ESS-DIVE: A DOE Earth and Environmental Data Repository

http://ess-dive.lbl.gov

Deb Agarwal, Charuleka Varadharajan, Shreyas Cholia, Cory Snavely, Valerie Hendrix, Joan Damerow, Fianna O'Brien, Hesham Elbashandy, Makayla Shepard, Bill Riley, Madison Burrus, Zarine Kakalia, Emily Robles, Maegen Simmonds, and Karen Whitenack

ESS-DIVE

Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

NERSC

NCEAS

The ESS-DIVE Team

Data Scientists and Software engineers

Digital Librarians

Environmental Scientists

Coordinator: Karen Whitenack

U.S. DEPARTMENT OF ENERGY | Office of Science

NERSC | BERKELEY LAB | NCEAS

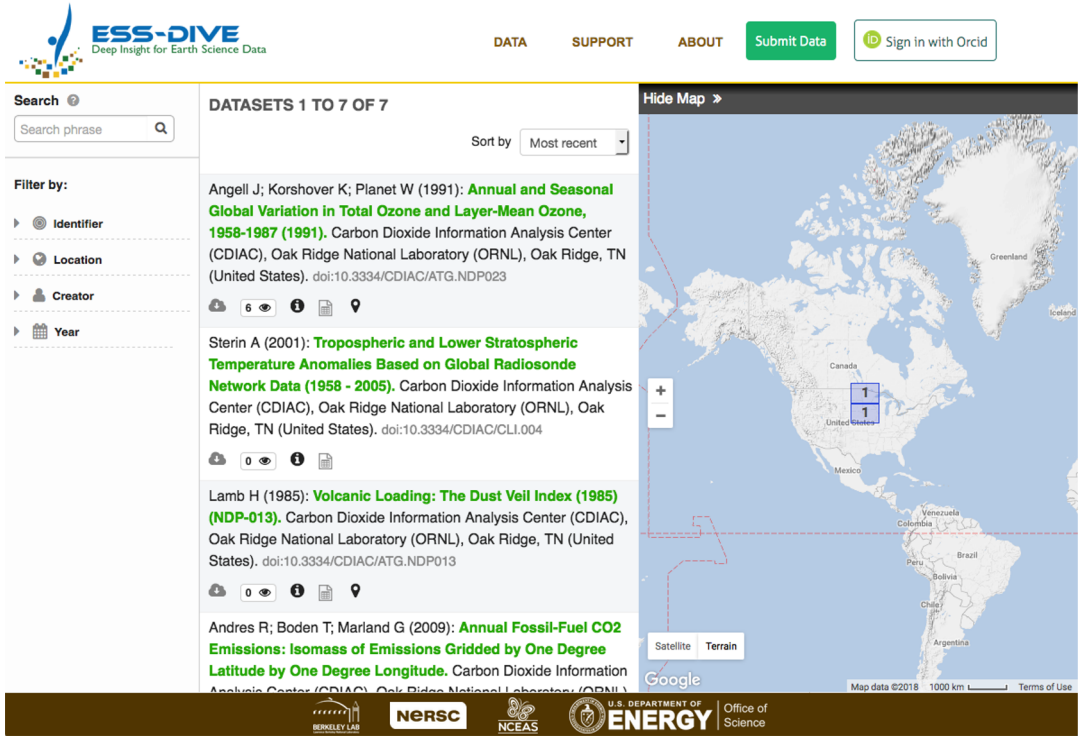
ESS-DIVE: a repository serving ESS data

- Key goals

  1. Broad partnership with the ESS community on repository capabilities and features

  2. Robust repository that provides long-term curation of ESS data

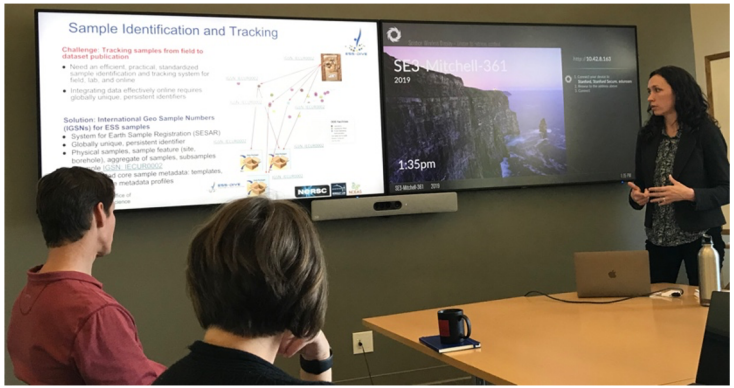
  3. Data that is Findable, Accessible, Interoperable, and Reusable

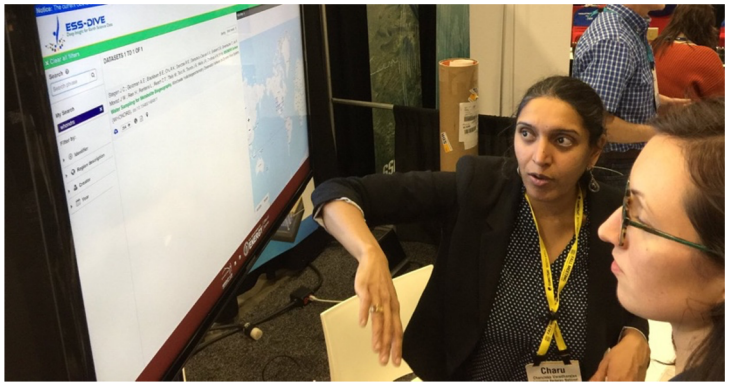


Community Engagement Activities

- Gathered ESS project requirements through: **Site visits, monthly webinars, advisory groups**
- Data and metadata **standards** development in partnership with community
- Demos and **tutorials**
- Representing ESS data activities in conferences/meetings

*Over **60** national and international meetings plus site visits to major ESS projects (3 years)*

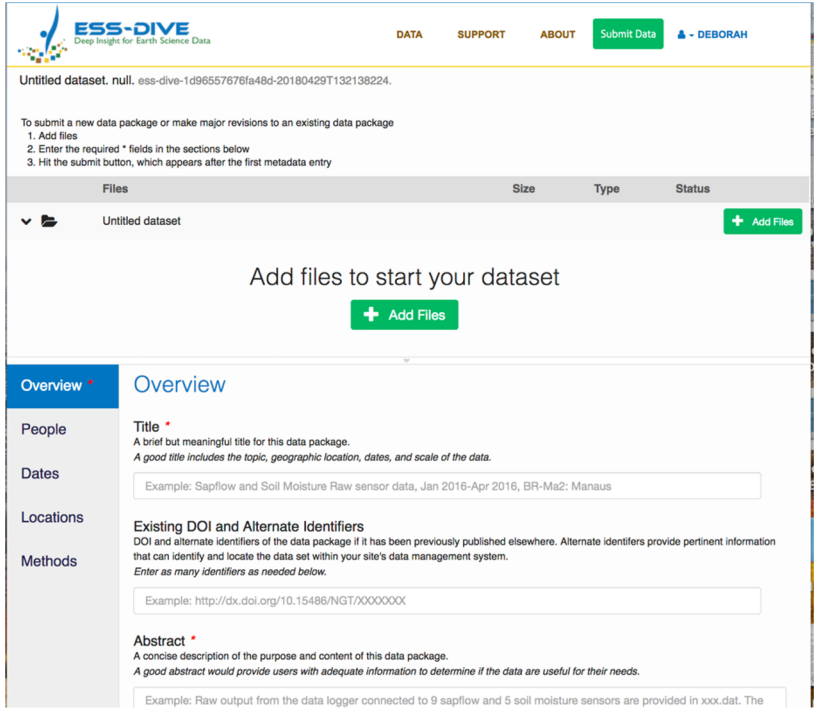




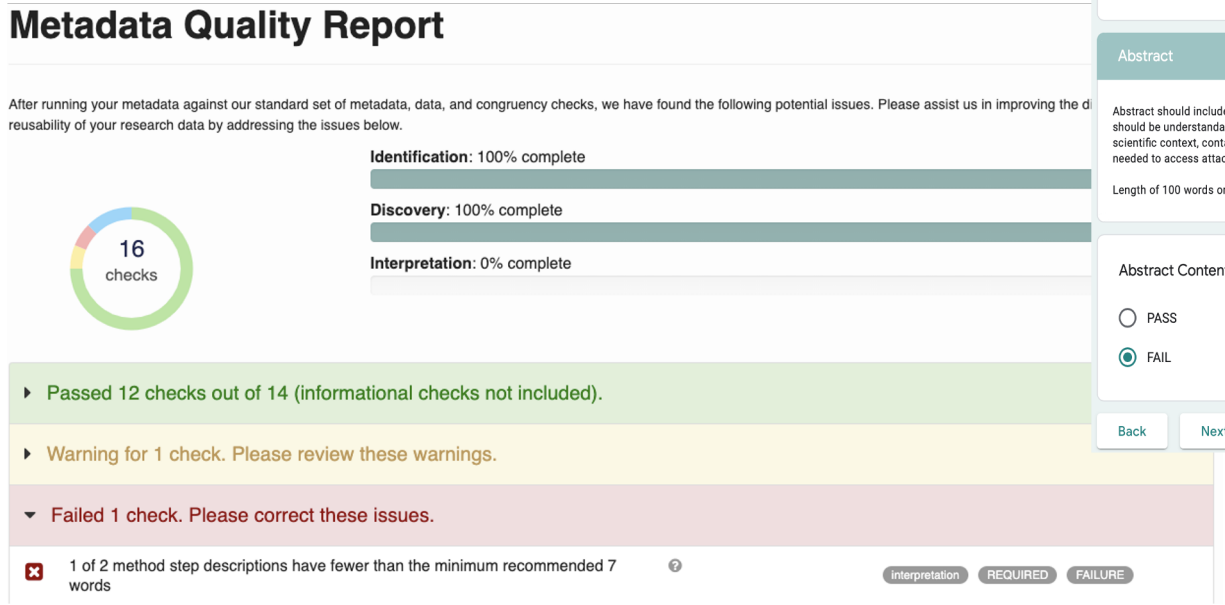
Submitting data: web portal and programmatic

