Sharing is encouraged

A workflow to standardize collection and management of large-scale data and metadata from environmental observatories

EGU22 Meeting | May 27, 2022

Dylan O'Ryan[1,2], Charuleka Varadharajan[1], Erek Alper[3], Kristin Boye[4], Madison Burrus[1], Danielle Christianson[1], Shreyas Cholia[1], Robert Crystal-Ornelas[1], Joan Damerow[1], Wenming Dong[1], Hesham Elbashandy[1], Boris Faybishenko[1], Valerie Hendrix[1], Douglas Johnson[3], Zarine Kakalia[1,5], Roelof Versteeg[3], Kenneth Williams[1], Catherine Wong[1,5], Deborah Agarwal[1]

1) Lawrence Berkeley National Laboratory 2) California State University, Sacramento 3) Subsurface Insights 4) SLAC National Accelerator Laboratory 5) University of California, Berkeley

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

O'Ryan et al., EGU22-13338

Watershed Function SFA Project

Watershed Function Scientific Focus Area

U.S. Department of Energy research project that seeks to determine how mountainous watersheds retain and release water, carbon, nutrients, and metals

Community Field Observatory in **East River, Colorado**

Over **30 collaborating institutions**

Diverse data types collected by WFSFA

Photo credit: Connor Scalbom

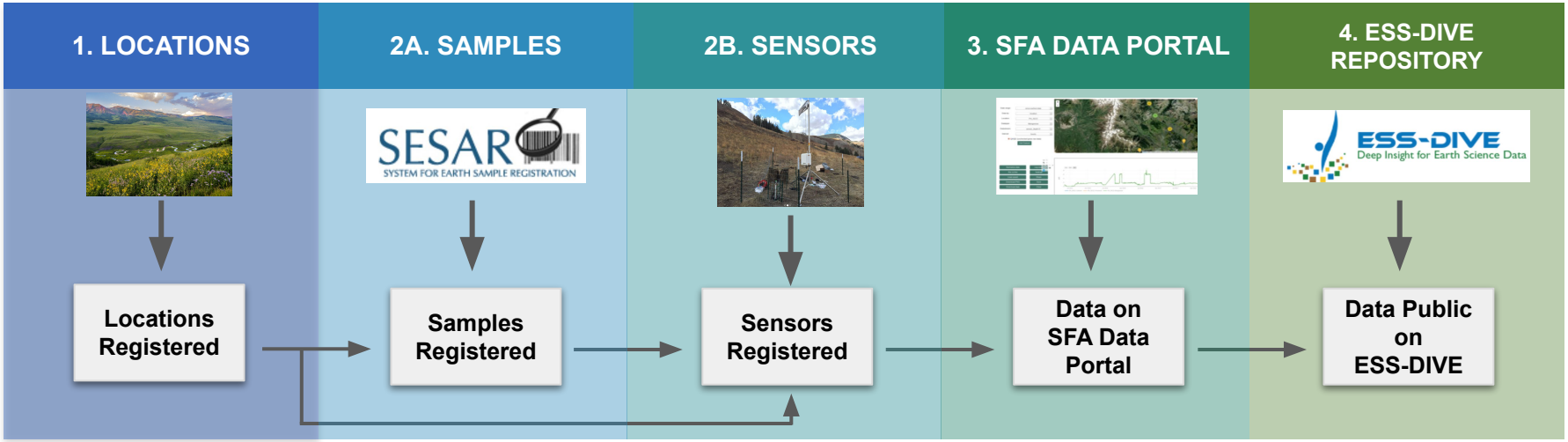
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Kakalia et al. 2021 | Hubbard et al. 2018

Data Management Workflow

Field-Data Workflow Diagram



1. Uniquely Registered Locations

- Manage sampling activity and data generation by registering *unique* locations with **Location_IDs**
- Location_IDs = **unique, persistent site identifiers** with consistent formatting (AA-XXXY)
- Registered with **detailed metadata**
- Location_IDs used to **track** sampling events and data

