Improving the management and reuse of water quality data for the DOE's Watershed Function Scientific Focus Area using community data standards

Dylan O'Ryan[1], Robert Crystal-Ornelas[2], Charuleka Varadharajan[2]
[1]San Joaquin Delta College,[2]Lawrence Berkeley National Laboratory

Abstract

FAIR and standardized data are increasingly needed and prioritized in earth and environmental science research. In an effort toward more standardized data, Lawrence Berkeley National Lab (LBNL) has begun developing community data standards, which provide guidelines for how data providers collect and store their data. During this internship, I gained experience in data management and assisted with several essential data standardization tasks. From these Community Partner-developed water quality reporting formats (community data standards), I created examples of water quality datasets in these reporting formats to demonstrate how data providers can follow these formats when submitting data to the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem data repository (ESS-DIVE). Progress that I made during the internship term helps to ensure water quality data submitted to ESS-DIVE follows Findable Accessible Interoperable and Reusable (FAIR) data practices. Future steps of community data standards will have to be integrated into current steps for data collection and ensure that necessary metadata is supplied. Additionally, LBNL has made steps to collect metadata information for sensors as part of the Watershed Function Scientific Focus Area (WFSFA). I worked with sensor metadata templates, to facilitate collecting information about sensors that are deployed as part of the WFSFA. Ultimately, data standards and metadata information will ensure that LBNL, ESS-DIVE, and the Department of Energy (DOE) follow FAIR standards.

Keywords: Data standardization, Community Data Standards, Metadata, FAIR data, Data Repositories, Open-Source Data

Introduction

Earth and environmental scientists are often required to deposit their data into repositories by funders or by journals where they submit their manuscripts. Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) is a Department of Energy (DOE) data repository that stores DOE-sponsored data from fieldwork, climate modeling projects, and other datasets. Data reusability is crucial because it can enable scientists and researchers to make predictions, analyze environmental changes and other environmental projects. For example, the DOE conducts climate modeling, which requires a lot of data to make predictions. In turn, these modeling predictions continue to improve with more data and time. Data stored needs to be machine-readable, easily traceable, and usable; therefore, changes and advancements in how data can be stored are being researched and developed. For example, one of the leading data principles that originated to combat these challenges is FAIR (Findable, Accessible, Interoperable, Reusable).

FAIR data principles originated out of the need to support the infrastructure for the reuse of data [1]. Reusability, which is one of FAIR's main focuses, can be achieved through data standards. Data standards are formal and accredited guidelines for structuring data that large oversight committees often govern (e.g., ISO, OCG, CF Conventions). Whereas community data standards (also called reporting formats) are what ESS-DIVE strives for, these formats are not "standards" in the formal sense but still involve communities of researchers to develop documents and templates that can ensure data is formatted consistently [2].

The Watershed Function Scientific Focus Area (hereafter, WFSFA) is a project based in the East River Watershed, located near Crested Butte, CO. This project aims to understand how environmental perturbations impact watersheds, discharge of water, and constituents of the watershed, which can potentially be applied to other watersheds [3]. Scientists working on the WFSFA are striving to make predictions of how mountainous watersheds react to environmental changes, and they are developing tools to predict droughts, early snowmelt, and other perturbations that affect the overall watershed function and health [3]. If the DOE sponsors a project or a scientist's research, scientists must store their data in ESS-DIVE. WFSFA is a DOE-sponsored project; therefore, this project is required to store data collected within ESS-DIVE's data repository.

ESS-DIVE's role is to archive environmental and earth sciences research data, preserving and improving the data in the process [4]. In order to make data stored in ESS-DIVE reusable, ESS-DIVE has worked with teams of Community Partners to develop a set of data formatting guidelines (reporting formats) for some of the diverse data stored in the repository; however, following the formats is not requirement to store data with ESS-DIVE. In order to make data stored in ESS-DIVE follow FAIR more closely, in 2019, ESS-DIVE partnered with six teams of Community Partners to develop data reporting formats for some of the diverse data stored in the repository. Some reporting formats are specific to research domains (e.g., water and soil samples), while others are generalized to a wide range of data (e.g., CSV files, sample collections). ESS-DIVE's goal of broad adoption of reporting formats within the scientific community will enable reusable and reliable data in the future.