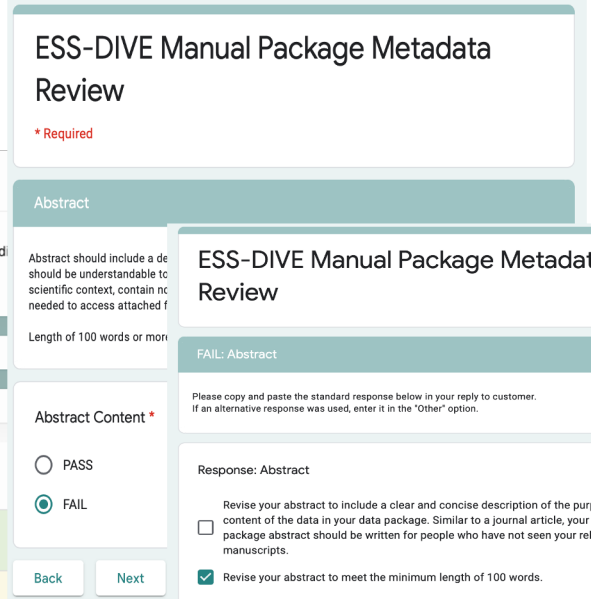
- **Upload** data and metadata to create a **data package** via web page or programmatic interface (API)
- Release data packages publicly with a digital object identifier (**DOI**)
- Data usage licensing **CCby4** and **CC0**
- Preview features on **http://data-sandbox.ess-dive.lbl.gov**

