



# ESS-DIVE

## *Environmental System Science Data Infrastructure for a Virtual Ecosystem*

Shreyas Cholia  
*Lawrence Berkeley National Laboratory*  
BERAC Meeting, April 2023

PIs: Charuleka Varadharajan (LBNL), Shreyas Cholia (LBNL), and Deb Agarwal (LBNL)

Team: Valerie Hendrix (LBNL), Joan Damerow (LBNL), Fianna O'Brien (LBNL), Hesham Elbashandy (LBNL), Mario Melara (LBNL), Shalki Shrivastava (LBNL), Madison Burrus (LBNL), Lauren Core (LBNL), Emily Robles (LBNL), Sarah Poon (LBNL), Cat Wong (LBNL), Dylan O'Ryan (LBNL), Matt Jones (NCEAS), Lavanya Ramakrishnan (LBNL), and Karen Whitenack (LBNL)



**NERSC**



DataONE



Office of  
Science

# What is ESS-DIVE?



<https://data.ess-dive.lbl.gov>

Environmental Systems Science Data  
Infrastructure for a Virtual Ecosystem:

Data Repository to **preserve**,  
expand **access** to, and improve  
the **usability** of diverse **Earth  
and Environmental Science  
(ESS) datasets**.

The screenshot shows the ESS-DIVE data portal interface. At the top, there are navigation tabs for DATA, PORTALS, PROJECTS, GET STARTED, ABOUT, and SUBMIT DATA. Below the navigation is a search bar and a filter section with categories like Project, Identifier, Region description, Creator, Year, and Access. The main content area displays a list of datasets, with the first few entries visible: Delgado D.; Barnes M.; Boehlke B.T.; Chen X.; Cornwell K.; Forbes B.; Fulton S.G.; Garayburu-Canuso V.A.; Goldman A.E.; Gonzalez B.I.; Grieger S.; Hammond G.E.; Jiang P.; Kaufman M.H.; Laan M.; Li B.; Li Z.; Lin X.; McKeever S.A.; Mudunuru M.K.; Muller K.A.; Myers-Pigg A.; Ottenburg O.; Pelly A.; Peta K.; Regier P.; Renteria L.; Rheeback A.; Schiebe T.D.; Son K.; Torgerson J.M.; Stegen J.C. (2020): Spatially Study 2020: Surface Water Samples, Cotton Strip Degradation, and Hydrologic Sensor Data across the Yakima River Basin, Washington, USA. River Corridor and Watershed Biogeochemistry SFA, ESS-DIVE repository. Dataset. doi:10.15485/1969566. Below the list is a map of the United States with a starburst overlay indicating 751 public datasets.



# ESS-DIVE Focus



## 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



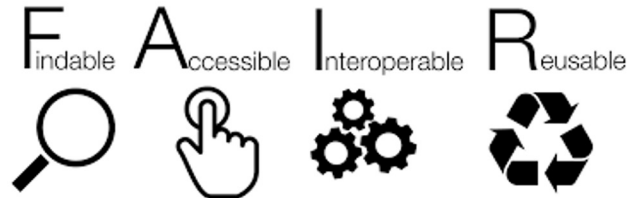
## Scaling

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



## Automation

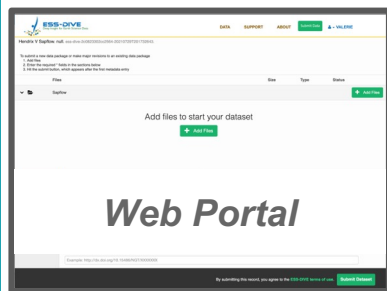
**Goal:** Apply automation techniques to ESS-DIVE processes to minimize human effort (human-in-the-loop).



# ESS-DIVE Data Lifecycle



## Upload



### ESS-DIVE Metadata Quality Report

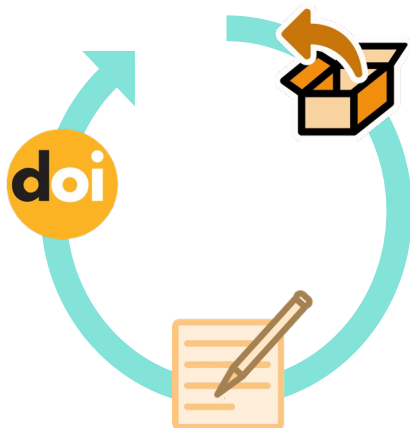
After running your metadata against our standard set of metadata, data, and congruency checks, we've research data by addressing the issues below.

19 checks

- Passed 11 checks out of 12 (informational checks not included).
- Warning for 0 checks.
- Failed 1 check. Please correct these issues.

The abstract is only 75 word(s) long but 100 or more is required.

