Utah State University
DigitalCommons@USU

Funded Research Records

9-11-2019

"Collaborative Research: Elements: Advancing Data Science and Analytics for Water (DSAW)"

Jeffery S. Horsburgh
Utah State University, jeffery.horsburgh@usu.edu

Follow this and additional works at: https://digitalcommons.usu.edu/funded_research_data

 Part of the Civil and Environmental Engineering Commons

Recommended Citation
Horsburgh, Jeffery S., ""Collaborative Research: Elements: Advancing Data Science and Analytics for Water (DSAW)"" (2019). *Funded Research Records*. Paper 118.
https://digitalcommons.usu.edu/funded_research_data/118

This Grant Record is brought to you for free and open access by DigitalCommons@USU. It has been accepted for inclusion in Funded Research Records by an authorized administrator of DigitalCommons@USU. For more information, please contact rebecca.nelson@usu.edu.

UtahStateUniversity
MERRILL-CAZIER LIBRARY

Elements: Advancing Data Science and Analytics for Water Science (DSAWS)

DATA MANAGEMENT PLAN

Types of Data and Materials to be Produced by the Project

Primary data collection or generation of new data through field or laboratory measurements is not anticipated as part of this project. The purpose of this project is to develop new cyberinfrastructure tools in the form of new Python software packages that better enable data science applications within the hydrology and water resources domain. The software we develop will be designed to operate on observational data of multiple types, including time series of hydrologic observations, geospatial datasets, including geographic feature dataset in shapefile format, geographic raster datasets in the form of GeoTiffs, and multidimensional space-time datasets in the form of NetCDF files. Our data science applications will use existing data of multiple types, including those previously listed and may produce derived datasets of the same types. This project will also produce new educational materials resulting from our proposed data science applications.

Data and Metadata Formats and Standards

The software we design and build for this project will make full use of the existing Dublin Core metadata standard (DCMI, 2012) used by the HydroShare data and model repository for describing data and modeling resources. It will also make use of data formats and types well known and accepted within the hydrology and water resources domain and as specified by HydroShare's Resource Data Model. For example, HydroShare uses Version 2 of the Observations Data Model (ODM2) for time series data (Horsburgh et al., 2016), the Network Common Data Form (NetCDF) for multidimensional space/time datasets (Rew and Davis, 1990), ESRI shapefiles for vector geospatial data, and the GeoTIFF format for geographic raster data. HydroShare's resource data model is an implementation of the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) standard (Lagoze et al., 2008). Resources are stored on disk and packaged for download using the BagIt hierarchical file packaging specification (Boyko et al., 2019). These standards are well known within the library, information science, and digital archiving communities. This combination of standard data formats, standardized metadata description, and standard packaging means that HydroShare resources are publishable and fully archivable.

The purpose of the Python packages we will develop is to facilitate data science applications using hydrologic and water resources data. Package functionality will enable moving hydrologic data Resources from HydroShare or USGS' NWIS system into performant data objects within a Python-based scientific computing environment for visualization and analysis. Metadata associated with these datasets (as supplied by HydroShare's resource data model or the USGS NWIS system) will be encoded as part of the Python object structure to ensure that important relationships between data and metadata are maintained. Users will be able to edit and modify metadata associated with datasets they manipulate in their Python computational environment and will then be able to copy resulting data products into the HydroShare system as new Resources using HydroShare's required and optional metadata. Any data products produced as part of our data science applications will be uploaded to the HydroShare repository and will be described using standard Dublin Core metadata elements as required by HydroShare.

Policies for Data and Research Products

A goal of the HydroShare system is to promote collaboration and sharing of data, models, and other research products. Any datasets resulting from our proposed data science applications will be made openly and publicly available under a Creative Commons license as Resources in HydroShare. Domain research results related to our proposed data science applications will be published in appropriate academic journals. All educational materials developed by this project will be openly and publicly available via a Creative Commons License as Resources in HydroShare, via the NSF-supported HydroLearn system, and referenced by CUAHSI's Data Driven education website.

As a general policy, source code developed by this project will be open source and will be distributed under the Berkeley Software Distribution 3-Clause Open Source License (BSD3). We will use open source code repositories in GitHub for our software development, which will enable us to coordinate development

activities of our project team and will enable potential contributions from individuals outside of the immediate project team who wish to contribute.

Plans for Archiving Data

For our software releases we will obtain snapshots of our GitHub repository(ies) and upload them to Zenodo to archive them and obtain citable digital object identifiers (DOIs). As an archival system, HydroShare will serve as the primary archival mechanism for all other data and research products created by this project. Curated research products published in HydroShare are citable for use in peer-reviewed journal articles, conference presentations and proceedings, and other formal publications using a formal DOI. HydroShare is operated by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) as an NSF-supported data facility, and all of its attendant systems are hosted on fault-tolerant, enterprise-class servers housed in the Renaissance Computing Institute's managed, climate controlled, UPS-backed Information Technology facility ensuring the reliability of the HydroShare system. HydroShare holds a growing body of knowledge that will be valuable to other research communities. Creating tools that facilitate sharing of reproducible and citable scientific results in HydroShare will broaden the impact of this project and encourage their use by a broader scientific community.