Applying Community Data Reporting Formats to Open-Source Water Quality Data

Dylan O'Ryan[1], Robert Crystal-Ornelas[1], Deb Agarwal[1], Kristin Boye[2], Shreyas Cholia[1], Joan Damerow[1], Wenming Dong[3], Kenneth Williams[4], and Charuleka Varadharajan[1]

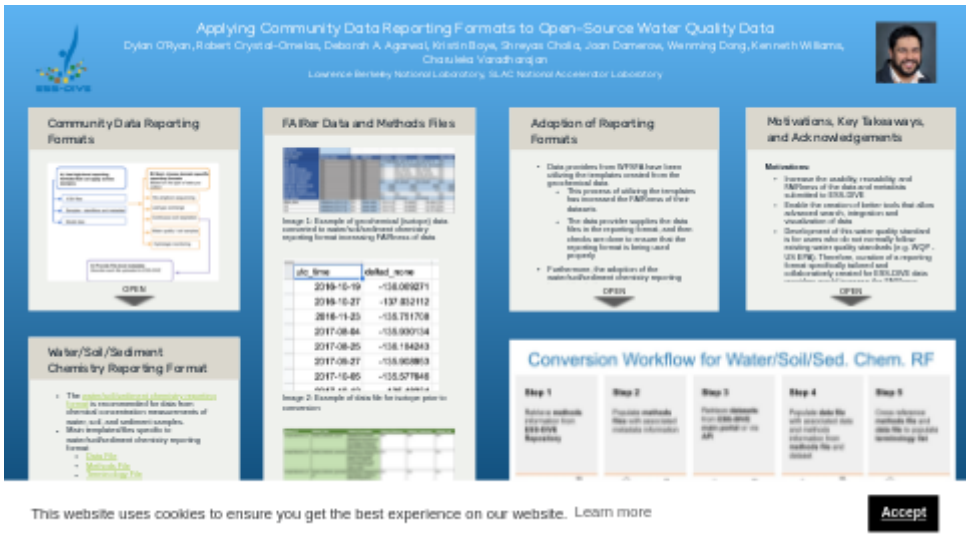
[1]Lawrence Berkeley National Laboratory
[2]SLAC National Acceleratory Laboratory
[3]Earth and Environment Sciences Area, Lawrence Berkeley National Laboratory
[4]Earth and Environment Sciences Area

November 24, 2022

Abstract

Data standardization can enable data reuse by streamlining the way data are collected, providing descriptive metadata, and enabling machine readability. Standardized open-source data can be more readily reused in interdisciplinary research that requires large amounts of data, such as climate modeling. Despite the importance given to both FAIR (Findable, Accessible, Interoperable, Reusable) data practices and the need for open-source data, a remaining question is how community data standards and open-source data can be adopted by research data providers and ultimately achieve FAIR data practices. In an attempt to answer this question, we used newly created water quality community data reporting formats and applied them to open-source water quality data. The development of this water quality data format was curated with several other related formats (e.g., CSV, Sample metadata reporting formats), aimed at targeting the research community that have historically published water quality data in a variety of formats. The water quality community data format aims to standardize how these types of data are stored in the data repository, ESS-DIVE (Environmental Systems Science Data Infrastructure for a Virtual Ecosystem). Adoption of these formats will also follow FAIR practices, increase machine readability, and increase the reuse of this data. We applied this community format to open-source water quality data produced by the Watershed Function Scientific Focus Area (WFSFA), a large watershed study in the East River Colorado, which involves many national laboratories, institutions, scientists, and disciplines. In this presentation, we provide a demonstration of a relatively efficient process for converting open-source water quality data into a format that adheres to a community data standard. We created examples of water quality data translated to the reporting formats that demonstrated the functionality of these data standards; descriptive metadata and sample names, streamlined data entries, and increased machine readability were products of this translation. As the community data standards are integrated within the WFSFA data collection processes, and ultimately all data providers of ESS-DIVE, these steps may enable interdisciplinary data discovery, increase reuse, and follow FAIR data practices.