## Publish



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

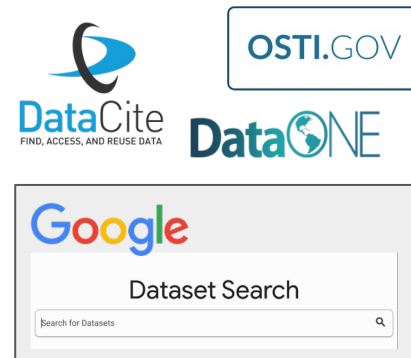
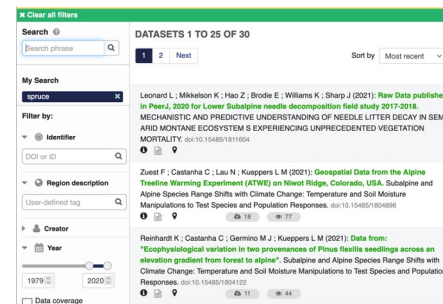
## Preserve

### Data Preservation Principles

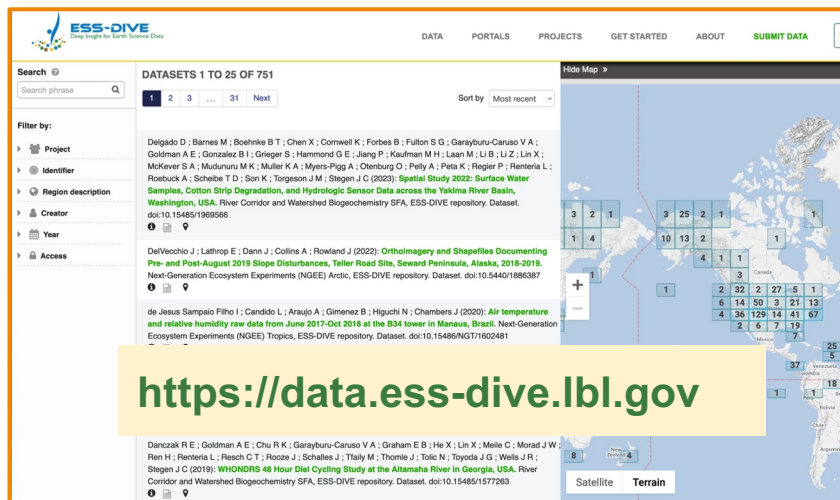
- Durable identifiers
- Quality metadata
- Change management rigor
- Redundancy
- Data auditing and reporting



## Discover

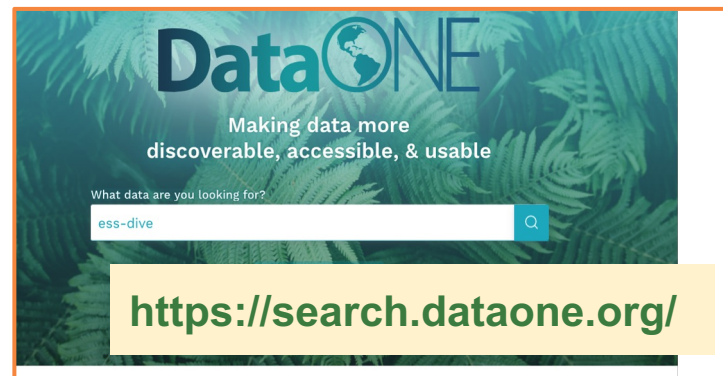


# Search and Access

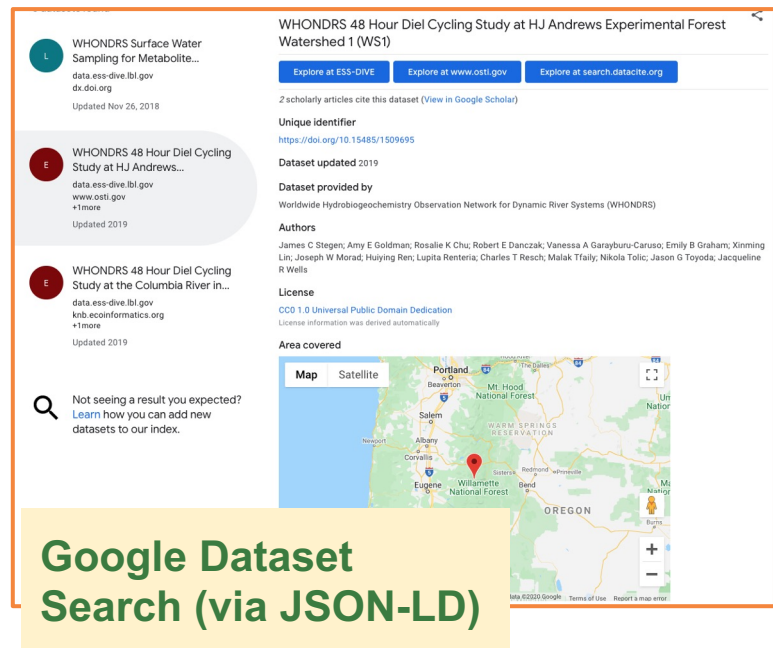


The screenshot shows the ESS-DIVE website interface. On the left, there is a search bar with the text "Search @" and a search phrase input field. Below the search bar, there are filter options for Project, Identifier, Region description, Creator, Year, and Access. The main content area displays "DATASETS 1 TO 25 OF 751" and a list of dataset entries. Each entry includes the creator's name, a brief description, and a DOI link. A map on the right side of the page shows the geographical distribution of datasets, with a grid overlay and a "Hide Map" button. A yellow box is overlaid on the bottom left of the screenshot, containing the URL <https://data.ess-dive.lbl.gov>.