The WFSFA data management framework (DMF) team seeks to develop workflows and tools to enable publishing WFSFA data using FAIR principles on ESS-DIVE. As part of a new data management workflow, the DMF team compiled a comprehensive list of locations with detailed metadata [5] [6]. As part of this continued work of publishing WFSFA data using FAIR principles, the DMF team is undertaking work of compiling sensor metadata associated with the list of locations within the WFSFA project.

As part of the missions of Earth and Environmental Sciences Area (EESA) and other associated projects, there is a two-fold set of objectives for the internship. These objectives are based on implementing or bettering the set of systems that are existing within the projects. The first objective is to work with a water quality reporting format developed by an ESS-DIVE Community Partner. This reporting format is designed to standardize how water quality data are entered into data files and ensure vital information is being supplied -- metadata. For this objective, the task was to integrate current water quality data collected as part of the WFSFA (e.g., ICP-MS, NPOC/DIC/TDN, Anion, Ammonia-N, Isotope) into the reporting format templates. In the process, suggestions will be made about the format's usability to the Community Partner, and expose WFSFA data providers to these reporting formats. Ultimately, this work strives to help data providers utilize these templates to supply their data in the reporting templates rather than their current forms. The relevance of creating reporting formats is to ensure that data files are following FAIR principles, enabling future reuse of these data. The second objective of this internship addresses sensor metadata. The WFSFA has sensors deployed in the East River, which are collecting diverse data from the watershed. As noted previously, the DMF team collected the location information of these deployed sensors; therefore, the second objective is to further the information that the WFSFA has on these sensors. Feedback will be provided on a newly developed sensor metadata template that will store sensor information from the WFSFA project. In order to populate the metadata, collaboration with multiple stakeholders within this project to ensure that the information collected will be accurate will have to be undertaken.

Progress

The technical approach to conducting these projects primarily involves open-source, publicly accessible programs and data. The steps of accessing and collaborating on documents was done utilizing Google Drive to access reporting formats drafts, create filled metadata and water quality reporting formats, as well as collaborate on shared documents. Datasets were assessed utilizing a Subsurface Insights (SSI) database which has an Application Program Interface (API) that can access water quality datasets from WFSFA. In addition to SSI, data packages were assessed utilizing ESS-DIVE's data portal to access and collect information about methods information (i.e., metadata) as well as datafiles. GitHub, a Version Control System, was utilized to view the collaborative development of community data standards and access publicly available documents and templates developed by the ESS-DIVE team (e.g., CSV, Water/Soil/Sed. Chem, Sample ID

Metadata) [7]. GitHub mainly was used for providing feedback to ESS-DIVE Community Partners, utilizing the "issue" feature on the platform.

The main internship results include five water quality datasets (Isotope, Ammonia-N, NPOC/DIC/TDN, Anion, and ICP-MS) that were converted to the water quality reporting format for WFSFA data providers. These files for data providers include: a dataFile, methodFile, and a term list [reference Figure 1 for workflow diagram]. These templates, or examples, were supplied to the data providers to show that these reporting formats that are developed can apply to the data they are providing to ESS-DIVE and relaying feedback from data providers to the Community Partner developing this reporting format. Several meetings with senior scientists at LBNL, the data providers, occurred to receive their feedback on the converted data. These reporting formats will also be used as a jumping-off point for the WFSFA data providers to utilize methods information on their future datasets. Additional internship results include the feedback given to the Community Partner developing this reporting format regarding clarifications that should be made before the finalization of the reporting formats. There have been steps, in regards to goal two, made in collecting metadata information for the sensors as part of the WFSFA. For instance, a meeting occurred to facilitate work in collecting metadata information for sensors deployed as part of the WFSFA.

