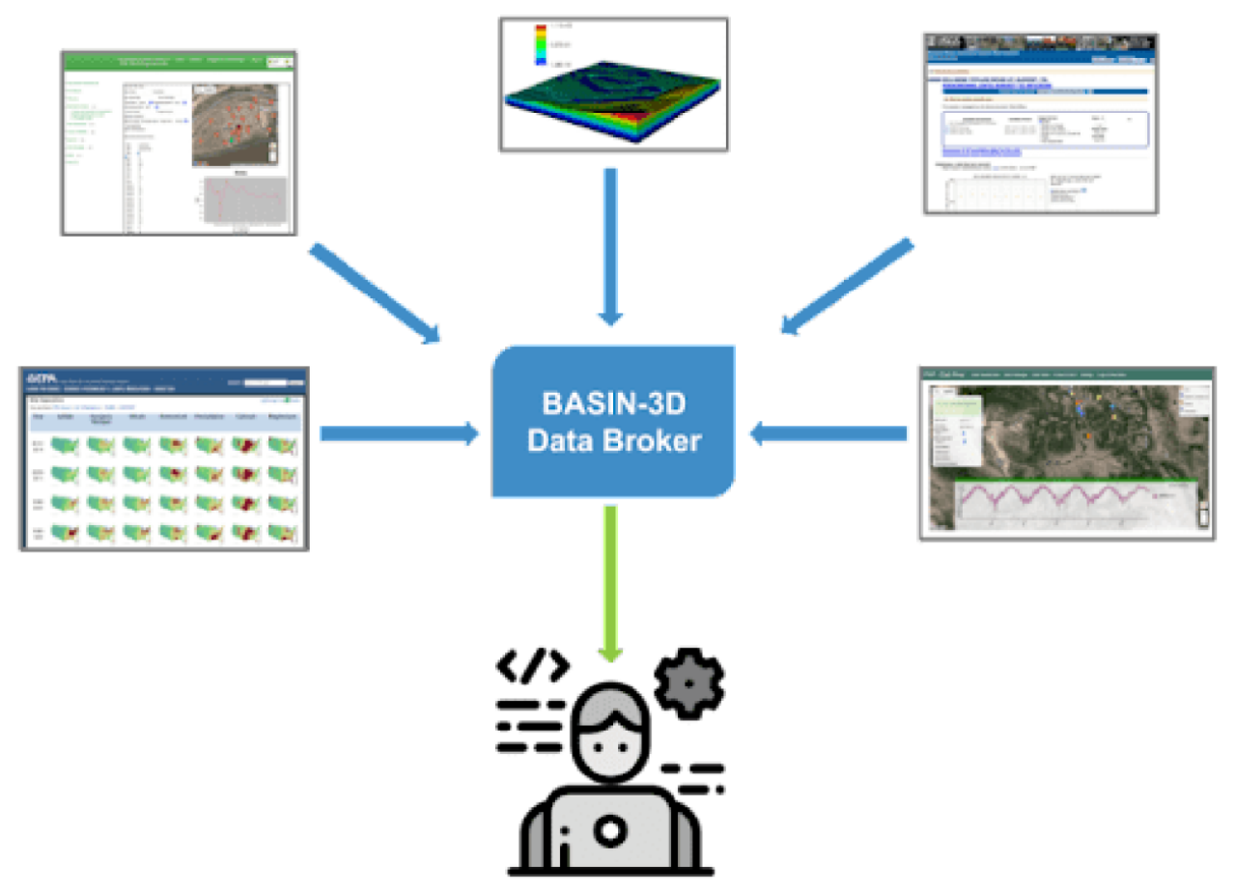
Take a deep dive into ESS-DIVE data at https://fusion.ess-dive.lbl.gov (https://fusion.ess-dive.lbl.gov)

Use ESS DIVE Deep Dive search to

- Find datasets relevant to your scientific research
- Understand if data is valid for your scientific goal
- Download data of interest