- ESS-DIVE data is automatically **replicated** to DataONE Federation
- Metadata/Data **searchable** via **DataONE** nodes
- We publish **JSON-LD Metadata** which pushes to **Google Dataset Search**



The image shows the DataONE logo, which consists of the word "DataONE" in a blue, sans-serif font, with a globe icon integrated into the letter "O". Below the logo, the text "Making data more discoverable, accessible, & usable" is displayed. A search bar is present with the text "ess-dive" and a search icon. The background features a green, leafy pattern. A yellow box is overlaid on the bottom right of the image, containing the URL <https://search.dataone.org/>.

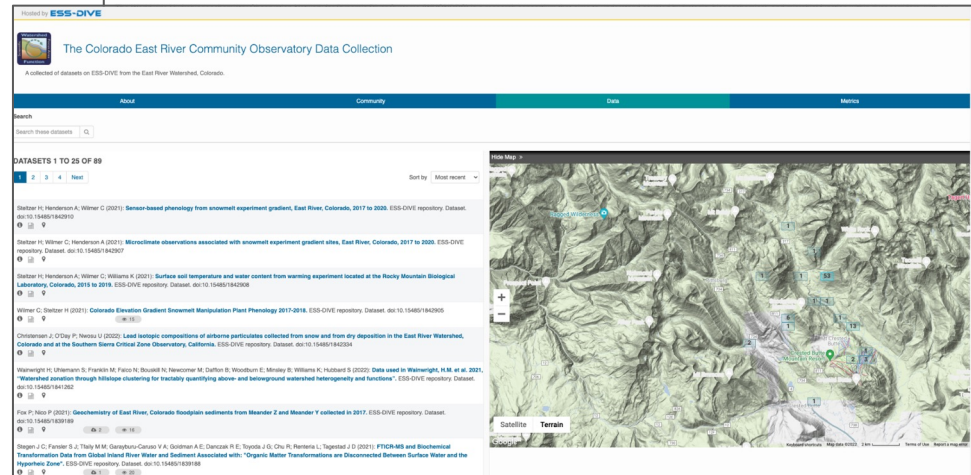
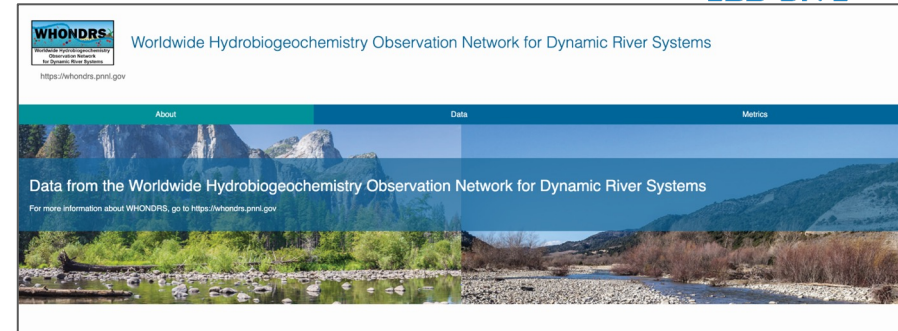


The screenshot shows a Google Dataset Search result for "WHONDRS 48 Hour Diel Cycling Study at HJ Andrews Experimental Forest Watershed 1 (WS1)". The result includes a list of three datasets, each with a unique identifier, a brief description, and a DOI link. The first dataset is "WHONDRS Surface Water Sampling for Metabolite..." with a DOI of 10.15485/1509695. The second dataset is "WHONDRS 48 Hour Diel Cycling Study at HJ Andrews..." with a DOI of 10.15485/1509695. The third dataset is "WHONDRS 48 Hour Diel Cycling Study at the Columbia River in..." with a DOI of 10.15485/1577263. The result also includes a map of the study area in Oregon, showing the location of the HJ Andrews Experimental Forest. A yellow box is overlaid on the bottom right of the image, containing the text "Google Dataset Search (via JSON-LD)".

# Evolving Project Data Management Tools



- Enable project-centric data **portals**
- Project-specific data **search**
- **Internal sharing within project teams** to promote collaboration
- **Administration** of users and **curation** of data packages
- Bulk data upload using API
- Project data **citations and metrics**



# ESS-DIVE formats and guidelines for submitting data developed in partnership with the community



<p>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 Flow sensor data, Jan 2016-Apr 2016, BR1</p> <p>Existing DOI and Alternate Identifiers DOI and alternate identifiers of the data package if it has been previously published elsewhere</p> <p><b>Package Level Metadata/JSON-LD</b></p>	<p>ESS-DIVE Manual Package Metadata Review</p> <p><b>Package Metadata Quality</b></p>	<p><b>SESAO</b> SYSTEM FOR EARTH SAMPLE REGISTRATION</p> <p><b>Sample IDs/Metadata</b></p>	<p><b>Model Data Archiving</b></p>
<p><b>Agarwal, Hendrix (LBNL)</b></p>	<p><b>Damerow (LBNL)</b></p>	<p><b>Damerow (LBNL)</b></p>	<p><b>Simmonds (LBNL)</b></p>