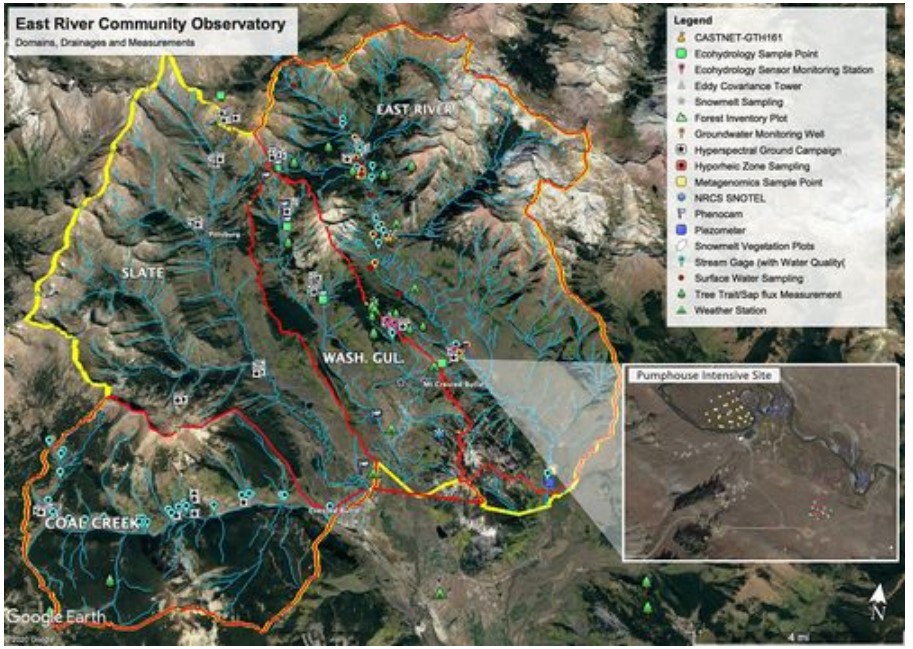


Figure from Kakalia et al. 2021

2A. Globally-Unique Sample Identifiers

- Registered **IGSNs** through **SESAR**
- **Globally unique** identifier paired with a sample
- **IGSNs** allow the team to track samples from collection to transport to analysis

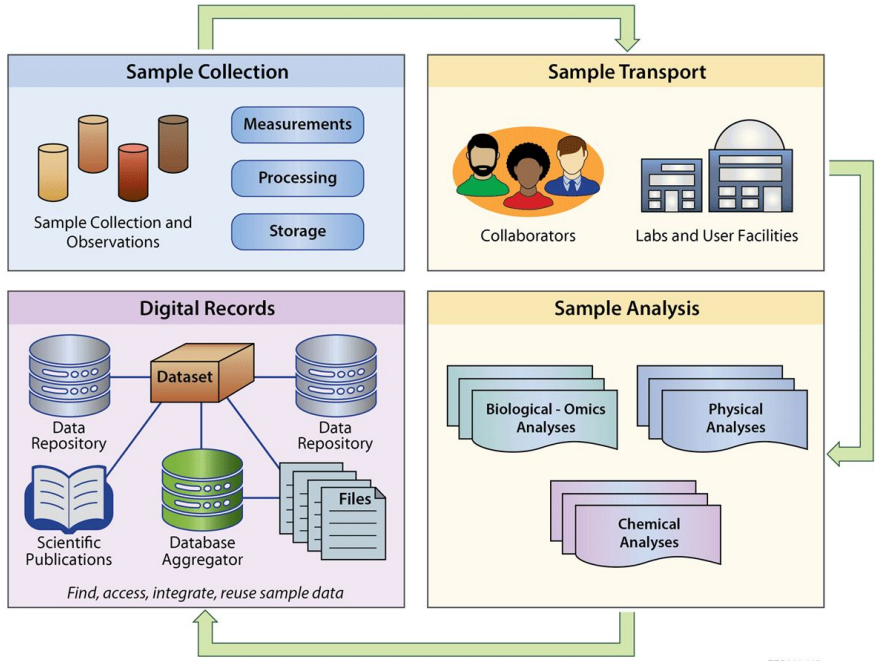
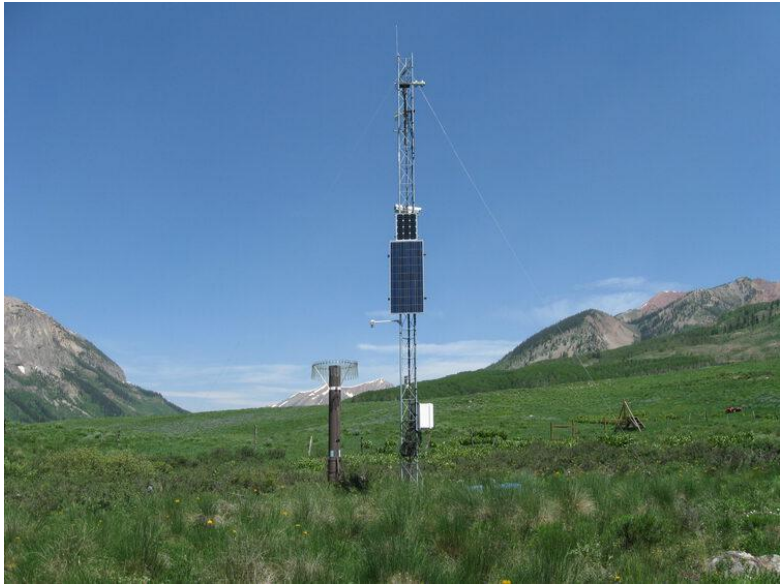