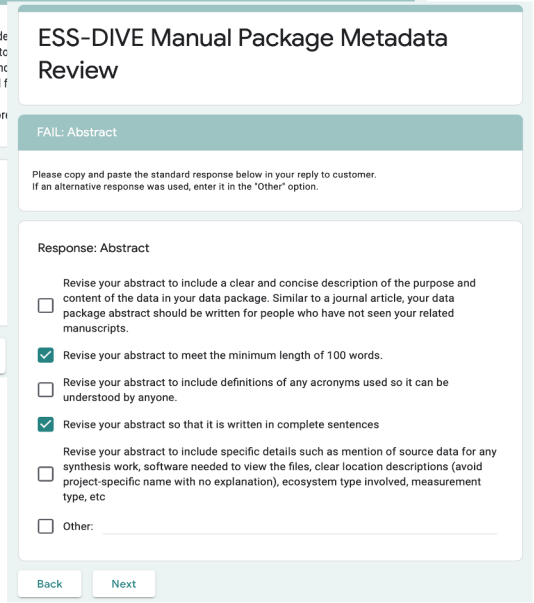
377 *data packages published (3 years)*



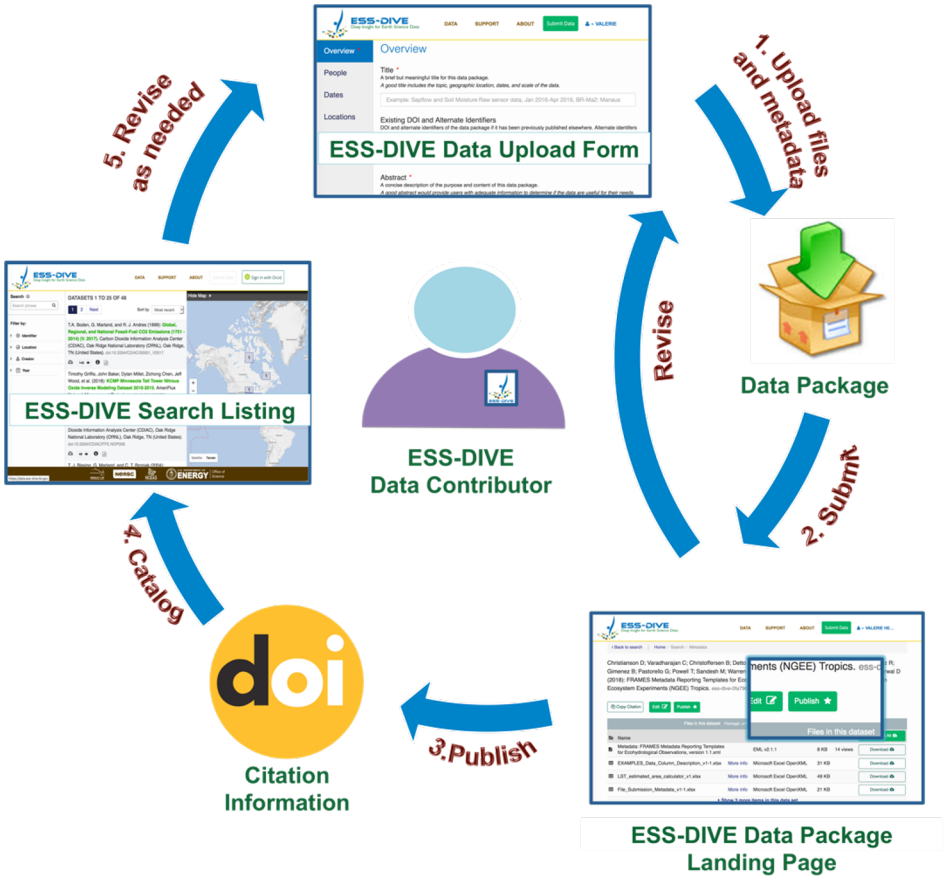
Data Package Quality Review



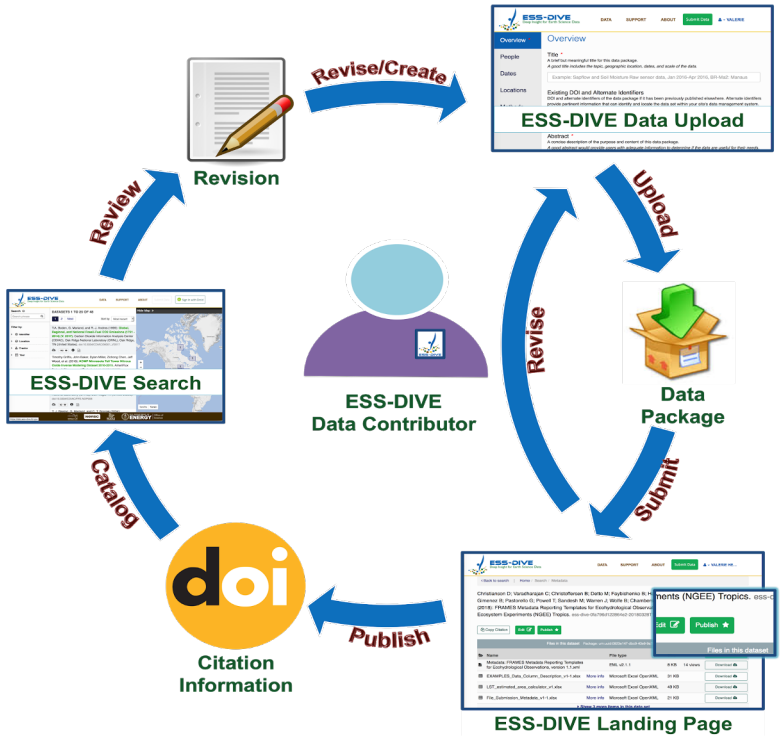


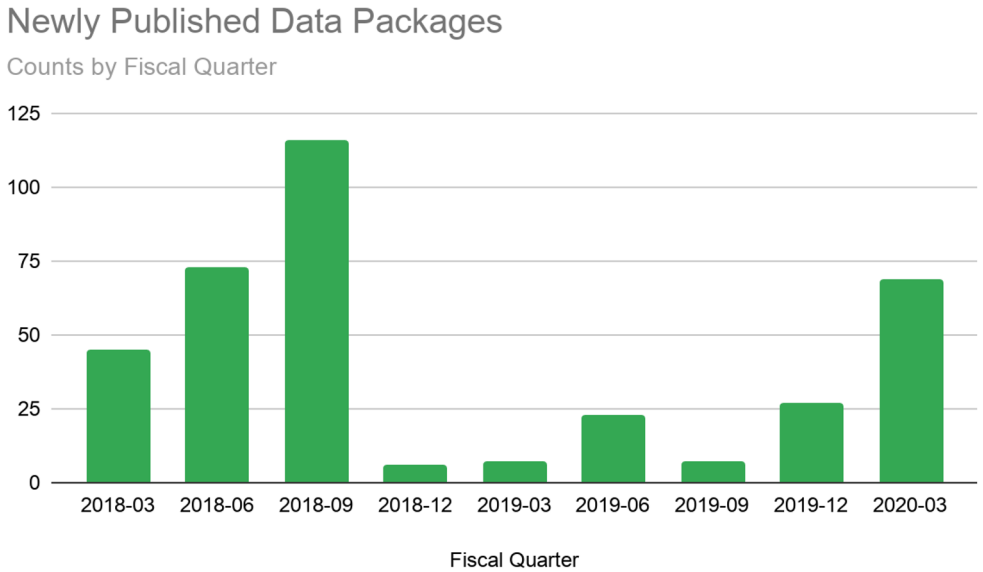


Publication Cycle

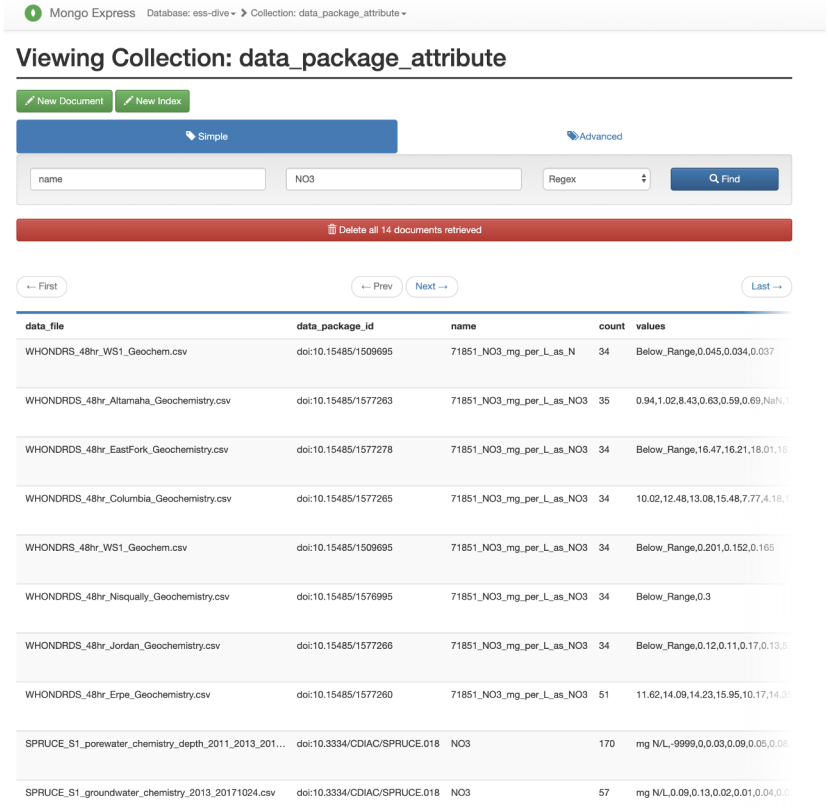


Data Publications





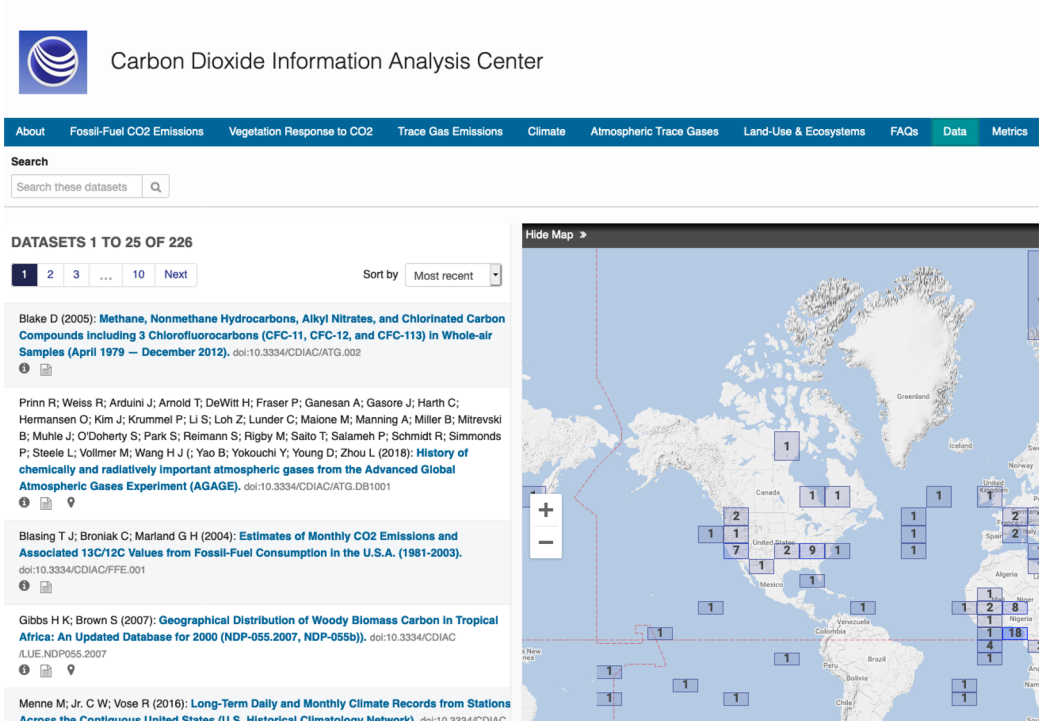
Prototype Data Fusion Database - Providing integrated search across data packages



- Purpose: make standardized data within data packages searchable
- Challenge: No standards yet adopted
- Prototype: reading the non-standardized comma separated value (csv) and Excel files
- Indexing everything found

Custom Data Portals – available in sandbox

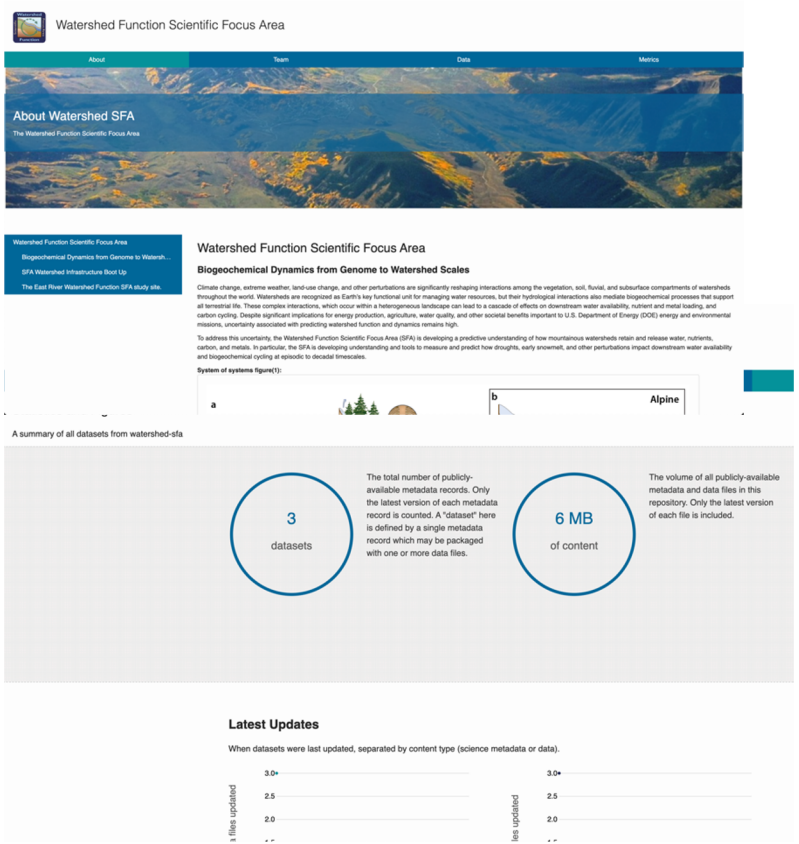
- Enable development of a topic or project specific view of ESS-DIVE
- Context web pages describing project or topic
- Custom data search
- Metrics specific to data
- Enables long-term view of project/topic including context
- Can be private or public



Upcoming Project Support Features:

Project Spaces: Project management and administration interfaces for use in managing data produced by projects.

- Enable project level curation and long-term maintenance
- Provide space for sharing data within project before it is made public
- Support data curation by project data manager
- Entry of data packages on behalf of data authors by project



ESS-DIVE Long-Term Vision
Connecting Users to Data

...
Water, water, every where,
And all the boards did shrink;
Water, water, every where,
Nor any drop to drink.
...

Data, data everywhere
and all the simulations drink
Data, data everywhere
All the users think

Developing Standards

ESS-DIVE Base Funds

- LBNL (Joan Damerow): **Sample metadata and identifiers**

ESS-DIVE Community Funds

- ORNL (Ranjeet Devarakonda): **Comma separated value files & File-level metadata**
- PNNL (Amy Goldman): **Hydrologic monitoring** data
- SLAC (Kristin Boye): **Sample-based Water quality** data
- PNNL (Ben Bond-Lamberty): **Soil respiration** data
- BNL (Alistair Rogers): **Leaf physiology** data
- ANL (Pamela Weisenhorn): **16S Amplicon** data

ESS-DIVE Collaboration Opportunities

- **Community funds:** What should be next?
- **Data Curation/Review:** Peer review?
- **Outreach:** Ambassadors?
- Other ideas?

Questions?

Contact: ess-dive-support@lbl.gov

ESS-DIVE

Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC

NCEAS

Persistent sample identifiers and documentation standards to support efficient sample management, interoperability, and data reuse in Environmental Systems Science

Joan Damerow, *Deb Agarwal, Kristin Boye, Eoin Brodie, Madison Burrus, Shreyas Cholia, Hesham Elbashandy, Ricardo Eloy Alves, Kim Ely, Amy E Goldman, Val Hendrix, Zarine Kakalia, Ken M Kemner, Annie B Kersting, Katharine Maher, Nancy Shiao-Lynn Merino, Fianna O'Brien, Zach Perzan, Emily Robles, Cory Snavely, Patrick Sorensen, James Stegen, Pamela Weisenhorn, Karen Whitenack, Mavrik Zavarin and Charuleka Varadharajan*

ESS-DIVE

Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC

NCEAS

Sample Naming and Tracking

COMMUNITY NEED

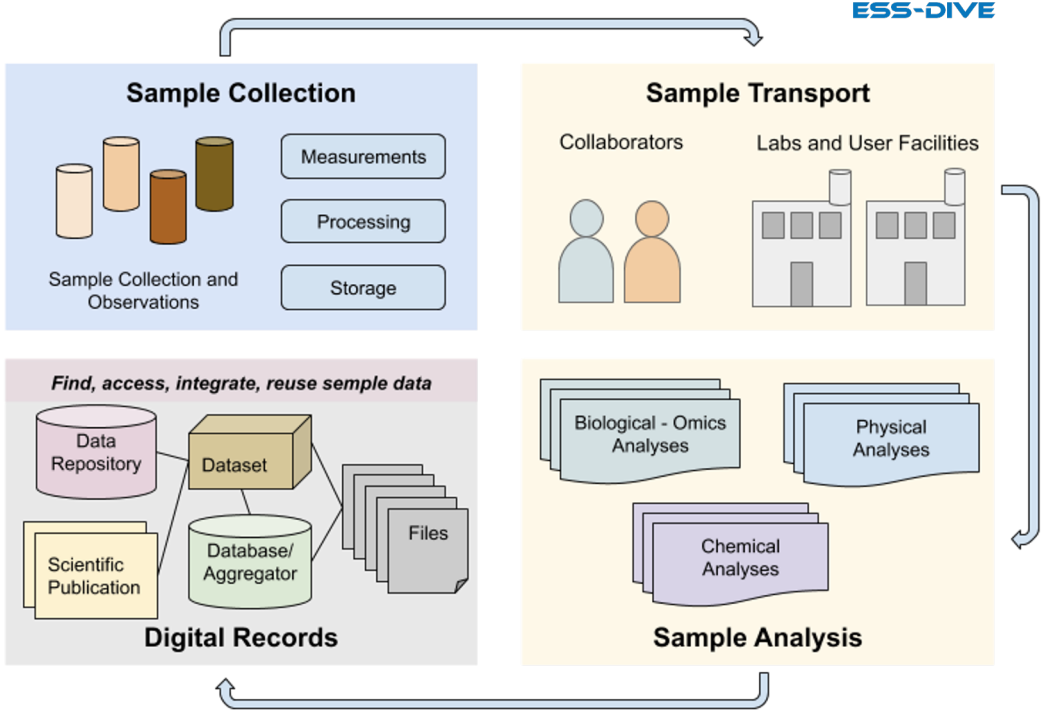
Practical, standardized sample ID/ tracking system as samples move to different physical and digital locations