Applying Community Data Reporting Formats to Open-Source Water Quality Data



Dylan O'Ryan, Robert Crystal-Ornelas, Deborah A. Agarwal, Kristin Boye, Shreyas Cholia, Joan Damerow, Wenming Dong, Kenneth Williams, Charuleka Varadharajan

Lawrence Berkeley National Laboratory, SLAC National Accelerator Laboratory









PRESENTED AT:

COMMUNITY DATA REPORTING FORMATS

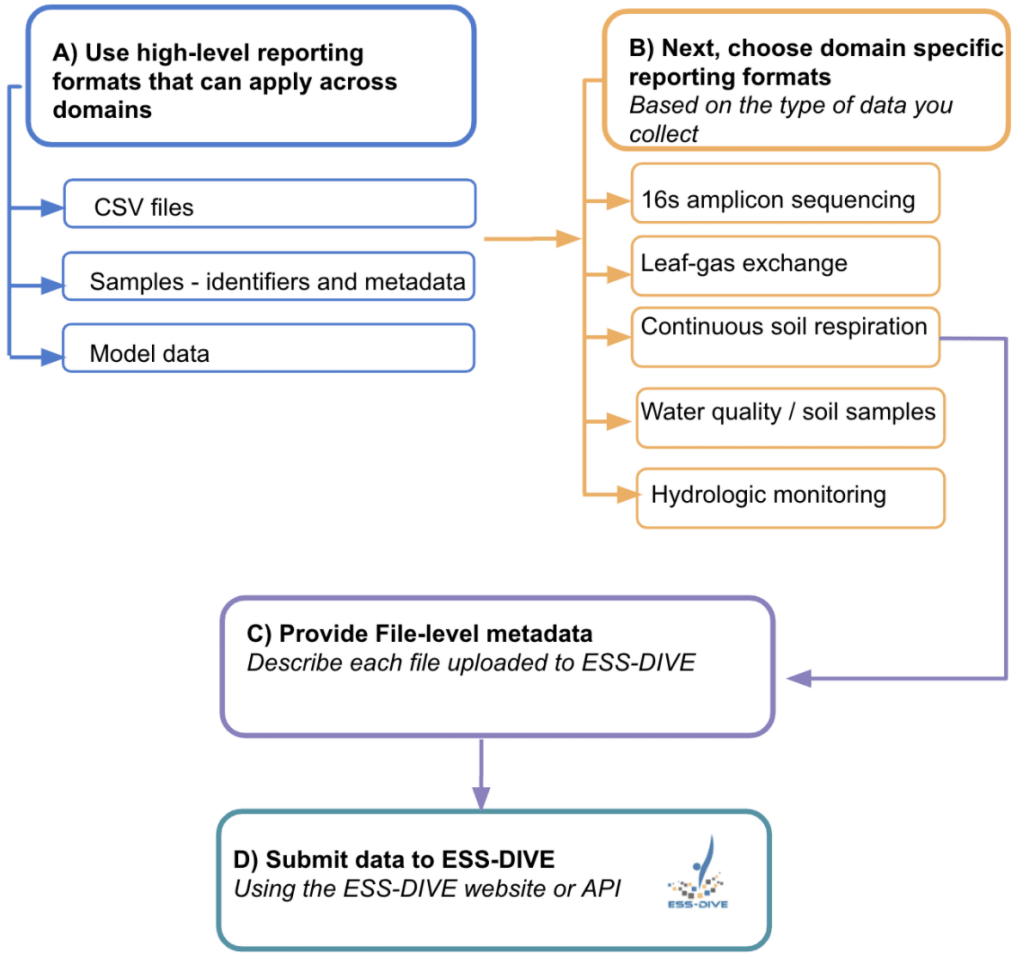


Diagram from ESS-DIVE: Workflow diagram showing the use of the reporting formats

- Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) (https://ess-dive.lbl.gov/), Lawrence Berkeley National Laboratory's Environmental Systems Science (ESS) data repository, collaborated with six teams of scientists in the US Department of Energy (DOE) ESS community to develop community data standards (http://github.com/ess-dive-community) (hereafter, reporting formats).

- These reporting formats were developed to standardize a wide range of data, increase FAIRness (Findable, Accessible, Interoperable, Reusable), and enable the creation of better tools that allow advanced search, integration and visualization of data.