<p><b>File-Level/ csv Metadata</b></p>	<p><b>Soil Respiration</b></p>	<p><b>Leaf Physiology</b></p>
<p><b>Velliquette, Heinz, Devarakonda (ORNL)</b></p>	<p><b>Bond-Lamberty, Pennington (PNNL)</b></p>	<p><b>Rogers, Ely (BNL)</b></p>
<p><b>Hydrologic Monitoring</b></p>	<p><b>Water/Soil Chemistry</b></p>	<p><b>Amplicon Sequencing</b></p>
<p><b>Goldman (PNNL)</b></p>	<p><b>Boye (SLAC)</b></p>	<p><b>Weisenhorn (ANL)</b></p>



**Community Space on GitHub:**  
<https://github.com/ess-dive-community>

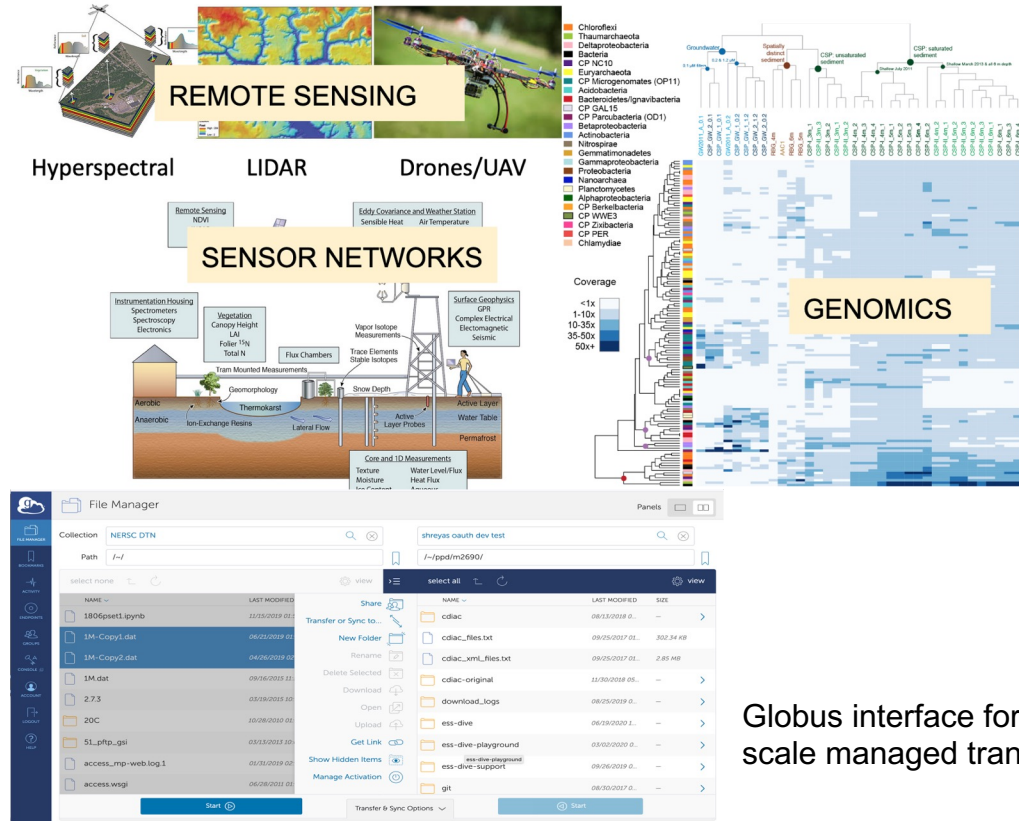


Crystal-Ornelas, R. *et al.* 2022. Enabling FAIR data in Earth and environmental science with community-centric (meta)data reporting formats. *Nature Scientific Data*.  
<https://doi.org/10.1038/s41597-022-01606-w>

# Supporting Growth In Data

Increasing need to support large files and datasets from Models, LIDAR, Sensor Networks etc.

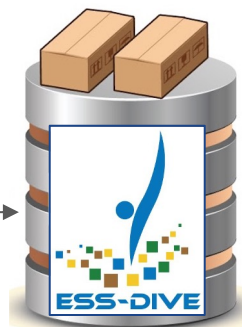
- Globus Data Transfer
- API Access
- Secondary direct access storage with hierarchical large datasets (Metadata in ESS-DIVE)



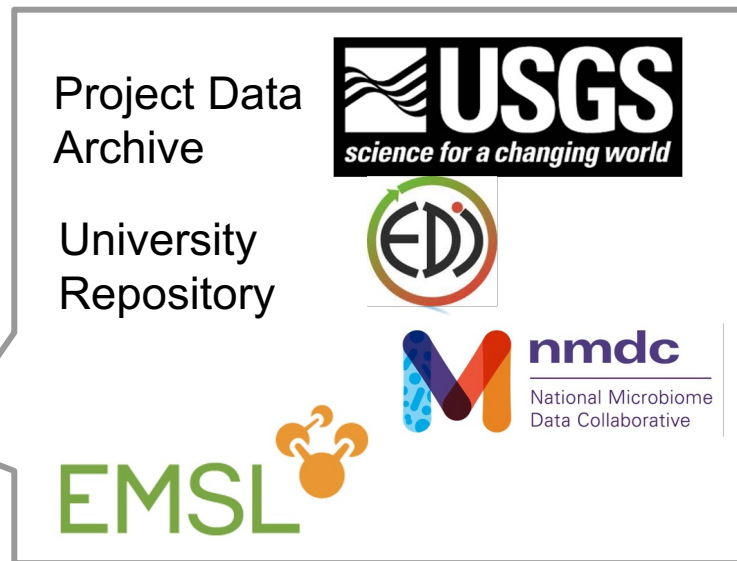
Globus interface for large scale managed transfers