Future Work

ESS-DIVE and EESA team members will continue to adapt and apply the water quality reporting formats. The data templates that I used and provided several rounds of feedback on will be used by other data providers that are part of the WFSFA. The ultimate long-term goal is to have most water quality data being supplied to ESS-DIVE in this reporting format. This task will require collaboration between multiple stakeholders within these projects, such as data providers from WFSFA and the ESS-DIVE team. In addition to further work involving the reporting formats, there will also be continued work to collect sensor metadata. Sensor metadata will be collected for newly deployed sensors as part of the WFSFA. This task is critical to ensure that the vital information is collected for new sensors (i.e., latitude/longitude, depth, variables collected). However, accomplishing this task will be challenging when looking at it from the data curation perspective for a myriad of reasons: researchers from various disciplines and institutions, differing equipment, sampling preferences, and collecting this information can be cost and time-intensive [8]. WFSFA team members will undertake this task to ensure that this information is continuously updated.

I will continue future work following the internship term, which includes developing a community blog post for ESS-DIVE's website to explain the process of converting existing datasets to the water quality reporting format. In addition, I will present at an ESS-DIVE Webinar addressing the conversion of datafiles to reporting formats. By continuing this work, I will keep working on my mission of outreach and integration into other mainstream data providers and the WFSFA project.

Impact on Laboratory

My accomplishments through the CCI internship and the future work I outlined in this paper are part of the missions of DOE and LBNL. The DOE's missions and objectives are to further scientific and technological innovation; therefore, creating reusable data is an essential component of these objectives [9]. Reusable data can contribute to climate, environmental, and energy technologies (e.g., climate predictions and modeling). The impacts from the objectives of the internship have allowed for new capabilities of the ESS-DIVE team. As a result of the work during the internship, several impacts were made: the water quality reporting format is easier to use, the WFSFA team is closer to adopting the reporting formats across multiple research projects happening along the East River Watershed, as well as data being structured to be reporting more consistently from WFSFA projects. Open-source data availability is a crucial step moving forward; therefore, having data files that are both easily understood by researchers collecting data and machine-readable by database software will need to be prioritized. Data stored on ESS-DIVE following FAIR principles will contribute to the lab and DOE's goal of expanding the availability of publicly accessible data.

Conclusion

In summary, the objectives I completed during my CCI internship fulfilled goals for EESA, WFSFA, and ESS-DIVE. As for the first goal of the internship, I created reporting format templates to receive feedback on the usability of these documents and in turn, to help integrate these formats into their future data curation. These reporting formats are designed to streamline the functionality of stored data, therefore, adoption of these formats will be extremely beneficial in the future for data analysis. As for the sensor metadata, I have made steps to collect this information from SSI. Adjustment to the sensor metadata collection may be needed if the information requested cannot be obtained from the API. During the internship term, the progress that I made strides to ensure water quality data is reported in a streamlined and consistent way while also collecting vital metadata information about sensors deployed as part of the WFSFA. Further advancements in data standards are required to ensure that all data supplied to ESS-DIVE will abide by FAIR data practices, in turn ensuring future use of data and curation of machine-readable data.

References

[1] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1). https://doi.org/10.1038/sdata.2016.18

[2] Sansone, SA., McQuilton, P., Rocca-Serra, P. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol* 37, 358–367 (2019). https://doi.org/10.1038/s41587-019-0080-8

[3] Hubbard, S. S., Williams, K. H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., Brodie, E., Carroll, R., Dafflon, B., Dwivedi, D., Falco, N., Faybishenko, B., Maxwell, R., Nico, P., Steefel, C., Steltzer, H., Tokunaga, T., Tran, P. A., Wainwright, H., & Varadharajan, C. (2018). The East River, Colorado, Watershed: A Mountainous Community Testbed for Improving Predictive Understanding of Multiscale Hydrological–Biogeochemical Dynamics. *Vadose Zone Journal*, *17*(1), 1–25. https://doi.org/10.2136/vzj2018.03.0061

[4] *ESS-DIVE - Environmental Systems Science Data Infrastructure for a Virtual Ecosystem*. ESS-DIVE. (n.d.). https://ess-dive.lbl.gov/.