- General Reporting Formats

  - Comma Seperated Values (CSV) (https://github.com/ess-dive-community/essdive-csv-structure)

  - File-level Metadata (FLMD) (http://github.com/ess-dive-community/essdive-file-level-metadata)

  - Sample ID Metadata (https://github.com/ess-dive-community/essdive-sample-id-metadata)

    - Package-level Metadata (https://github.com/ess-dive-community/essdive-package-metadata)

- Domain-Specific Reporting Formats
    - Leaf-gas exchange (https://github.com/ess-dive-community/essdive-leaf-gas-exchange)
    - Continuous soil respiration (https://github.com/ess-dive-community/essdive-soil-respiration)
    - Hydrological monitoring (https://github.com/ess-dive-community/essdive-hydrologic-monitoring)
    - Water/Soil/Sediment Chemistry (http://github.com/ess-dive-community/essdive-water-soil-sed-chem)
    - ...

- Workflow diagrams (see above) and help documentation (http://docs.ess-dive.lbl.gov/contributing-data/data-reporting-formats) has been developed to ensure that data providers have support while using these reporting formats

WATER/SOIL/SEDIMENT CHEMISTRY REPORTING FORMAT

- The water/soil/sediment chemistry reporting format (http://github.com/ess-dive-community/essdive-water-soil-sed-chem) is recommended for data from chemical concentration measurements of water, soil, and sediment samples.
- Main templates/files specific to water/soil/sediment chemistry reporting format
  - Data File (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/Detailed_Instructions_Data_File.md)
  - Methods File (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/Detailed_Instructions_Methods_File.md)
  - Terminology File (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/Detailed_Instructions_Terminology_File.md)
- The water/soil/sediment chemistry reporting format also uses formats and recommendations from the File-level Metadata (FLMD) (http://github.com/ess-dive-community/essdive-file-level-metadata) and CSV (http://github.com/ess-dive-community/essdive-csv-structure) reporting formats
  - Required FLMD file (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/templates/flmd_template.csv) and data dictionary (dd) file (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/templates/dd_template.csv)

FAIRER DATA AND METHODS FILES

#Columns	7						
#Header_Rows	11						
Field_Name	Sample_Name	IGSN	Material	dateTime	deltao18	deltad	data_notes
Unit	N/A	N/A	N/A	YYYY-MM-D	ppt_(per-mil, %)	ppt_(per-mil, %)	N/A
Unit_Basis	N/A	N/A	N/A	N/A	parts_per_thousar	parts_per_thousand	N/A
MethodID_Storage	N/A	N/A	N/A	N/A	Stor_sample	Stor_sample	N/A
MethodID_Preservation	N/A	N/A	N/A	N/A	Refrig_sample	Refrig_sample	N/A
MethodID_Preparation	N/A	N/A	N/A	N/A	samplePreparation	samplePreparation_	N/A
MethodID_Analysis	N/A	N/A	N/A	N/A	Anl_sample	Anl_sample	N/A
Analysis_DetectionLimit	N/A	N/A	N/A	-9999	-9999	-9999	-9999
Analysis_Precision	N/A	N/A	N/A	-9999	0.025	0.1	-9999
MethodID_DataProcessing	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Data_Status	N/A	N/A	N/A	raw	processed	processed	N/A
#Start_Data	tuttleSpring_2016-10-19	N/A	stream water	2016-10-19	-18.369955	-136.069271	N/A
N/A	tuttleSpring_2016-10-27	N/A	stream water	2016-10-27	-18.580289	-137.032112	N/A
N/A	tuttleSpring_2016-11-23	N/A	stream water	2016-11-23	-18.260694	-135.751708	N/A

Image 1: Example of geochemical (isotope) data converted to water/soil/sediment chemistry reporting format increasing FAIRness of data

utc_time	deltad_none
2016-10-19	-136.069271
2016-10-27	-137.032112
2016-11-23	-135.751708
2017-08-04	-135.930134
2017-08-25	-136.184243
2017-09-27	-135.908953
2017-10-05	-135.577646