# API and External Integration



## External Repositories



ESS-DIVE metadata enable links to related external data

### 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	<a href="http://doi.org/10.5281/zenodo.4445383">http://doi.org/10.5281/zenodo.4445383</a>

# Interoperability with other BER Systems



- Cross linking data across repositories eg. NMDC links back to ESS-DIVE
- Sample data interoperability working group - develop standards and common identifiers to connect sample data across systems
  - ESS-DIVE, NMDC, KBase, EMSL, JGI
- Collaborating to establish common data submission pipelines
  - Submit data to one system -> automatically extract and push to partner repository

## Study Details



**DOI**  
10.25585/1488224



**ID**  
gold:Gs0135149


**Funding sources**  
This material is based upon work supported as part of the Watershed Function S Department of Energy, Office of Science, Office of Biological and Environmental R AC02-05CH11231. A portion of this research was performed under the Facilities I (FICUS) program and used resources at the DOE Joint Genome Institute and the I (grid.436923.9), which are DOE Office of Science User Facilities. Both facilities ar Environmental Research program and operated under Contract Nos. DE-AC02-05 (EMSL).


**Sample count**  
53

## External Resources

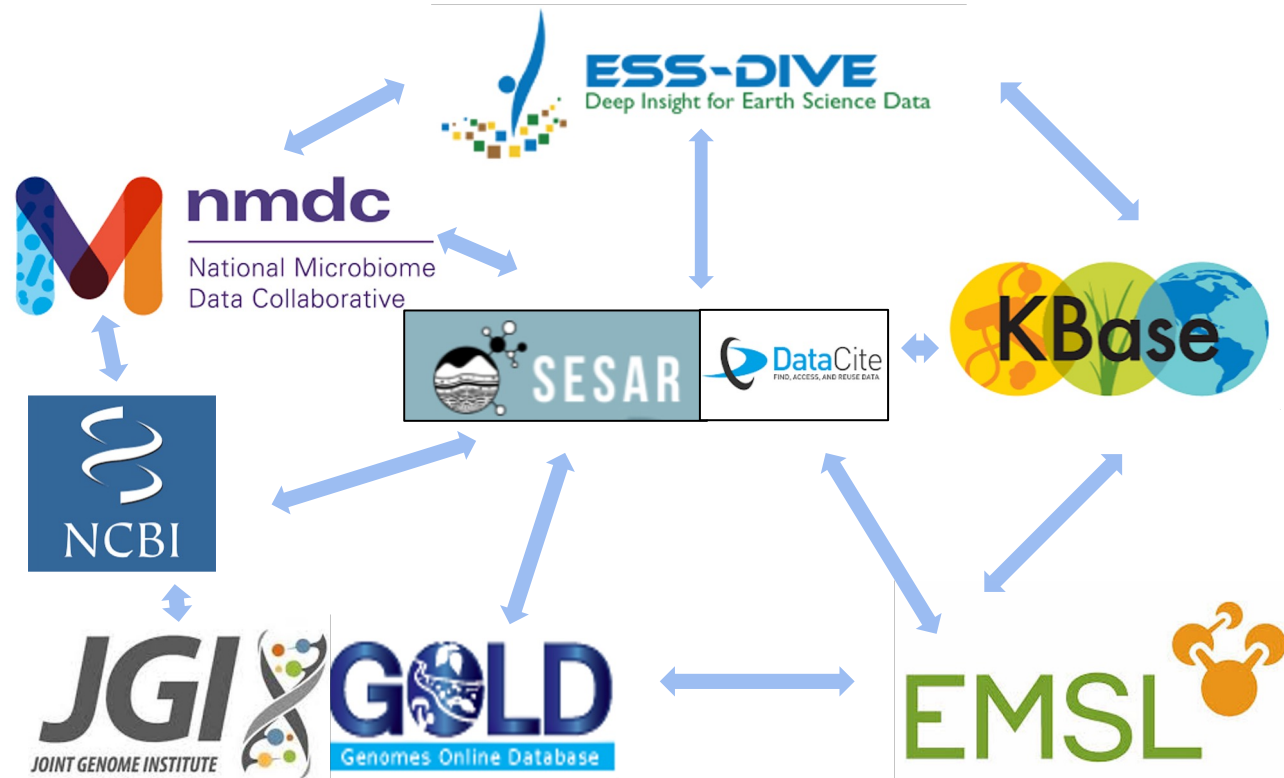
**Additional data**

  Open in GOLD  
<https://gold.jgi.doe.gov/study?id=Gs0135149>

 ESS DIVE Dataset  
<https://identifiers.org/doi:10.21952/WTR/1573029>

 ESS DIVE Dataset  
<https://identifiers.org/doi:10.15485/1577267>

# Persistent Identifiers for Tracking Data



## Persistent Identifiers

1. Link and **expand access** pathways
2. **Exchange relevant metadata** across platforms
3. **Sample tracking, Data Reuse**
4. Goal: enable **integrated search**