[5] Kakalia, Z., Varadharajan, C., Alper, E., Brodie, E., Burrus, M., Carroll, R., Christianson, D., Hendrix, V., Henderson, M., Hubbard, S., Johnson, D., Versteeg, R., Williams, K., & Agarwal, D. (2021). The East River Community Observatory Data Collection: Diverse, multiscale data from a mountainous watershed in the East River, Colorado. *Hydrological Processes*, *35*(6). https://doi.org/10.22541/au.160157556.64095872

[6] Varadharajan C ; Kakalia Z ; Banfield J ; Berkelhammer M ; Brodie E ; Christianson D ; Dafflon B ; Carbone M S ; Carroll R ; Chadwick K D ; Christensen J ; Enquist B J ; Fox P ; Henderson M ; Gochis D ; Kueppers L ; Powell T ; Matheus Carnevali P ; Singha K ; Sorensen P ; Tokunaga T ; Versteeg R ; Wilkins M ; Williams K ; Worsham M ; Wu Y ; Agarwal D (2020): Location Identifiers, Metadata, and Map for Field Measurements at the East River Watershed, Colorado, USA. Watershed Function SFA. doi:10.15485/1660962

[7] Crystal-Ornelas, R., et. al. A guide to using Version Control Systems for developing and versioning data standards and reporting formats. Unpublished manuscript (under review). [8] Varadharajan, C., *et al.* "Challenges in Building an End-to-End System for Acquisition, Management, and Integration of Diverse Data From Sensor Networks in Watersheds: Lessons From a Mountainous Community Observatory in East River, Colorado," in *IEEE Access*, vol. 7, pp. 182796-182813, 2019, doi: 10.1109/ACCESS.2019.2957793. [9] Department of Energy. (n.d.). *About Us*. energy.gov. https://www.energy.gov/about-us.

Appendix

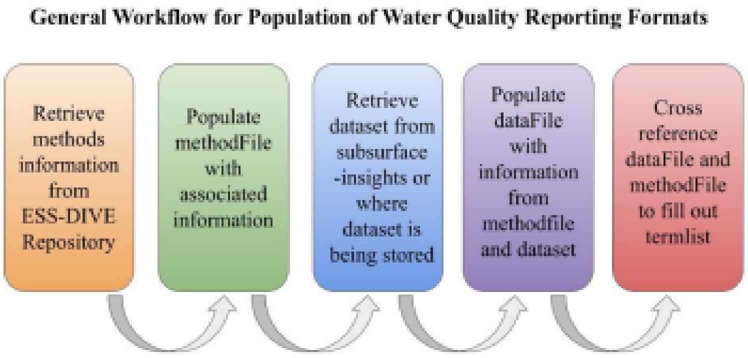
Participants of Project:

Name	Associated Institution Roles
Rob Crystal-Ornelas	LBNL Primary Mentor for Internship
Charuleka Varadharajan	LBNL Secondary Mentor for Internship
Kristin Boye	SLAC Utilized water/soil/sed chem reporting formats
Wenming Dong	LBNL Utilized data for reporting formats
Kenneth Williams	LBNL Utilized data for reporting formats
Roleof Versteeg	SSI Utilized API to access water quality data

Terms/Abbreviations:
Application Programming Interface = API
Data Management Framework = DMF
Department of Energy = DOE
Earth and Environmental Sciences Area = EESA
Environmental Systems Science Data Infrastructure for a Virtual Ecosystem = ESS-DIVE
Findable, Accessible, Interoperable, Reusable = FAIR
Lawrence Berkeley National Laboratory = LBNL
Subsurface Insights = SSI
Watershed Function Scientific Focus Area = WFSFA

Figures:

Figure 1: Workflow diagram for population of water quality reporting formats


General Workflow for Population of Water Quality Reporting Formats
Retrieve methods information from ESS-DIVE Repository
Populate methodFile with associated information
Retrieve dataset from subsurface -insights or where dataset is being stored
Populate dataFile with information from methodfile and dataset
Cross reference dataFile and methodFile to fill out termlist

Acknowledgements

Special thanks to Dr. Rob Crystal-Ornelas for his mentoring throughout the internship term. Additional thanks to Dr. Charuleka Varadharajan, the ESS-DIVE team, ESS-DIVE Community Partners, Dr. Varadharjan's Data Science team, WFSFA team, and WD&E staff for their support during the term of the internship.

The Watershed Function SFA and ESS-DIVE are funded by the U.S Department of Energy, Office of Science, Biological and Environmental Research Program.

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.