Funded Research and Data

9-15-2017

# Cyberinfrastructure for Advancing Hydrologic Knowledge Through Collaborative Integration of Data Science, Modeling and Analysis

David Tarboton

# Data Management Plan

## Types of Data

Primary data collection through field or laboratory measurements are not anticipated as part of this project. However, the purpose of HydroShare is to promote collaboration and sharing of data, models, and research products, and, as such, HydroShare is necessarily designed to support the full data life cycle. HydroShare users are already extensively using and sharing existing data from many different sources as research products and as inputs to shared models and analyses. These data include time series of observations from fixed monitoring locations (e.g., streamflow gages, weather stations, etc.), geospatial datasets (e.g., raster and feature datasets of land use, digital elevation models, etc.), and time varying gridded data such as radar-based precipitation products. Additional types of data and modeling resources that can be uploaded, shared, managed, and published in HydroShare include collections, web apps, model programs, model instances, and scripts. Finally, HydroShare includes a "generic" resource into which users can upload multiple types of files and describe them with metadata. Thus, HydroShare enables the sharing of not only data and model codes, but also other collaborative products such as codes for analytical tools and research results that may consist of images, documents, and other digital materials resulting from group collaborations.

To facilitate community access to the full range of anticipated data types and sources, HydroShare enables sharing of individual investigator datasets and derived data products. Continued development of the required repositories, protocols, and methods for doing so, along with the proposed new advances and functionality for HydroShare are part of the intellectual contribution of this project.

## Data and Metadata Standards

HydroShare makes full use of existing and emerging standards for sharing environmental datasets. All HydroShare resources (i.e, datasets, models, scripts, etc.) are described with metadata that conform to the Dublin Core metadata standard (DCMI, 2012), conform to a data model that is an implementation of the Open Archives Initiative's Object Reuse and Exchange (OAI-ORE) standard (Lagoze et al., 2008), and are stored on disk and packaged for download using the BagIt hierarchical file packaging specification (Boyko et al., 2012). These standards are well known within the library, information science, and digital archiving communities. We have also adopted standard file formats for the content files of known resource/data types. 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, ESRI shapefiles for vector geospatial data, and the GeoTIFF format for raster datasets. This combination of standard data formats, standardized metadata description, and standard packaging means that HydroShare resources are publishable and fully archivable. Indeed, an important aspect of our work on HydroShare has been to advance standardized data formats and metadata descriptions for contributed datasets, models, and other research products.

## Policies for Data and Research Products

The goal of HydroShare is to promote collaboration and sharing of data, models, and research analyses. Groups of researchers may wish to share data, model instances, or simulation results within their group before they are published externally. HydroShare provides users with the choice to create public or private resources and public or private collaboration groups, accessible only to selected users, within which these activities can take place. Authentication and access control have been fully integrated within HydroShare, and users can already choose how to share resources with other users or the larger community. Our experience has been that collaborative activities may result in multiple intermediate research products, only some of which may be considered publishable by the researchers. As such, HydroShare has functionality for creating formal, tracked versions of resources. Users can choose the

Creative Commons License under which their resources are shared, and HydroShare has already established the facilities required to formally publish data, models, and simulation results, enabling individual researchers to select and publish their results as they see fit. Formally published resources are made immutable and receive a citable digital object identifier (DOI).

Access to private resources and private research groups is at the discretion of resource and group owners. Final research results can be made freely and publicly available when they are deemed publication ready by the author. All HydroShare resources have a landing page the displays the resource's metadata and contents, including attribution information (i.e., authors and contributors, funding agency credits, etc.) and a formal citation. HydroShare users must agree to a formal publication agreement prior to formally publishing a resource. This agreement was developed in collaboration with the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) and specifies the terms and conditions under which users can publish resources in HydroShare.

As a general policy that is already in place with our team, all source code developed by this project is (and will be) created using an open development model and will be distributed under the BSD 3-Clause Open Source License. We have already established a GitHub organization for HydroShare, and all HydroShare source code is already publicly available within the repositories organized under this organization. GitHub and the software development best practices we have already established enable us to coordinate our development activities across multiple Universities and engage developers and contributors from outside of the immediate project team who wish to contribute.

**Plans for Archiving Data**

Researchers curate and archive data, derived data products, scripts, model programs, and model instances as research products within HydroShare to better enable collaboration and reproducibility of research results. As an archival system, HydroShare itself serves as the primary archival mechanism for curated products created by community researchers. Curated research products published in HydroShare are citable for use in peer-reviewed journal articles, conference presentations and proceedings, and other formal publications. HydroShare and all of its attendant systems are hosted on fault-tolerant, enterprise-class servers dedicated to this project and housed in RENCI's managed, climate controlled, UPS-backed IT facility ensuring the reliability of the HydroShare system.

HydroShare is already gaining momentum and will eventually contain a hydrology body of knowledge that will be valuable to other research communities. As a sustainability mechanism (see our Sustainability Plan Section), we will be investigating partnerships with institutional repositories for long term storage and archival of published resources. Ideally, we will transfer formally published resources to an institutional repository for archival, easing the burden on HydroShare's production website to maintain this role.