# Supporting data integration & access



## MODELS

Improvement of watershed models to include chemical and biological processes

ecosys



PFLOTRAN

## DATA

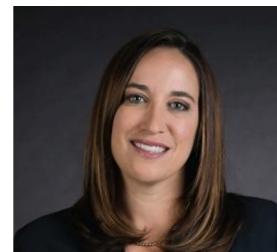


Data assembly, integration, and storage



Reaction-Scale Integration

**GROW**  
Genome Resolved Open Watersheds



Kelly Wrighton



Mikayla Borton



**DISTRIBUTED SCIENCE APPROACH**

Geochemistry, hydrology, metabolites, metagenomes, and metatranscriptomes

U.S. DOE. 2019. Open Watershed Science by Design: Leveraging Distributed Research Networks to Understand Watershed Systems Workshop Report, DOE/SC-0200, U.S. Department of Energy Office of Science.

# GROW-WHONDRS

## Initiative (Project)

<https://www.pnnl.gov/projects/WHONDRS>

US Department of Energy River Corridor Science Focus Area,  
Worldwide Hydrobiogeochemical Observation Network for  
Dynamic River Systems (WHONDRS)



## Sampling Site

<http://igsn.org/IEWDR000Q>

Site Name: S19S\_0009

## Sample Collection

## Source Sample Record

<http://igsn.org/IEWDR0133>

Sample Name: S19S\_0009\_Sediment-D  
Sampled Feature Type: stream  
Material: Sediment



Subsample (JGI/GOLD) Metagenome

<https://gold.jgi.doe.gov/project?id=Gp0503504>



Link (NCBI) BioSample

<https://www.ncbi.nlm.nih.gov/biosample/SAMN17619>



Link (EMSL) - Project

<https://www.emsl.pnnl.gov/project/48473>



Link (NMDC) - Project

<https://data.microbiomedata.org/details/study/gold:Gs0114663>



Link - Journal Publication

<https://doi.org/10.1128/mSystems.00151-18>

WHONDRS: a Community Resource for Studying  
Dynamic River Corridors



Link (ESS-DIVE) - Dataset

<https://doi.org/doi:10.15485/1729719>

WHONDRS Summer 2019 Sampling Campaign:  
Global River Corridor Sediment FTICR-MS,  
Dissolved Organic Carbon, Aerobic Respiration,  
Elemental Composition, and Grain Size.