SOLUTION

Integrating and preserving sample data effectively in digital environment requires globally unique, persistent identifiers

International Geo/General Sample Numbers (IGSNs) for ESS samples

- Standardized core sample metadata, templates
- Sample landing page - metadata profile, link to related samples and publications



IGSN: IECUR0002

Sample ID and Metadata Pilot Test Methods

Field test of IGSN PID and metadata

Sample Metadata Research and Crosswalk

Publish data packages in ESS-DIVE

1

2

3

4

5

6

Map sample relationships and history

PID for sample tracking and publication

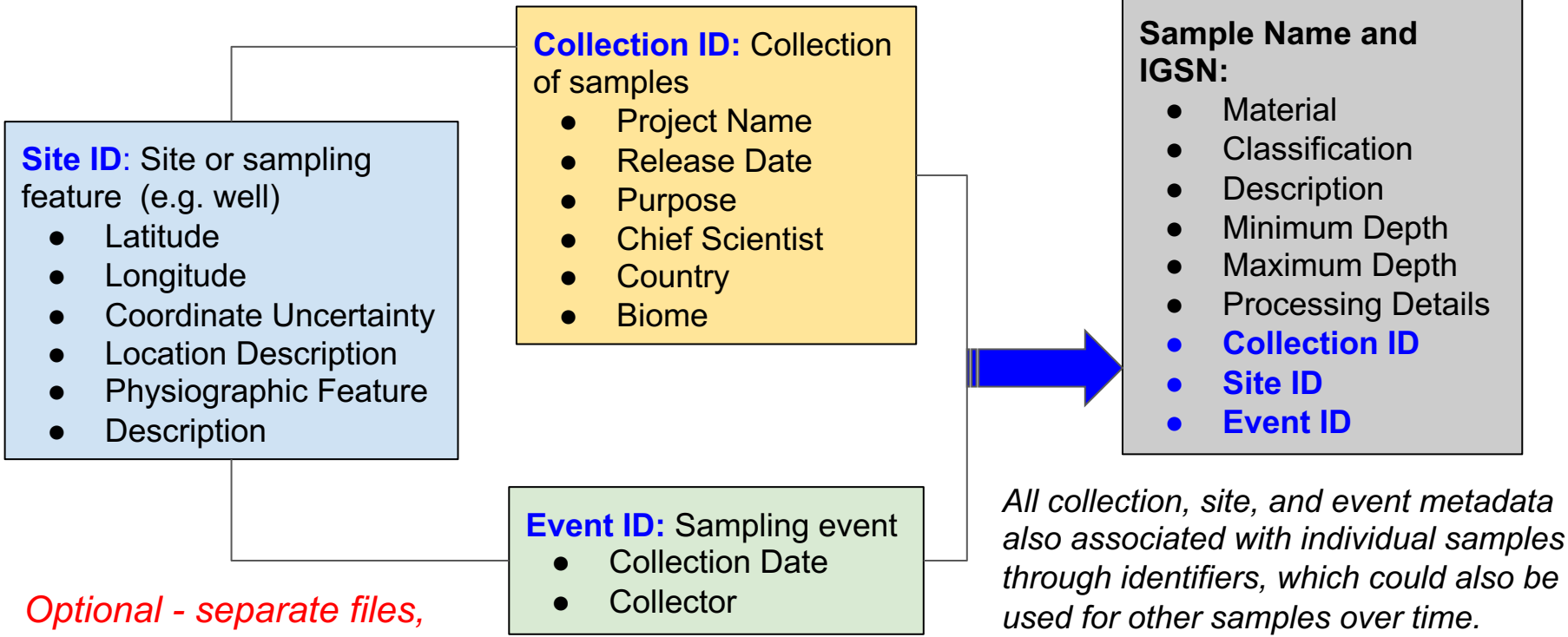
Incentive tools to support sample management and reuse.

8 Projects, ~ 3,000 Registered IGSNs

Review proposed sample ID and metadata standard:

- Crosswalk that compares existing sample metadata standards and templates
- Updated SESAR IGSN sample metadata guide accounting for pilot test feedback
- Updated SESAR IGSN spreadsheet template for sample metadata

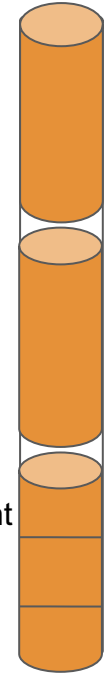
Providing Metadata and Linking Samples at Different Levels



Optional - separate files, terminology file mapping existing terms to standard, if needed

All collection, site, and event metadata also associated with individual samples through identifiers, which could also be used for other samples over time.

Assigning and linking IDs to sets or chains of highly related samples and subsamples



Soil Core

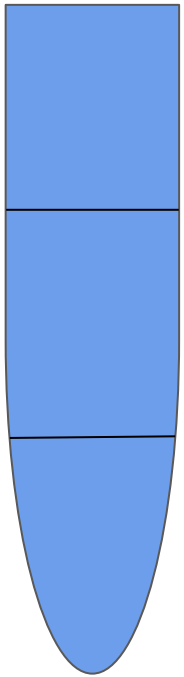
Core Section

Homogenized subsamples sent for analyses

Collection of water samples at one place and time

Varying depths

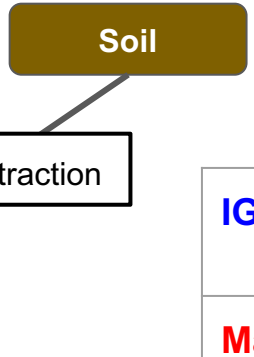
Individual containers sent for analyses



Relationship	1. Standard IGSNs	2. Extend primary IGSN	3. Collection or Site ID linked to IGSN
FeatureOfInterest (Parent)	*IGSN:* IEBWE0001	*IGSN:* IEBWE0001	Collection ID: S19S_0006
↓ isSampleOf (Parent/child) ↓	IGSN: IEBWE000P IEBWE000Q	*extensions:* IEBWE00011M IEBWE00012M	*IGSN:* IEBWE000P IEBWE000Q
isSampleOf (child)	*IGSN:* IEBWE000R IEBWE000S IEBWE000T ...	*IGSN:* IEBWE00012MA IEBWE00012MB IEBWE00012MC ...	*extensions:* IEBWE000PA IEBWE000PB IEBWE000PC ...

****Subsequent IDs assigned in external labs or user facilities are separate but always associated with the sample PID***

Sample Interoperability with NMDC, KBase, JGI, EMSL



IGSN: IEBWE000P ⟷ **BioSample: SAMN04007148**

IGSN	Source_material_id = IGSN Link
Material	env_medium* = ENVO; isolation_source
populate from material	organism* (e.g. soil metagenome)
Physiographic feature	Env_local_scale* = ENVO
	Env_broad_scale* = ENVO
Country; Locality	Geo_loc_name* = GAZ
	samp_mat_process

- Links between IDs: Source material ID and subsample/analysis ID
- IGSN metadata → MIxS/MIMS
- Use of standard ontologies - mapping ontologies to terms in use
- Use cases: WHONDRS, LBNL TES SFA FICUS

Questions?

Contact: ess-dive-support@lbl.gov

ESS-DIVE

Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC

NCEAS

OAK RIDGE National Laboratory

Developing CSV and File-level Metadata Standards

ORNL and ESS-DIVE
Community Fund Standards

May 11, 2020

Terri Velliquette <velliquettet@ornl.gov>
Jessica Welch
Michael Crow
Ranjeet Devarakonda
Susan Heinz

ORNL Community Fund Project is to develop two standards:

- a generic standard for the structure and content requirements of data stored in a comma-separated values file referred to as a CSV file.
- a metadata schema to describe the contents, scope, and structure of each data file within the ESS-DIVE repository referred to as file-level metadata or FLMD.

These standards will promote:

- machine-readability for the efficient extraction of file-level discovery and usable metadata.
- interoperability to facilitate the development of a future fusion database.