Image 2: Example of data file for isotope prior to conversion

Method_ID	Method_Type	Method_Description	Method_Reference	Method_ Instrument	Method_Lab
sampleCollection_01	Sample_Collection_source	Samples are collected directly from the source for the tributary streams and groundwater monitoring wells (following pumping)	N/A	N/A	N/A
sampleCollection_02	Sample_Collection_automatic	Samples are collected from an automated sampler bottle in the case the Pumphouse-ISCO and Coal 11-ISCO surface water location.	N/A	N/A	N/A
sampleCollection_03	Sample_Collection_groundwater	Samples are collected from groundwater monitoring wells through above ground or below ground pumps used to recover fluids from the well	N/A	N/A	N/A
sampleCollection_04	Sample_Collection_isotope	Collection of isotope samples	N/A	N/A	N/A

Image 3: Methods file with filled in metadata from the isotope dataset

Data Files (http://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/Detailed_Instructions_Data_File.md)

- Geochemical data from the Watershed Function Scientific Focus Area (WFSFA) project (https://watershed.lbl.gov/) was used to test the water/soil/sediment chemistry reporting format

    - ICP-MS, Ammonia-N, NPOC/DIC, TDN, anion, and isotope data types were converted

- Before the conversion process, the data files lacked desired information for the reusability and understandability of the dataset (Image 2)

- Fields in the data file increases the FAIRness of the data
    - Descriptive sample names, Analysis_Precision, IGSNs, and Data_status.

- Suggestions from the CSV reporting format increases the machine-readability of the data (empty fields noted with N/A or -9999)

Methods File (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/Detailed_Instructions_Methods_File.md)

- Methods file is where information on the samples' collection methods, analysis, storage, and other vital information is stored

- Methods file fields
    - Method_ID
        - User determined ID for steps in collecting, processing, analyzing, or storing the sample
    - General Methods information
        - Method_type (e.g., analysis, processing, storage, preservation) and method_description (method step/description)
    - Specific Methods information
        - Method_temp, Method_light, Method_instrument, Method_lab...

Terminology File (https://github.com/ess-dive-community/essdive-water-soil-sed-chem/blob/main/Detailed_Instructions_Terminology_File.md)

- The terminology file is structured to allow for one project/team to have one master terminology library to standardize terms used.

	A	B	C	D	E	F
1	**Term**	**Unit**	**Definition**	**Column_or_Row_Long_Name**	**Data_Type**	**Term_Type**
2	Column_Name	N/A	header for row of column names	N/A	text	Row_header
3	Sample_Name	N/A	header for column with sample names	N/A	text	Column_header
4	imidacloprid	micrograms_per_liter	concentration of imidacloprid	N/A	numerical	Column_header
5	units	N/A	header for row with units	N/A	text	Row_header
6	MethodID_Analysis	N/A	code for the analysis method used	N/A	text	Row_header
7	imidacloprid_Notes	N/A	notes specific to imidacloprid values	N/A	text	Column_header
8	BD	N/A	below detection limit	N/A	text	dataFlag
9	BV_SIQ	N/A	broken vial, sample integrity questionable, data not reliable	N/A	text	dataFlag

Image 4: From water/soil/sediment chemistry reporting format, showing the use of the terminology file for a master list for a project

ADOPTION OF REPORTING FORMATS

- Data providers from WFSFA have been utilizing the templates created from the geochemical data
  - This process of utilizing the templates has increased the FAIRness of their datasets
  - The data provider supplies the data files in the reporting format, and then checks are done to ensure that the reporting format is being used properly
- Furthermore, the adoption of the water/soil/sediment chemistry reporting format (and other reporting formats) by other data providers within the WFSFA and other ESS projects will increase the datasets' usability

MOTIVATIONS, KEY TAKEAWAYS, AND ACKNOWLEDGEMENTS

Motivations:

- Increase the usability, reusability, and FAIRness of the data and metadata submitted to ESS-DIVE
- Enable the creation of better tools that allow advanced search, integration and visualization of data
- Development of this water quality standard is for users who do not normally follow existing water quality standards (e.g. WQP - US EPA). Therefore, curation of a reporting format specifically tailored and collaboratively created for ESS-DIVE data providers would increase the FAIRness, reusability, and usability of their datasets