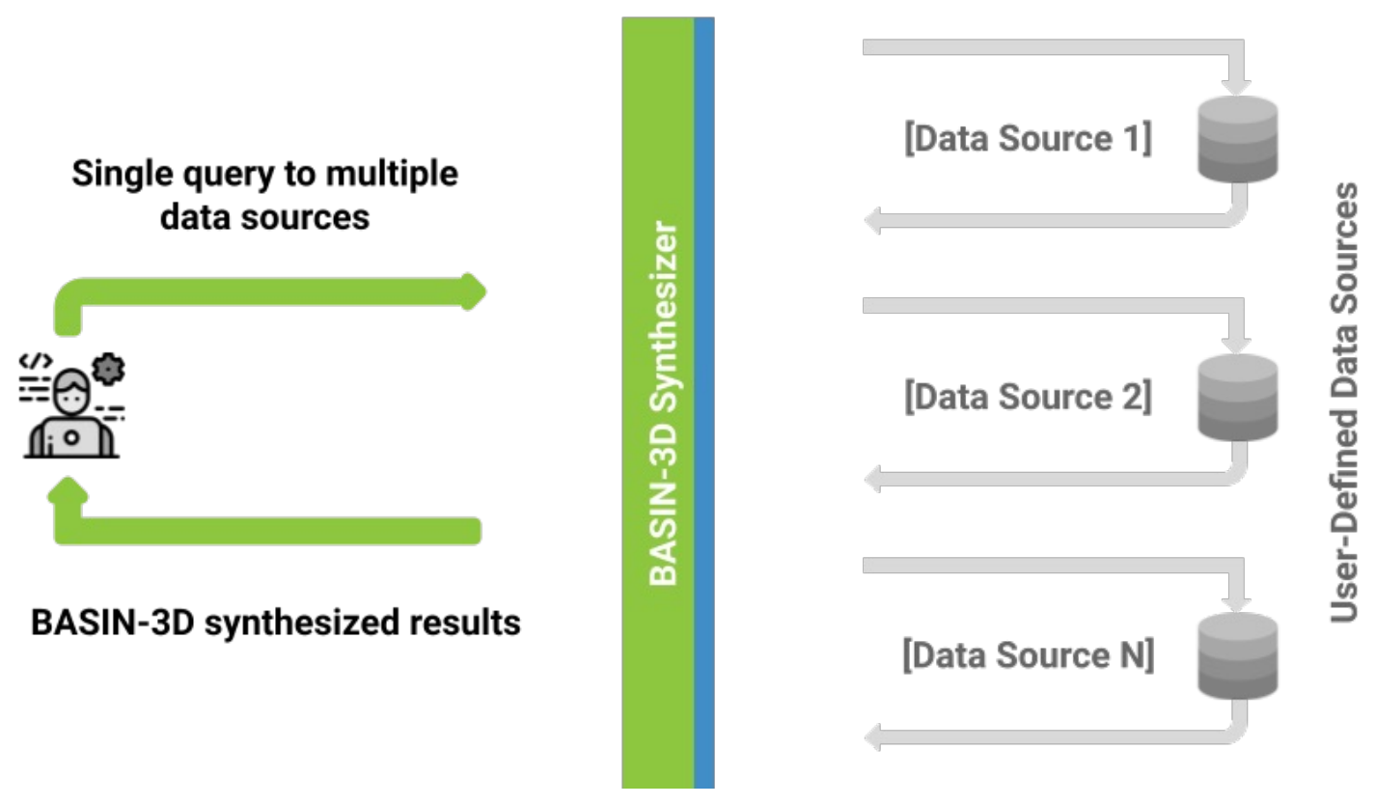
Data Synthesis with BASIN-3D

BASIN-3D is a software ecosystem that *synthesizes diverse earth science data* from a variety of remote data sources on demand and presents results in a harmonized format without the need for storing data in a single database.



BASIN-3D allows scientists to make a single query to multiple data sources and receive data in one format

BASIN-3D can currently synthesize ESS-DIVE timeseries data that use hydrological reporting formats with USGS NWIS, EPA WQX data (https://github.com/BASIN-3D). *Let us know if you have an ESS timeseries data integration use case!*


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

References

- Varadharajan, C., Hendrix V.C., Christianson D.S., 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
- Crystal-Ornelas, R., Varadharajan, C.*, O'Ryan, D. 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
- Damerow, JE, Varadharajan, C, et al. (2021). Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences. Data Science Journal, https://doi. org/10.5334/dsj-2021-011

Acknowledgements

This research is supported as part of the ESS-DIVE repository and funded by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award no. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

BERKELEY LAB

EARTH & ENVIRONMENTAL SCIENCES

ESS-DIVE

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY | Office of Science