Link (KBase) - Samples Narrative

<https://kbase.us/n/109073/41/>



Link (EMSL) - Metabolomics Data

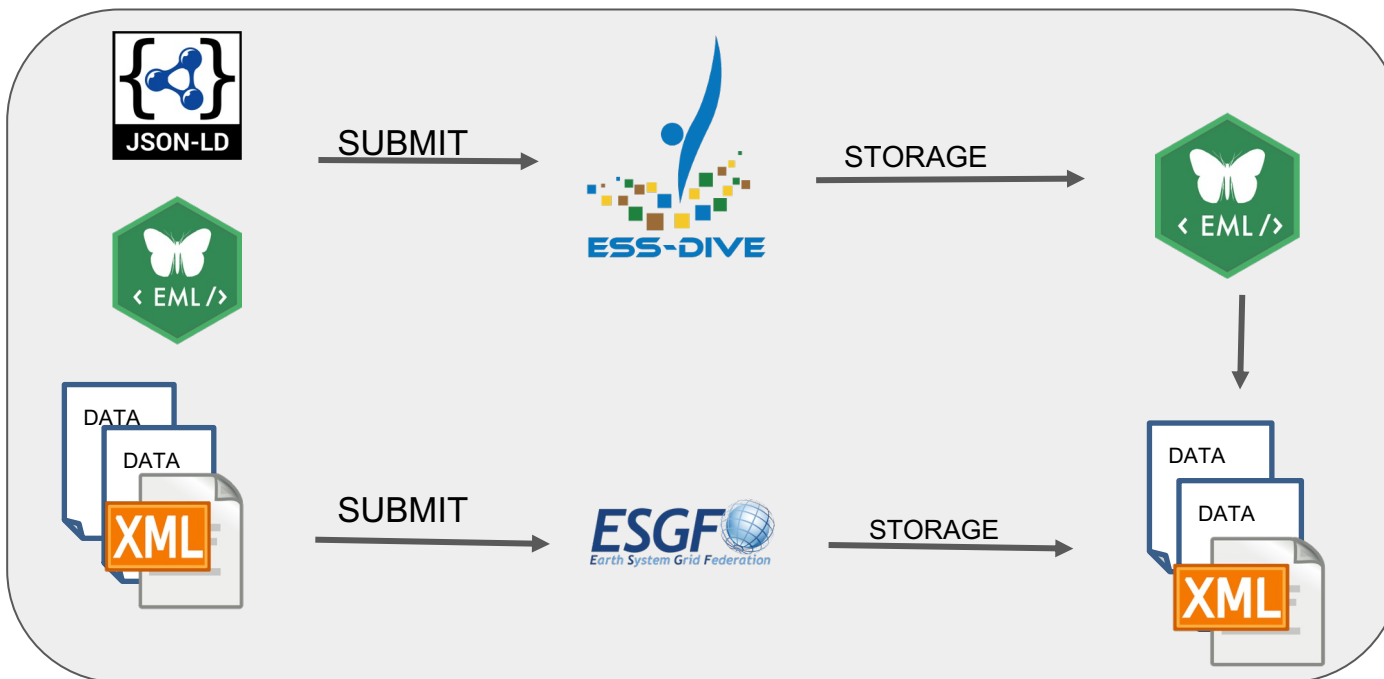
Description: FTICR, Metabolomics



# ESGF ESS-DIVE Integration



Making ESGF Metadata searchable in ESS-DIVE



Cross-walk to map ESGF metadata to ESS-DIVE EML standard

Metadata submitted and searchable in both systems

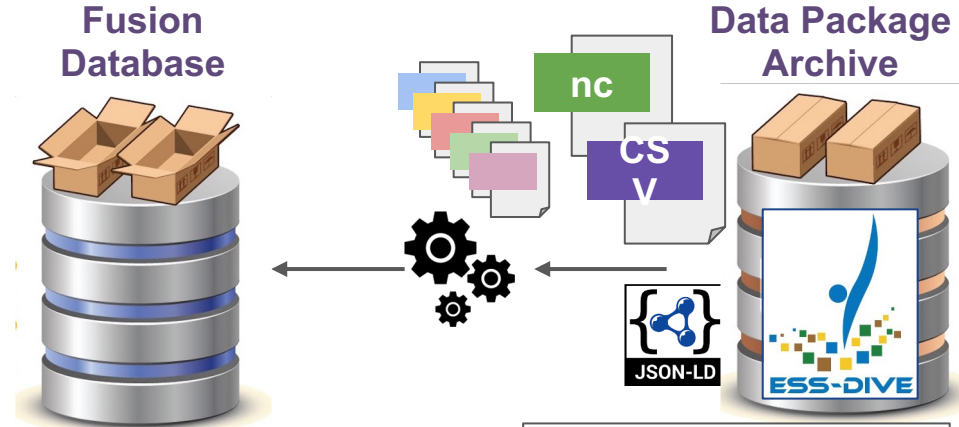
Use ESS-DIVE external dataset linking to link out to data in ESGF

Prototype work - index and link to datasets in ESGF from ESS-DIVE (Mario Melara, Sasha Ames 2020)

# Ongoing Development: Fusion Database

A way to make *any* standardized *data* searchable.

- Support established standards starting with ESS-DIVE file-level metadata reporting format
- Introspect data files and extract information that can be searchable
- Integrated search for scientific data and its metadata.



The screenshot shows the ESS-DIVE search interface. The search term "Nitrogen" is entered in the search bar. The results show "DATASETS 1 TO 25 OF 396". The first result is "Nitrogen" with a file icon. The second result is "Total nitrogen" with a file icon. The third result is "EARTH SCIENCE > LAND SURFA..." with a file icon. The fourth result is "EARTH SCIENCE > BIOSPHERE ..." with a file icon. The search results are displayed in a table format with columns for "Region description", "Creator", and "Year".

**Integrated Search**

The screenshot shows the ESS-DIVE analytics and visualization interface. It displays a code editor with the following code:

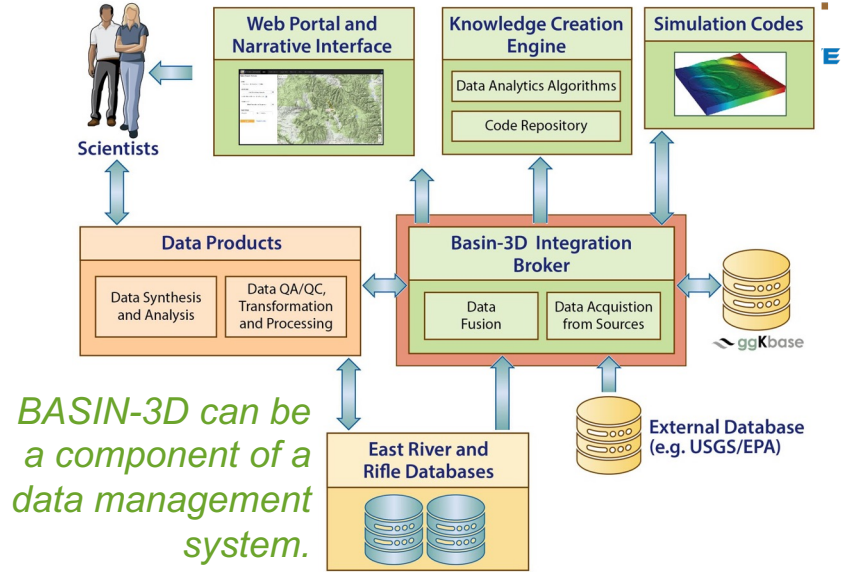
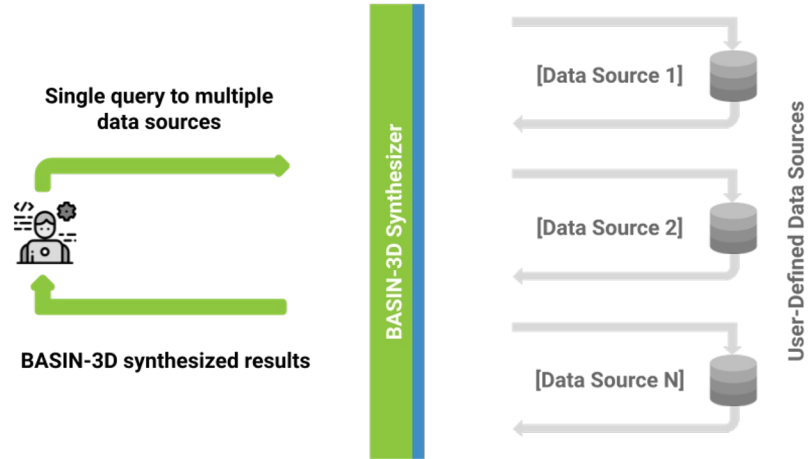
```
[9]: for x, v in usgs_metadata['USGS-01463500_DO'].items():  
    print(f'{x} = {v}')  
  
data_start = 2000-10-01 00:00:00  
data_end = 2019-08-31 00:00:00  
records = 6056  
basin_3d_variable = DO  
units = mg/L  
statistic = MEAN  
temporal_aggregation = DAY  
quality = CHECKED  
sampling_medium = WATER  
sampling_feature_id = USGS-01463500  
datasource = USGS  
datasource_variable = 00300  
***  
Distance (m/3h)  
Temperature (deg C)  
Conductivity (uS/cm)  
pH (pH/L)
```