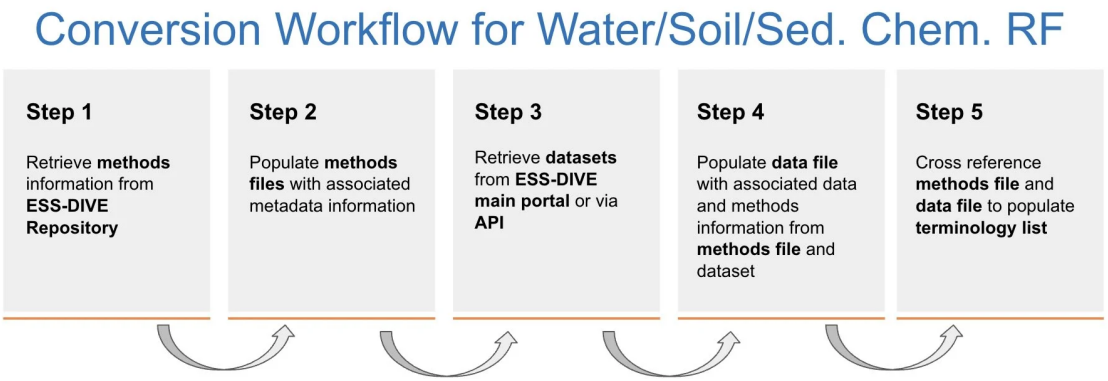
Key Takeaways:

- Use of the water/soil/sediment chemistry reporting format is easy and straightforward
- ESS-DIVE's reporting formats enable FAIRness of data, and may enable for interdisciplinary data discovery and increase reuse
- As use of reporting formats increase for datasets submitted to ESS-DIVE, this will enable ESS-DIVE to create future tools increasing data discovery for data users
- Reusing the data file template with associated header information for the same data type is recommended, as long as the associated metadata does not change. Therefore, reusing this information quickens the process of conversion.

Funding Acknowledgements:

ESS-DIVE is funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Earth and Environmental Sciences Division, Data Management program under contract number DE-AC02-05CH11231. ESS-DIVE uses resources of the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internship (CCI) program



AUTHOR INFORMATION

The presenting author, Dylan O'Ryan, is a student at CSU Sacramento studying Envrionmental Studies B.S. Dylan O'Ryan is also a Student Research Assistant at Lawerence Berkeley National Laboratory (LBNL). At LBNL, he works with the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) as part of the Community Engangement team focusing on reporting format roll-out, IGSN registration, and data publication review. He also works with the Watershed Function Scientific Focus Area (WFSFA) data team focusing on location and sensor metadata, integration of reporting formats within the project, and data publication reviews.

Dylan's interests are focused around water quality data collection, community monitoring programs (CBM), data standards, and FAIR data practices.

ABSTRACT

Data standardization can enable data reuse by streamlining the way data are collected, providing descriptive metadata, and enabling machine readability. Standardized open-source data can be more readily reused in interdisciplinary research that requires large amounts of data, such as climate modeling. Despite the importance given to both FAIR (Findable, Accessible, Interoperable, Reusable) data practices and the need for open-source data, a remaining question is how community data standards and open-source data can be adopted by research data providers and ultimately achieve FAIR data practices.

In an attempt to answer this question, we used newly created water quality community data reporting formats and applied them to open-source water quality data. The development of this water quality data format was curated with several other related formats (e.g., CSV, Sample metadata reporting formats), aimed at targeting the research community that have historically published water quality data in a variety of formats. The water quality community data format aims to standardize how these types of data are stored in the data repository, ESS-DIVE (Environmental Systems Science Data Infrastructure for a Virtual Ecosystem). Adoption of these formats will also follow FAIR practices, increase machine readability, and increase the reuse of this data. We applied this community format to open-source water quality data produced by the Watershed Function Scientific Focus Area (WFSFA), a large watershed study in the East River Colorado, which involves many national laboratories, institutions, scientists, and disciplines.

In this presentation, we provide a demonstration of a relatively efficient process for converting open-source water quality data into a format that adheres to a community data standard. We created examples of water quality data translated to the reporting formats that demonstrated the functionality of these data standards; descriptive metadata and sample names, streamlined data entries, and increased machine readability were products of this translation. As the community data standards are integrated within the WFSFA data collection processes, and ultimately all data providers of ESS-DIVE, these steps may enable interdisciplinary data discovery, increase reuse, and follow FAIR data practices.