Funded Research Records

9-13-2019

Earth Cube Data Capabilities: Collaborative Research: Deep Integration of Reproducibility in Community Portals

David G. Tarboton
Utah State University, david.tarboton@usu.edu

Data Management Plan

Types of Data

Primary data collection through field or laboratory measurements are not anticipated as part of this project. The products from this project will be software that is either (1) part of Sciunit or (2) deployed as part of HydroShare or installed on platforms linked to HydroShare to support reproducible research. Use cases will include observational data and model and analysis results, or benchmarks.

Software Management

Expected software includes code in C, C++, and Python. 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. HydroShare source code is already publicly available in github (https://github.com/hydroshare/hydroshare). Sciunit is hosted in bitbucket (https://bitbucket.org/geotrust/sciunit2) 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.

Data Management

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. HydroShare has 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.

This project will adhere to the same data and metadata standards as adopted by HydroShare. In particular the Sciunit interface with HydroShare will be based on standards adopted by HydrShare.

Policies for Data and Research Products

The goal of HydroShare is to promote collaboration and sharing of data, models, and research analyses. The goal of this project is to enable and enhance the reproducibility of model and analysis workflows that are published in HydroShare. Groups of researchers may wish to share data, model instances, or simulation results within their group before they are published externally. Upon completion of this project, this functionality will extend to Sciunit objects in support of reproducibility. 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.

Plans for Archiving Data

Github, with snapshots to Zenodo of specific code commits for which we will obtain a DOI will serve as the archival system for code products from this research. HydroShare will serve as the Archival system for data and model results from 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. 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.

References

Boyko, A., J. Kunze, J. Littman, L. Madden and B. Vargas, (2012), The BagIt File Packaging Format (v0.97) Network Working Group Internet Draft, http://tools.ietf.org/html/draft-kunze-bagit-10, accessed 2/6/2014.

DCMI, (2012), DCMI Metadata Terms, Dublin Core Metadata Initiative (DCMI), http://dublincore.org/documents/dcmi-terms/, accessed 9/15/2016.

Horsburgh, J. S., A. K. Aufdenkampe, E. Mayorga, K. A. Lehnert, L. Hsu, L. Song, A. S. Jones, S. G. Damiano, D. G. Tarboton, D. Valentine, I. Zaslavsky and T. Whitenack, (2016), "Observations Data Model 2: A community information model for spatially discrete Earth observations," Environmental Modelling & Software, 79: 55-74, http://dx.doi.org/10.1016/j.envsoft.2016.01.010.

Lagoze, C., H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson and S. Warner, (2008), Open Archives Initiative Object Reuse and Exchange: ORE User Guide – Primer, http://www.openarchives.org/ore/1.0/primer, accessed 10/23/2012.