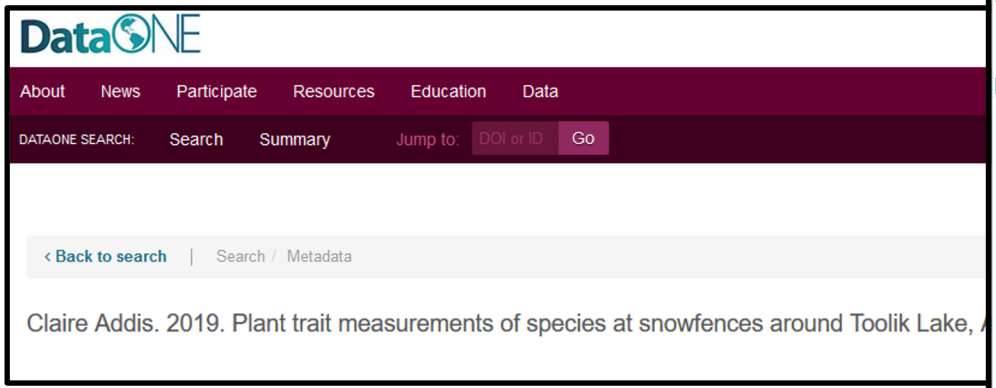
ORNL is managed by UT-Battelle, LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

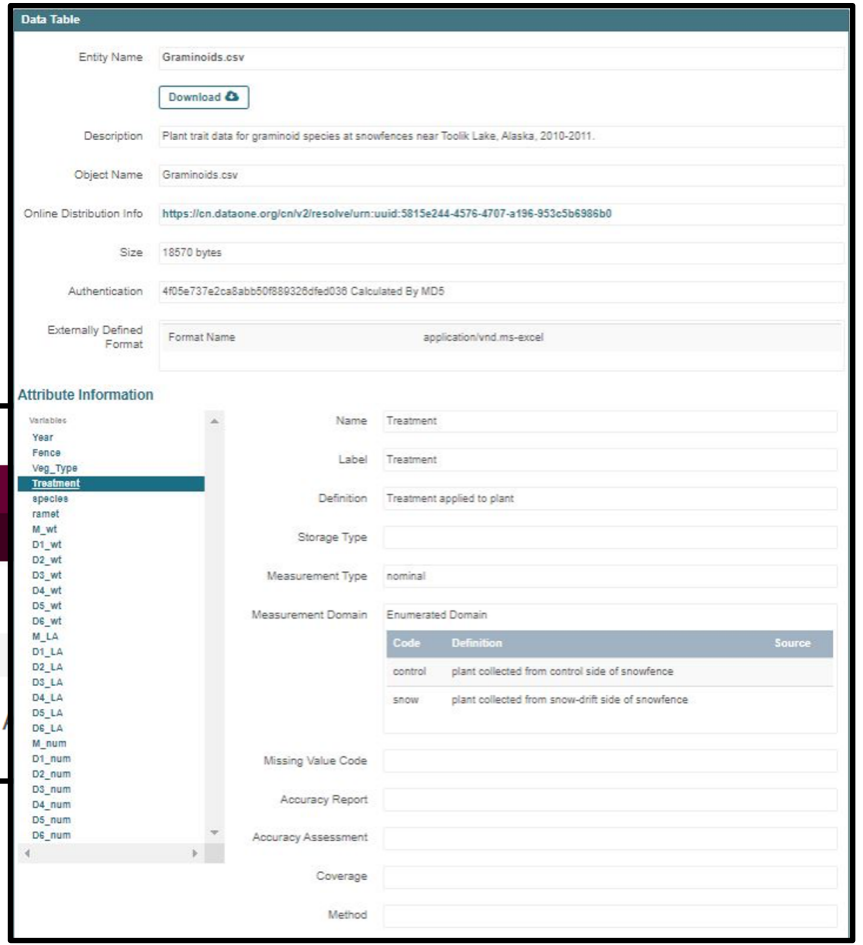
An Example of File-level Metadata (FLMD)

FLMD can provide:

- granular information that can enable data users to compare data files within a data package
- the ability to search for and locate files across an overarching data collection



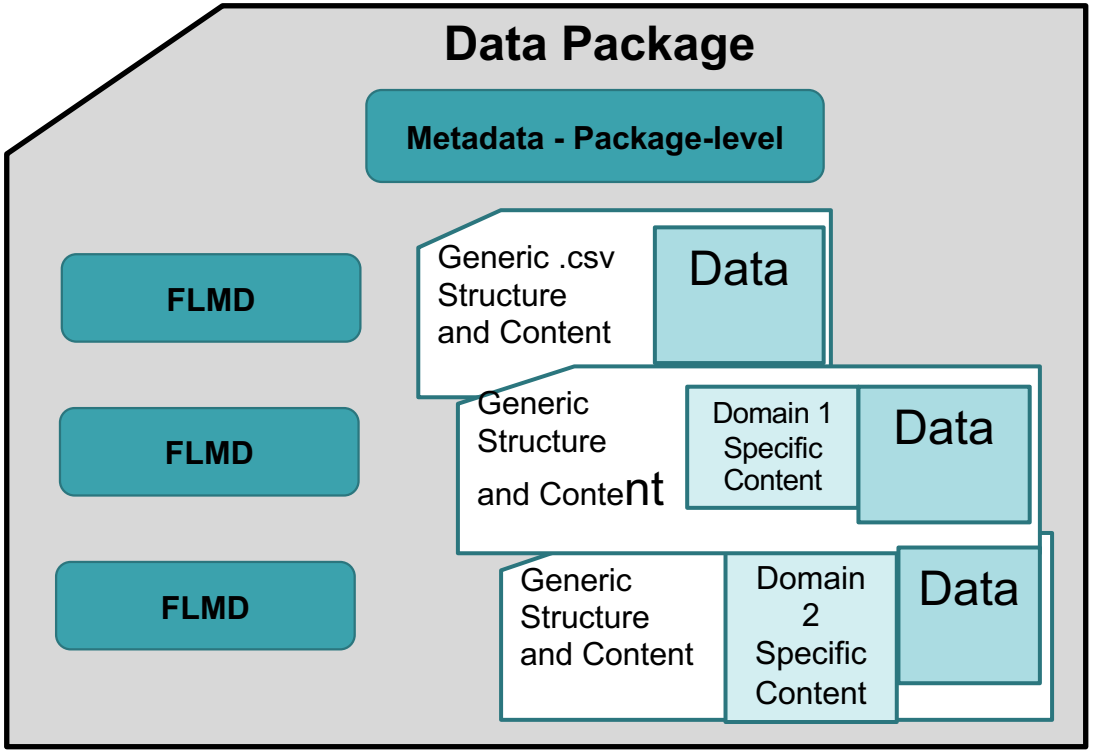
https://search.dataone.org/view/doi:10.18739/A26Q1SH2V



OAK RIDGE National Laboratory

CSV Standards and Format

- The generic structure and content requirements are the foundation for standardized and consistent data and metadata reporting. The goal is to have general data and domain-specific content and data fit within the larger common file structure.



Next Steps

Template:

- Develop example template for *.csv file

Format Check:

- Implement CSV file format verification script

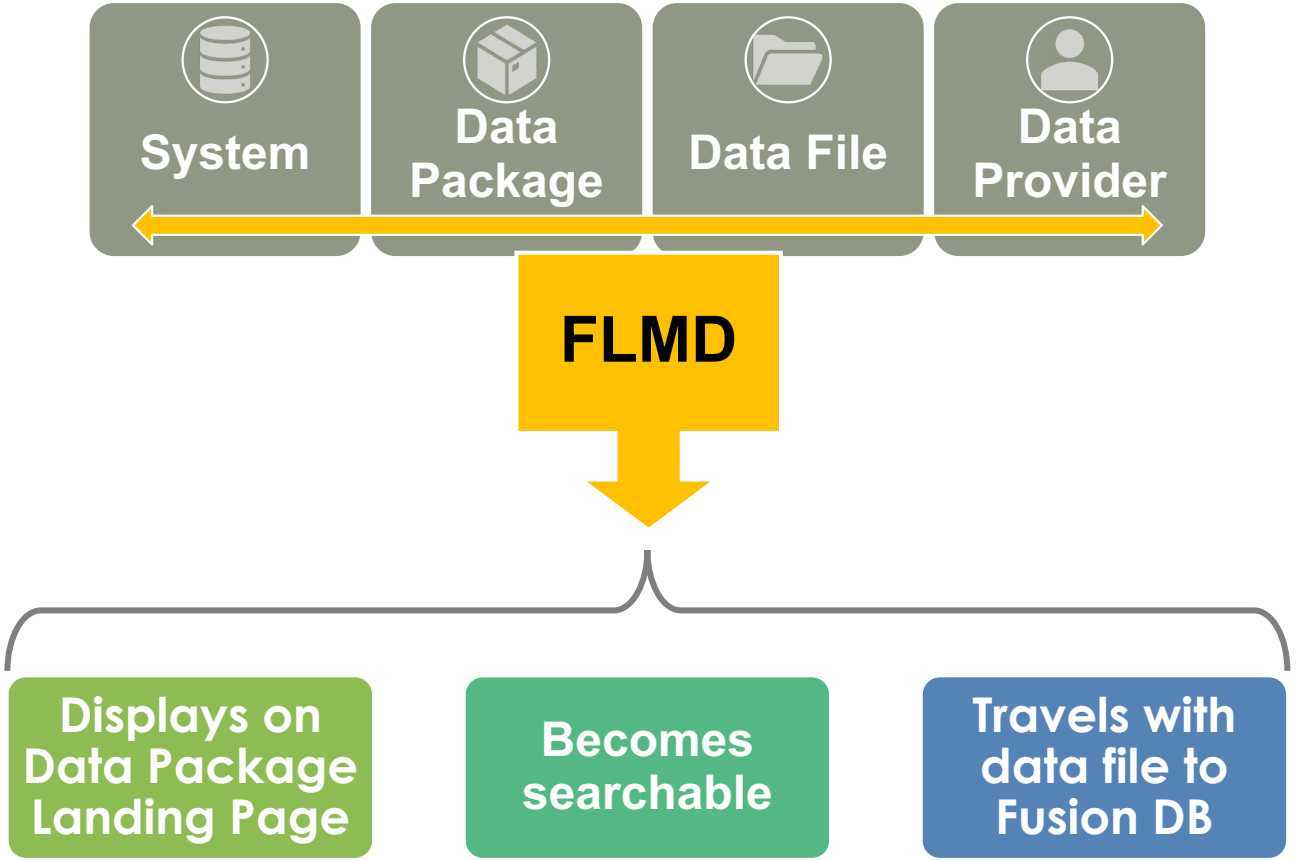
File Level Metadata Extraction Tool:

- Implement a script that will extract FLMD from a standard *.csv file and store the metadata for ESS-DIVE Search applications.