The interface also displays several data plots: a time series plot of Distance (m/3h), a time series plot of Temperature (deg C), a time series plot of Conductivity (uS/cm), and a time series plot of pH (pH/L). The plots are labeled with "ESS-01463500" and "ESS-01463500" in the legend.

**Analytics & Viz**

# BASIN-3D: A framework for powering real-time data integration

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



BASIN-3D can be used within data infrastructures or by individual researchers using python-compatible tools like Jupyter notebooks.

BASIN-3D design is a plugin model to obtain data from any public or private network-connected source. It contains a generalized data model and the machinery to manage the data acquisition plugins.

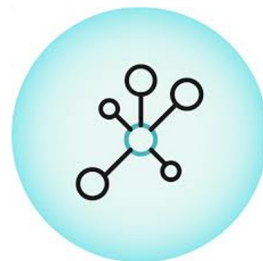
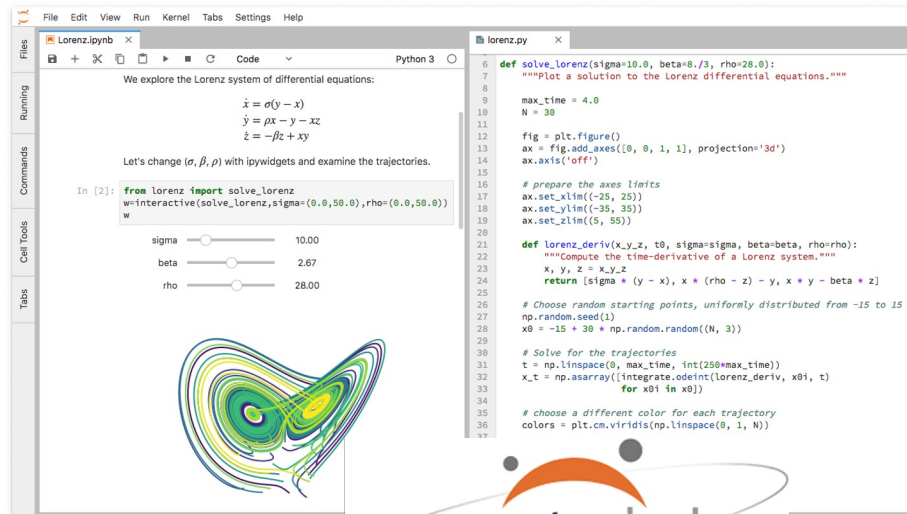
BASIN-3D development informed data transformation and synthesis concepts for the ESS-DIVE Fusion DB.



# Future Work on Data Integration and Compute



- Community Jupyterhub platform for users to directly operate on ESS-DIVE data.
- Automation of data publication pipeline across repositories.
  - eg. data on ESS-DIVE -> metadata for NMDC
- Related Identifiers to Support Linking Datasets and Citations
- Automated Data Quality Checking
- Dedicated Storage platform for synthesis/analysis (Ceph)





# Connect With Our Team!

<https://ess-dive.lbl.gov>  
[ess-dive-support@lbl.gov](mailto:ess-dive-support@lbl.gov)



To stay updated:

 [ess-dive-community@lbl.gov](mailto:ess-dive-community@lbl.gov)  
 [@essdive](https://twitter.com/essdive)



## Acknowledgements

Advisory Groups: ESS-DIVE Archive Partnership Board, ESS Cyberinfrastructure Working Groups  
ESS-DIVE is supported by the U.S. DOE - Office of Biological and Environmental Research, EESSD Data Management Program