Figure from Damerow et al. 2021

2B. Registered Sensors

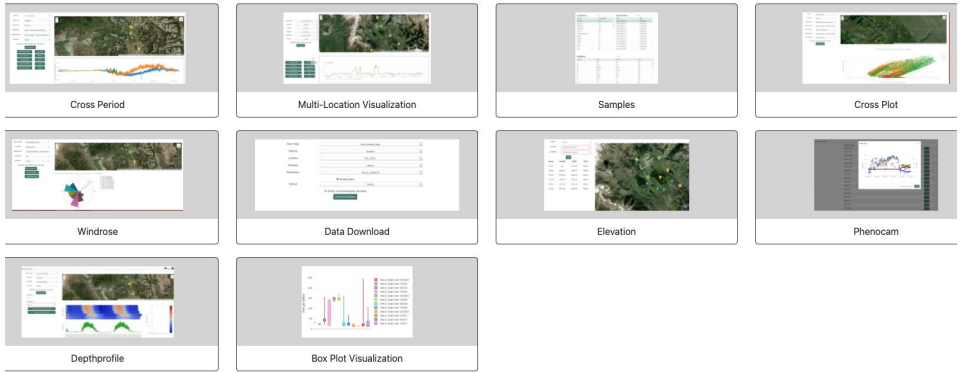
- Many sensors can be **deployed** at a single location and collect different data
- **WFSFA Sensor Listing** catalogs sensors in the East River, important metadata, and parent **Location_IDs**
- Lays the foundation for future capabilities such as advanced telemetry through **5G** and **edge computing**



https://wfsfa-data.lbl.gov/watershed/sites/ER-KLP1

3. Internal Sharing of Datasets

- **WFSFA Internal Data Portal** for field observations, data products, model output, and publication datasets
- Internal sharing enables **transparency** and **increases** use of data
- Datasets **internally reviewed** by WFSFA Data Team



4. Public Sharing of Datasets

- WFSFA Project uses **ESS-DIVE** (U.S. DOE Earth and environmental science data repository) for publication of datasets
- **Public sharing** of project datasets furthers transparency and public use of data
- ESS-DIVE **reviews datasets** for metadata and dataset requirements

Geochemical Data Standardization

Standardization vital to FAIRness of Project Data

- **Water-Soil-Sediment Chemistry community data standard** developed by U.S. Department of Energy's ESS-DIVE data repository (https://github.com/ess-dive-community/essdive-water-soil-sed-chem)

- Standardization enables use beyond initial intent
- Standardized geochemical data enables **modeling** and **analysis** of data

Crystal-Ornelas et al. In Review

Formatted Data Files increase FAIRness

Sample_Name	dateTime	DIC_mg/l
Bradley-2021-06-11	2021-06-11	16.36
Bradley-2021-06-18	2021-06-18	17.24
Bradley-2021-06-24	2021-06-24	19.11
Bradley-2021-07-02	2021-07-02	19.53
Bradley-2021-07-09	2021-07-09	19.79
Bradley-2021-07-15	2021-07-15	19.99
Copper-2021-06-11	2021-06-11	12.55
Copper-2021-06-18	2021-06-18	11.94
Copper-2021-06-24	2021-06-24	12.52
Copper-2021-07-02	2021-07-02	13.3
Copper-2021-07-09	2021-07-09	13.38
Copper-2021-07-15	2021-07-15	14.12
EAQ-2021-06-11	2021-06-11	21.06
EAQ-2021-06-18	2021-06-18	20.9
EAQ-2021-06-24	2021-06-24	21.7
EAQ-2021-07-02	2021-07-02	23.57
EAQ-2021-07-09	2021-07-09	25.11
EAQ-2021-07-15	2021-07-15	26.55
EBC-2021-06-11	2021-06-11	16.03
EBC-2021-06-18	2021-06-18	16.41
EBC-2021-06-24	2021-06-24	17.65
EBC-2021-07-02	2021-07-02	17.77
EBC-2021-07-09	2021-07-09	18.91
EBC-2021-07-15	44392	19.72
GLS-1-2021-06-11	2021-06-11	46.94