Possible Sources for FLMD and Potential Uses of FLMD

FLMD

- Can be populated from multiple sources (following standard formats)
- Can display with each data file on the Data Package Landing Page
- Can be searched
- Can travel with the data file if pulled into a fusion database



Pacific Northwest
NATIONAL LABORATORY

Hydrologic Monitoring Standards

May 11, 2020

Amy Goldman, Huiying Ren, Josh Torgeson, Huifen Zhou

ESS Cyberinfrastructure Working Group Annual Meeting

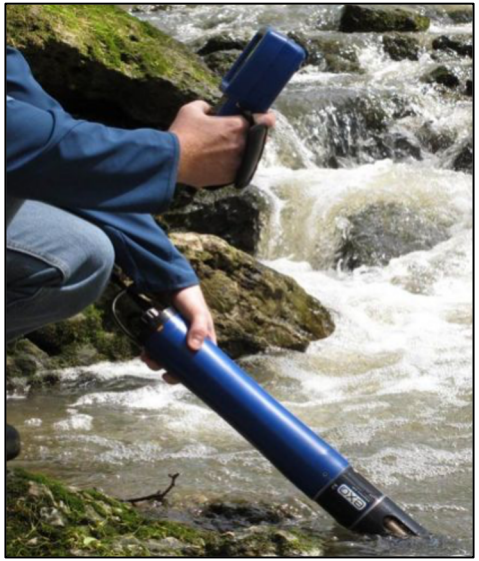
U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Hydrologic monitoring data often lack consistency and metadata

Scope: Sensors used in the field to measure parameters in water

Parameters: "Water level", temperature, electrical and specific conductivity, pH, dissolved oxygen



Challenges

- Rarely define terms
- Often not machine readable
- Rarely have information about instrumentation

Examples of ambiguous column headers	Possible questions from data users
Elevation	Surface water elevation? Riverbed elevation? Using what datum?
Level_m	Level of what? Relative to what?
Well depth	Depth to water from the well casing? Water elevation in the well?

Interested in taking a closer look and providing feedback? https://tinyurl.com/HydroStandard

www.xylem-analytics.com.au%2Fproductsdetail.php%3FYSI-EXO2-Multiparameter-Sonde

Three-file system captures sensor metadata and allows flexible terminology

Data File	Sensor Metadata File	Terminology File
• *Time stamps • Measurements • Alphanumeric code linking each parameter at a specific time point to sensor metadata file • Alphanumeric data flag codes	• Alphanumeric sensor codes from data file • Sensor construction (make, model, serial number, accuracy, range, calibration, QAQC, notes) • Deployment detail (lat/long, *water body type, context of sensor placement, description of deployment)	• *Terms used in headers, data flags, missing value cells • *Definition of terms • *Mapping the used terms to the terms suggested by the standard

Required fields are marked with *

Interested in taking a closer look and providing feedback? https://tinyurl.com/HydroStandard

Columns below are a subset of the standard's suggestions. Most are optional. Required columns are marked with*

Not the preferred term

Sensor code column for the parameter matches data column term

Data File					
*DateTime_UTC	cond	WaterTemp_deg.C	DataFlag	Sensor_cond	Sensor_WaterTemp
2020-05-14 13:30:00	252	15.2	DataFlag_01	S1_SpC	S3_T

Sensor Metadata File								
SensorID	SensorMake	SensorModel	SensorSN	SensorAccuracy	*Latitude_dec.deg	*Longitude_dec.deg	*WaterBodyType	SensorContext
S1_SpC	Campbell Scientific	CS547A	41593	+/- 5% of reading	43.3195	-119.2593	groundwater [ENVO:01001004]	Well

Terminology File		
*Term	*Definition	*MapToSuggested
cond	Specific conductivity in microSiemans per centimeter	SpC_uS.per.cm

Interested in taking a closer look and providing feedback? https://tinyurl.com/HydroStandard

ESS-DIVE
Deep Insight for Earth Science Data

Chemical concentrations in water, soil, and sediment samples

A community funded project for developing data and metadata standards and templates for ESS-DIVE archived data

Kristin Boye, Associate Staff Scientist, SLAC

U.S. DEPARTMENT OF ENERGY
Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Relationships and structure overview

[redacted]

Standard by Joan Damerow

Data
- Sample information
- Methods information
- Observable/variable
- Unit
- Time stamp
- Data quality/flags
- Measurement value

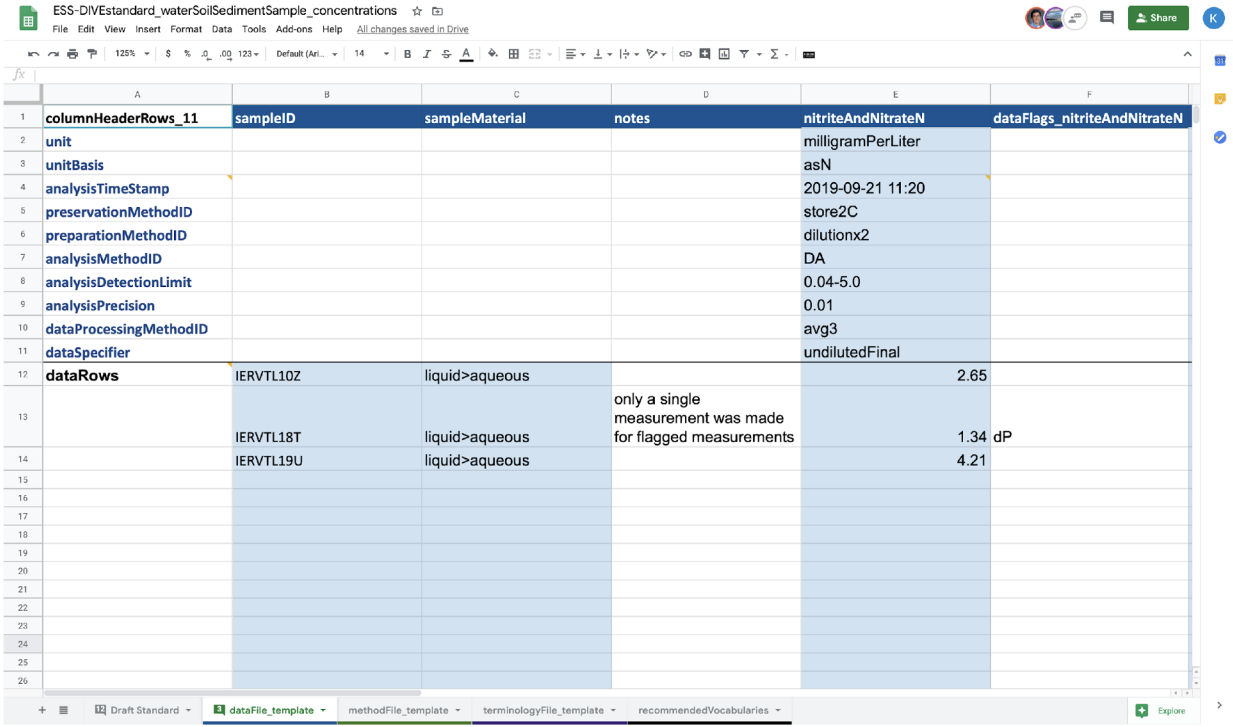
Methodology
- Preservation
- Preparation
- Analysis
- Data Processing

Terminology
- Codes
- Definitions

Examples:
- Quality flags
- Missing data value
- Alternate vocabulary

Current Data File Template

Data is reported in columns with key metadata included as header rows

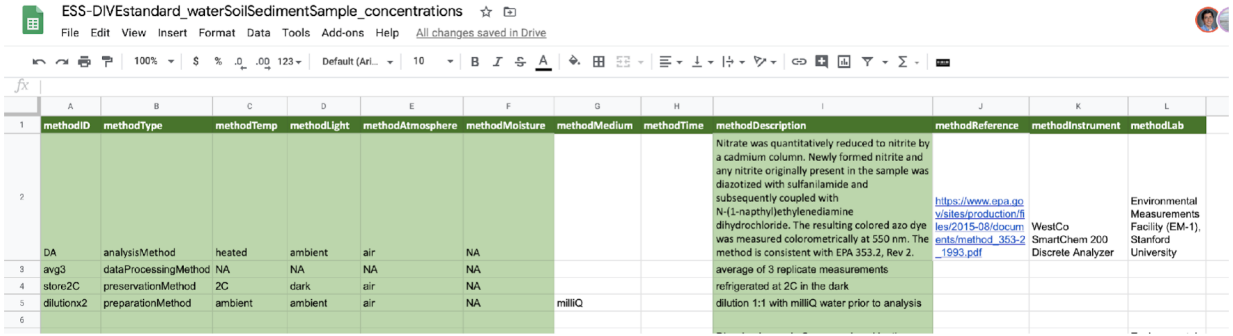


	A	B	C	D	E	F
1	columnHeaderRows_11	sampleID	sampleMaterial	notes	nitriteAndNitrateN	dataFlags_nitriteAndNitrateN
2	unit				milligramPerLiter	
3	unitBasis				asN	
4	analysisTimeStamp				2019-09-21 11:20	
5	preservationMethodID				store2C	
6	preparationMethodID				dilutionx2	
7	analysisMethodID				DA	
8	analysisDetectionLimit				0.04-5.0	
9	analysisPrecision				0.01	
10	dataProcessingMethodID				avg3	
11	dataSpecifier				undilutedFinal	
12	dataRows	IERVTL10Z	liquid>aqueous		2.65	
13		IERVTL18T	liquid>aqueous	only a single measurement was made for flagged measurements	1.34	dP
14		IERVTL19U	liquid>aqueous		4.21	

Current Methods and Terminology File Templates

Methods are linked through methodIDs

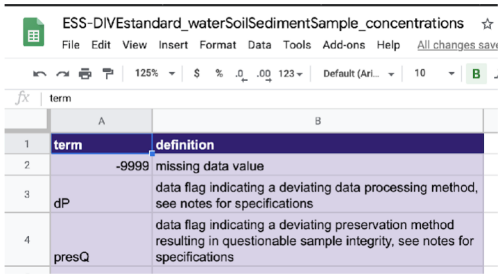
(defined by user or registered IGSNs or other permanent identifier)


ESS-DIVEstandard_waterSoilSedimentSample_concentrations

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

	A	B	C	D	E	F	G	H	I	J	K	L
1	methodID	methodType	methodTemp	methodLight	methodAtmosphere	methodMoisture	methodMedium	methodTime	methodDescription	methodReference	methodInstrument	methodLab
2	DA	analysisMethod	heated	ambient	air	NA			Nitrate was quantitatively reduced to nitrite by a cadmium column. Newly formed nitrite and any nitrite originally present in the sample was diazotized with sulfanilamide and subsequently coupled with N-(1-napthyl)ethylenediamine dihydrochloride. The resulting colored azo dye was measured colorometrically at 550 nm. The method is consistent with EPA 353.2, Rev 2.	https://www.epa.gov/sites/production/files/2015-08/documents/method_353-2_1993.pdf	WestCo SmartChem 200 Discrete Analyzer	Environmental Measurements Facility (EM-1), Stanford University
3	avg3	dataProcessingMethod	NA	NA	NA	NA			average of 3 replicate measurements			
4	store2C	preservationMethod	2C	dark	air	NA			refrigerated at 2C in the dark			
5	dilutionx2	preparationMethod	ambient	ambient	air	NA	milliQ		dilution 1:1 with milliQ water prior to analysis			
6												

Flags, codes, alternative vocabularies etc. are reported in a separate terminology file


ESS-DIVEstandard_waterSoilSedimentSample_concentrations

File Edit View Insert Format Data Tools Add-ons Help All changes save

term

	A	B
1	term	definition
2	-9999	missing data value
3	dP	data flag indicating a deviating data processing method, see notes for specifications
4	presQ	data flag indicating a deviating preservation method resulting in questionable sample integrity, see notes for specifications

PLEASE provide input!!!

This is a living document and a standard in progress - it will be most **applicable**, **appropriate**, **efficient**, and **functional** if it is adopted and accepted, which means *adapted* to **YOUR** needs and practices!!!

PLEASE reach out to me with comments or to be part of the review team for this standard

kboye@slac.stanford.edu

Continuous soil respiration and other chamber fluxes

May 11, 2020

Ben Bond-Lamberty and Stephanie Pennington

ESS-DIVE

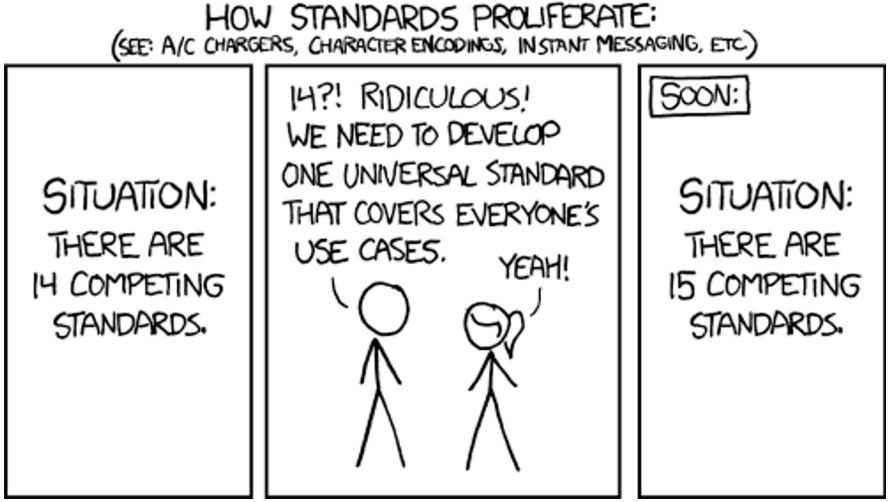
Deep Insight for Earth Science Data

Orphaned datasets limit science

- Large fluxes of greenhouse gases (GHGs) between soils and the atmosphere
- Continuous (automated) measurements increasingly common
- No standard or repository for them
- Potentially *huge* amount of data bearing on dynamics and response of earth system

Philosophy, approach, standards

- "As simple as possible, but not simpler"
- Can't be all things to all people
- Leverage historical experience
- Learn from recent ISRaD and SIDb databases
- Consultation with ICOS, FLUXNET, Ameriflux, manufacturers
- Community engagement and surveys



Philosophy, approach, standards

- Location in *space* and *time*
- Measured flux(es)
- Provenance
- Contributors and reference

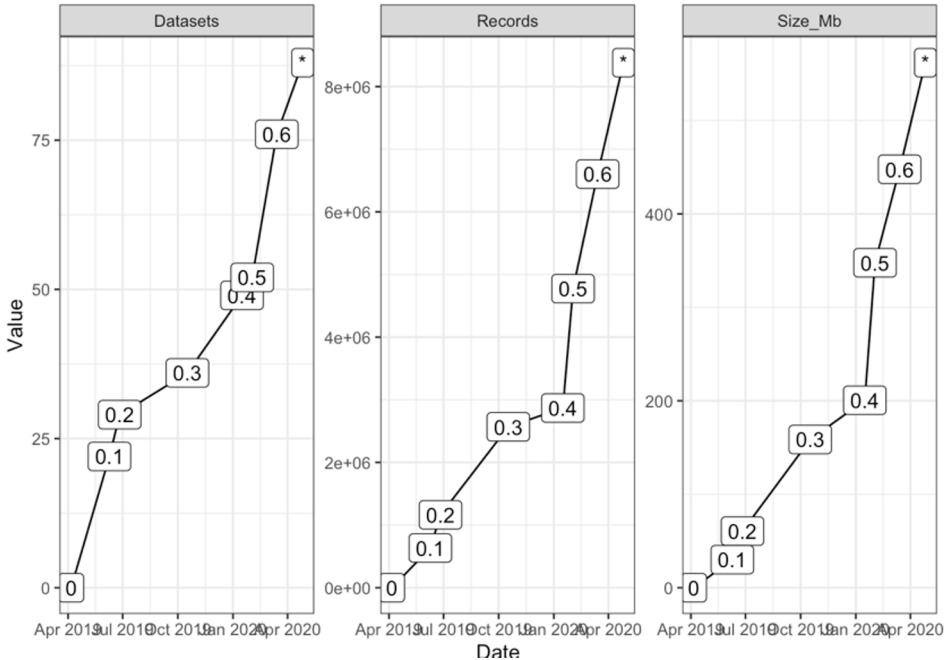
- Instrument, errors, ecosystem, errors, ancillary data…

- Each field added has a **cost** that has to be paid in perpetuity!

Field name	Description	Class	Units	Required
CSR_DATASET	Dataset name	character		*
CSR_SITE_NAME	Site name	character		*
CSR_LONGITUDE	Decimal longitude of site, positive = north	numeric	degrees	*
CSR_LATITUDE	Decimal latitude of site, positive = east	numeric	degrees	*
CSR_ELEVATION	Elevation of site	numeric	m	*
CSR_TIMEZONE	Timezone code, from https://en.wikipedia.org/wiki/List_of_tz_database_time_zones	character		*
CSR_IGBP	IGBP class, from http://www.eomf.ou.edu/static/IGBP.pdf	character		*
CSR_NETWORK	Site network name	character		
CSR_SITE_ID	Site ID in network	character		
CSR_INSTRUMENT	Measurement instrument (i.e. model)	character		*

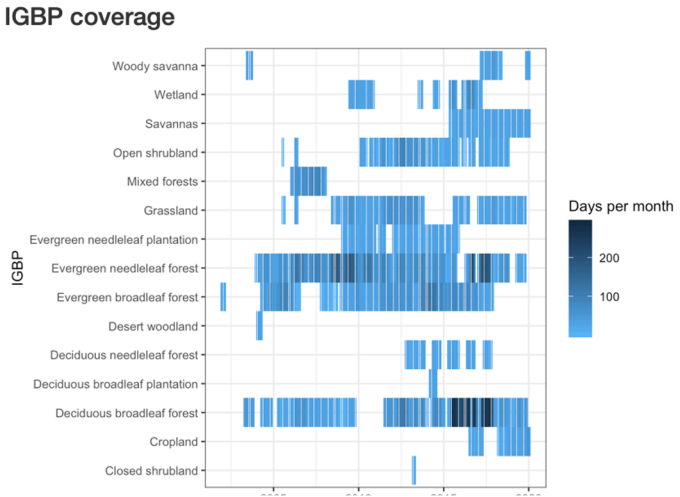
Concurrent evolution

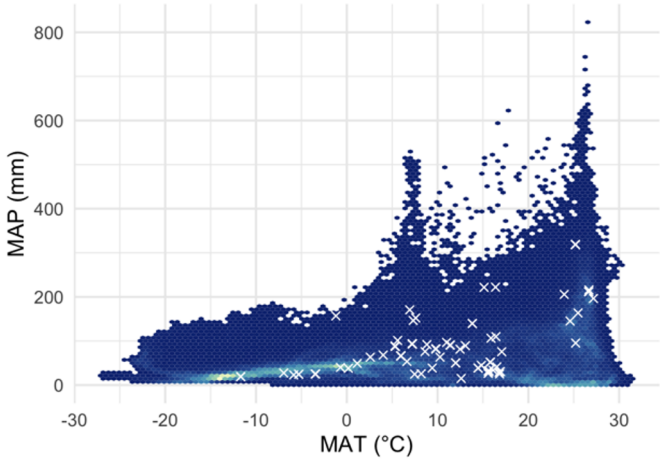
- Building *slowly* and *concurrently* : data and standard
- This has given real-world experience with kinds and formats of data
- ...and therefore the costs and benefits of different standards and formats
- With well-designed tools and software this process is worth it



An increasingly mature data standard and database

- Data standard increasingly refined
- Concrete mapping to Ameriflux BADM
- *ESSD* manuscript in prep
- Data available now!
- https://github.com/bpbond/cosore





ESS-DIVE data and metadata standards

Leaf physiology

Kim Ely & Alistair Rogers

Barrow Environmental Observatory, AK

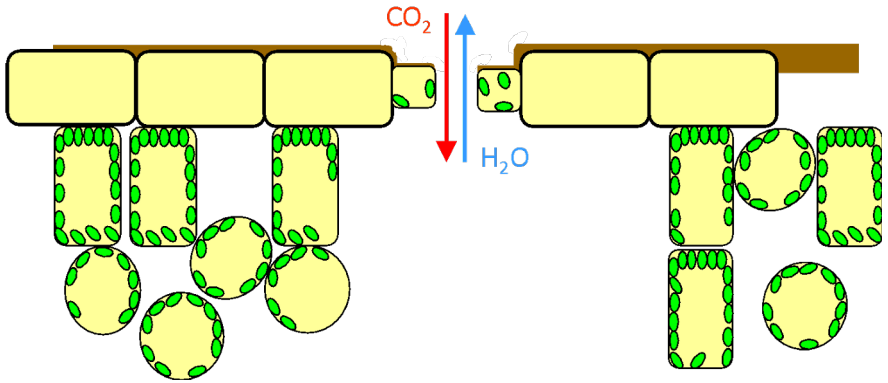
U.S. DEPARTMENT OF ENERGY | Office of Science