#Columns	8				
#Header_Rows	11				
Field_Name	Sample_Name	IGSN	Material	dateTime	DIC_mg.L
Unit	N/A	N/A	N/A	YYYY-MM-DD	mg/L
Unit_Basis	N/A	N/A	N/A	N/A	milligrams_per_Liter
MethodID_Storage	N/A	N/A	N/A	N/A	Storage_Sample
MethodID_Preservation	N/A	N/A	N/A	N/A	Refrig_Sample
MethodID_Preparation	N/A	N/A	N/A	N/A	Filt_Sample
MethodID_Analysis	N/A	N/A	N/A	N/A	DIC_Anl
Analysis_DetectionLimit	N/A	N/A	N/A	N/A	0.88
Analysis_Precision	N/A	N/A	N/A	N/A	0 - 50
MethodID_DataProcessing	N/A	N/A	N/A	N/A	PRCS_TOC-VCPH
Data_Status	N/A	N/A	N/A	raw	processed data
#Start_Data	Bradley-2021-06-11	IEWFS006F	Surface water	2021-06-11	16.36
N/A	Bradley-2021-06-18	IEWFS006G	Surface water	2021-06-18	17.24
N/A	Bradley-2021-06-24	IEWFS006H	Surface water	2021-06-24	19.11
N/A	Bradley-2021-07-02	IEWFS007C	Surface water	2021-07-02	19.53
N/A	Bradley-2021-07-09	IEWFS007D	Surface water	2021-07-09	19.79
N/A	Bradley-2021-07-15	IEWFS007E	Surface water	2021-07-15	19.99
N/A	Copper-2021-06-11	IEWFS006I	Surface water	2021-06-11	12.55

Detailed Methods File increase understandability

Method_ID	Method_Type	Method_Description
Filt_Sample	Field Filtering of Sample	Water samples for DOC, DIC, and TDN analysis were filtered using 0.45 micrometer Millipore filters
Storage_Sample	Storage of Filtered Samples	Filtered samples were collected in no-headspace 40 mL glass vials with polypropylene open-top caps and butyl rubber septa (VWR TraceClean)
Refrig_Sample	Refrig Storgage Samples [Preservation] (QA)	Samples were transported to laboratory on ice and stored in 4 degree C refrigerator until analysis.

File-level Metadata and Data Dictionary Files

File_Name	File_Description	Standard	Location_ID
locations.csv	Location information for the sites containing data. This file contains both Location_IDs and the latitude and longitude measurements associated with the Location_IDs.	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	N/A
cc_irw1_deltad.csv	Data for liquid water delta2H	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	CC-IRW1
cc_irw1_deltao18.csv	Data for liquid water delta18O	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	CC-IRW1
cc_lmr1_deltad.csv	Data for liquid water delta2H	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	CC-LMR1
cc_lmr1_deltao18.csv	Data for liquid water delta18O	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	CC-LMR1
coal_11_deltad.csv	Data for liquid water delta2H	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	Coal-11
coal_11_deltao18.csv	Data for liquid water delta18O	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	Coal-11
er_avy1_deltad.csv	Data for liquid water delta2H	ESS-DIVE Reporting Format for Comma-separated Values (CSV) File Structure v1.0.0	ER-AVY1

Column_or_Row_Name	Unit	Definition	Column_or_Row_Long_Name	Data_Type
utc_time	YYYY-MM-DD	Date samples were collected in the field	N/A	date
location_name	N/A	Location_ID for the sites at the WFSFA field site	N/A	text
latitude	decimal degree	Latitude of the locaton where data was collected	N/A	numeric
longitude	degree	Longitude of the location where data was collected	N/A	numeric
dic_mg/l	milligrams per liter (mg/l)	mean values for dissolved organic carbon (DOC) in milligrams per liter (mg/l)	N/A	numeric
npoc_mg/l	milligrams per liter (mg/l)	mean values for dissolved organic carbon (DOC) analyzed as non-purgeable organic carbon (NPOC) in milligrams per liter (mg/l)	N/A	numeric

Lessons Learned from Data Standardization

- **Conversion** and **use** becomes easier
- In our experience, **~ 1 hour** to convert an existing dataset

- **Reuse** templates when data and methods are unchanged, then **populate** with variable specific information

Image from WSFA Field Information Portal: https://wfsfa-data.lbl.gov/watershed/sites/ER-BRD1

Thanks for Listening

Contact Information:
dporyan [at] lbl.gov
dylanporyan [at] gmail.com

Acknowledgement: This work was supported by the US Department of Energy, Office of Biological and Environmental Research as part of the LBNL Watershed Function Scientific Focus Area

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

WFSFA Project Resources

Project Website: https://watershed.lbl.gov/

Watershed Function SFA Field Information Portal: https://wfsfa-data.lbl.gov/watershed/

East River Data Portal: https://data.ess-dive.lbl.gov/portals/east-river-watershed

Kakalia et al. on WFSFA Datasets: https://onlinelibrary.wiley.com/doi/10.1002/hyp.14243

Hubbard et al. on WFSFA Project: https://acsess.onlinelibrary.wiley.com/doi/full/10.2136/vzj2018.03.0061

Story Map on WFSFA Project: https://eesa.lbl.gov/a-watershed-moment/

WFSFA Project Resources

WFSFA Field Data Workflow: on https://watershed.lbl.gov/