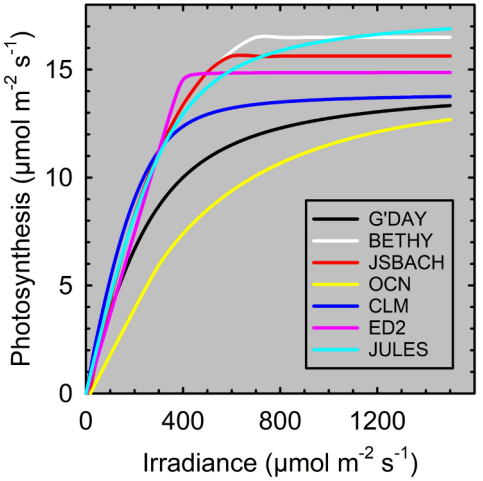
ESS-DIVE

BROOKHAVEN NATIONAL LABORATORY

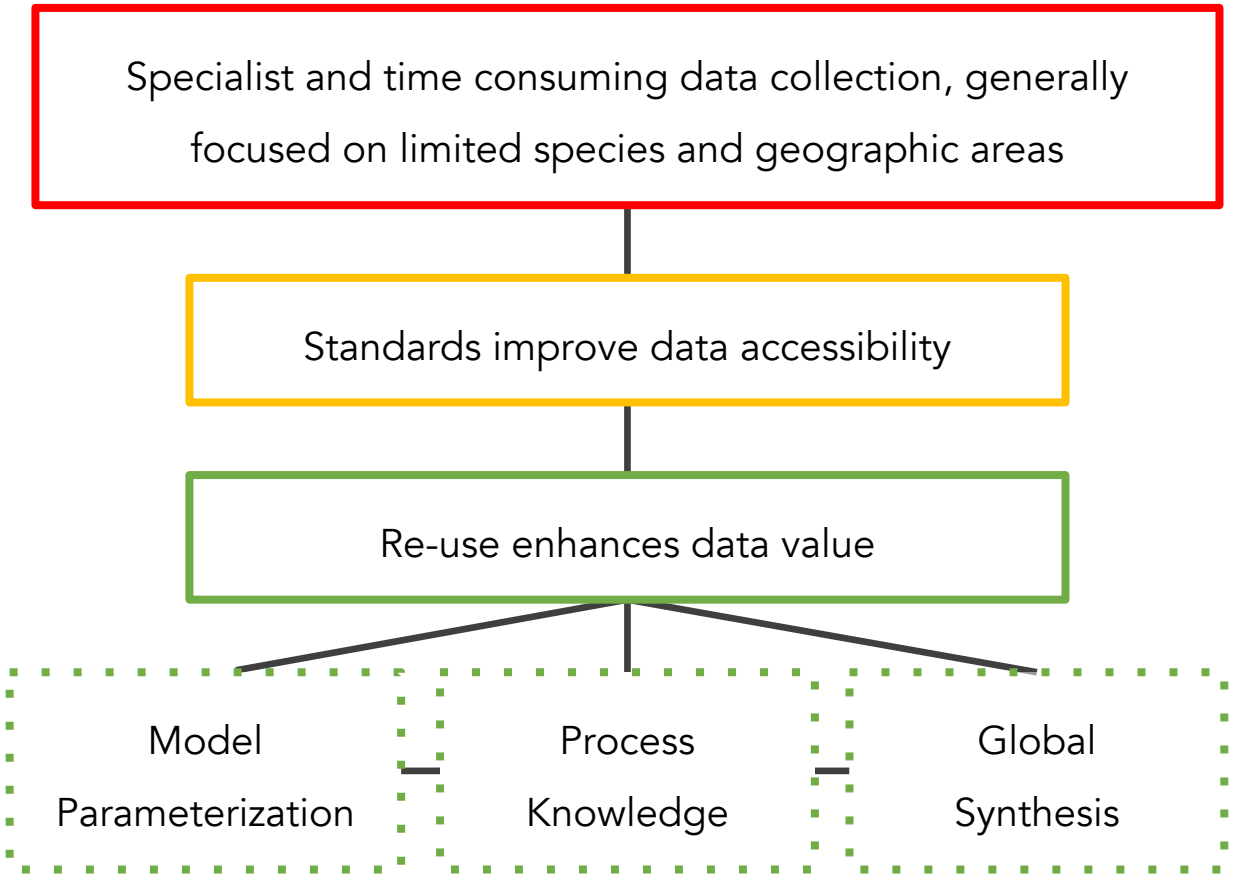
About leaf physiology data

- Provides mechanistic understanding of plant and ecosystem fluxes of carbon and water
- Provides important parameterization for terrestrial biosphere models
- Necessary to understand the response of plants to global change



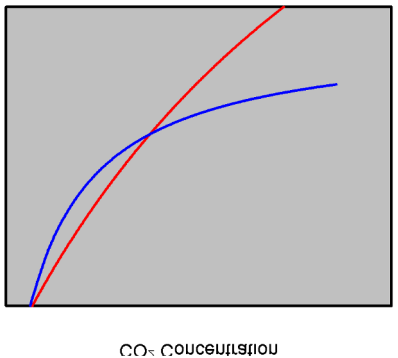


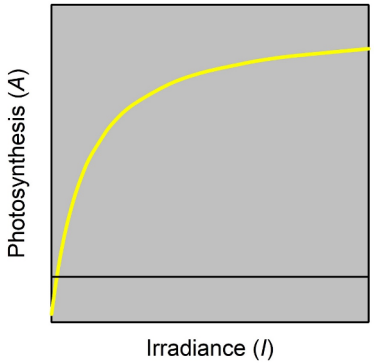
Data challenges



Data standard scope

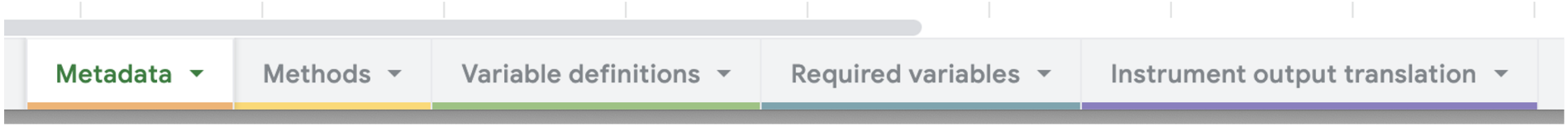
- Measurements made on leaves with portable gas exchange instruments
- Photosynthesis, respiration, and stomatal conductance
- Response to changing temperature, CO_2, and light levels
- Derived parameters used in models, e.g., $V_{c,max}$, J_{max}





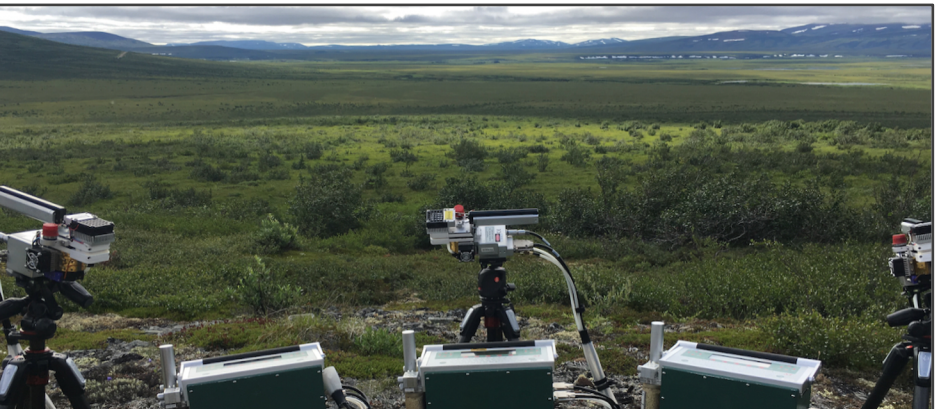
Data standard components

- General metadata, as prescribed by file-level and sample metadata standards
- Detailed metadata for methods and experimental parameters
- Standardization of required data variables: names, definitions & units
- Required variables for different measurement types
- Inclusion of raw data to future-proof data collection effort



Current status and plan

- Research of existing standards complete; draft standard prepared
- Next, review with BNL and ESS-DIVE CF
- Then, invite broad community review of v2.0
- Implement resulting standards
- Publish description of standard; manuscript in preparation
- Contributors will be invited to co-author



Feedback

Contributions and comments welcome

Alistair Rogers arogers@bnl.gov
Kim Ely
kely@bnl.gov

Rio Chagres draining into the Caribbean Sea, Panama

16S Amplicon Data and Workflows

May 11, 2020

Pamela Weisenhorn

ESS-DIVE

Deep Insight for Earth Science Data

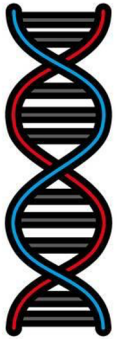
16S Amplicon data

16S data is widely used to examine microbial community composition (and sometimes to predict function)

Although valuable, there are many difficulties in comparing data across studies or repurposing data for other uses

Data standard needs to capture both the sample extraction and preparation as well as bioinformatic processing to ensure reproducibility and support data re-use

Data standard



Two common formats

.biom and .csv

Improvements to existing approaches:

- Capture of consensus sequence information
- Key metadata related to extraction and sample preparation
- Key metadata related to bioinformatic processing

Coordination with KBase

KBase is actively developing analysis and visualization tools for amplicon data

KBase importers will require amplicon data that adheres to the ESS-DIVE data standard in content and structure

KBase will allow export of data in ESS-DIVE compatible format

KBase is also developing templates to support the import of related sample metadata for analysis

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Questions?

Contact: ess-dive-support@lbl.gov

ESS-DIVE

Deep Insight for Earth Science Data

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC

NCEAS