

Utah State University

DigitalCommons@USU

Funded Research Records

7-21-2020

Collaborative Research: Network Hub: Enabling, Supporting, and Communicating Critical Zone Research

Jeffery S. Horsburgh

Utah State University, jeff.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, J. S. (2020). Collaborative Research: Network Hub: Enabling, Supporting, and Communicating Critical Zone Research. Utah State University. <https://doi.org/10.26078/y3yp-dd97>

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 digitalcommons@usu.edu.



Data Management Plan
Collaborative Research: Network Hub:
Enabling, Supporting, and Communicating Critical Zone Research

This Data Management Plan describes our approach for establishing data aggregation, archival, and dissemination for the new network of Critical Zone (CZ) Thematic Clusters, as well as services we will provide for legacy CZO data. Our goal is to assist Cluster data managers and investigators in curating high-quality data and research products that are Findable, Accessible, Interoperable, and Reusable by any potential user, conforming to the FAIR Data Principles (Wilkinson et al., 2016). We are committed to secure archival of primary data and ensuring accessibility of Network research products for the broad scientific community, but we also understand that our Data Management Plan is dependent on the cooperation of the Cluster PIs and data managers. It is our intent to work closely with each Cluster team and with the PIs to develop a strong CZ community that encourages cooperation and collaboration among the Clusters and with the Coordinating Hub. Moreover, we intend to provide services and cyberinfrastructure (CI) that are attractive to Clusters, thereby providing them with the best and easiest solution for their data archival and dissemination. We will create the CZ Network Hub as a unifying platform for investigation, discovery, training, and education. Data and research products from across the Collaborative Network will be archived in perpetuity, easily discoverable, and openly available.

Types of data: Given that we do not yet know the location or focus of the new Clusters, we can only anticipate the types of data to be collected. However, based on past CZ studies and our direct interaction with CZOs through CUAHSI and EarthChem data archival efforts, we anticipate there will be data types common to all clusters, including perhaps: 1) documents such as site descriptions, white papers, publications, presentations, curricular materials, and reports; 2) data generated from fixed points with single or repeated sampling, such as time series data (e.g., atmospheric moisture, groundwater table depth, sapflow, soil chemistry, soil moisture, water chemistry, snowpack, etc.); 3) synoptic or survey sampled data, such as distributed points sampled simultaneously once or at regular or irregular intervals - e.g., stormflow sampling events, ecosystem surveys, ground penetrating radar, or seismic surveys; 4) geospatial datasets, including land use, stream networks, surficial and bedrock geology, topography, and vegetation; and 5) time varying regular or irregular gridded data - e.g., radar precipitation products or output from groundwater models. Modeling products will also play a key role in advancing understanding of how the CZ forms, operates, and evolves; provide a mechanism to generalize and extend observation results beyond individual CZ sites; and allow prediction of the CZ response to changes. Archiving models to enhance reproducibility and further development is key. Our targeted workshops with Cluster data managers and investigators planned during the first two years of the project will be focused on identifying and, to the extent possible, standardizing data types, formats, and metadata descriptions.

Data and metadata standards: Our goal is to create and promote standards and CI that enable Cluster data managers to compile consistent and interoperable data both within and across Clusters. Based on our experience working with CZOs and other major field data collection efforts, we know that accomplishing this will be challenging. Clusters may have different observed variables, different methods to make measurements, different observation intervals, and varying approaches for metadata. Geospatial and laboratory chemistry datasets may come in a wide variety of formats with little consistency in metadata or the types of data available at each site. Thus, it will be critical for our team to engage Cluster data managers and PIs in our early project workshops to establish both common data and metadata archival standards to the extent possible to enhance compatibility across Clusters. The proposed Data Submission Portal will promote these standards through automated validation of submitted data and metadata prior to deposition into the appropriate long-term repository and through a controlled vocabulary supported by the CZ Community.

We will promote common data formats that are open (e.g., CSV files versus Excel files) and metadata that provide at least the minimum elements required by the repository into which the data are deposited - e.g., HydroShare uses standard Dublin Core metadata elements (DCMI, 2012) for describing datasets. We

will work with Cluster data managers to ensure that submitted data are well formatted and described to meet interoperability and reusability benchmarks. We will also facilitate collaboration among Cluster data managers to standardize vocabularies (e.g., naming of observed variables), file naming conventions, spatial reference systems for geospatial datasets, sensor types (where possible), and data collection intervals to facilitate integration of data across sites.

Policies for data and research products: Data collection, quality assurance/quality control, and annotation with metadata will be the responsibility of Cluster teams but will be conducted using protocols developed in collaboration with the Coordinating Hub. All datasets submitted by Clusters to the Data Submission Portal will be made freely and publicly available in the appropriate long-term repository (e.g., HydroShare, EarthChem, OpenTopography, etc.) using the data format and metadata standards described above. These repositories will be the primary dissemination mechanism for data from the Clusters. We will encourage and, to the extent possible, facilitate rapid publication of Cluster data to enable widespread use. All finalized data and research products will be made publicly available. Researchers may choose to apply an embargo of up to two years, following practices established by the existing CZO network. Up to the point of publication, researchers may also choose to restrict access to a selected group. All data and research products published by Clusters will include appropriate attribute and citation information. Access to published data and products may include an access/use agreement that conforms with established practices or policies to be agreed upon by the collaborating Clusters. This will ensure that the intellectual rights of Cluster investigators are protected while granting redistribution rights to the Coordinating Hub for purposes of open, long-term data sharing. Synthesis teams may want access to data that is not yet public. We will work with Clusters and the synthesis teams to broker agreements that allow access to data for synthesis but protect the interests of those who created the data. All source code developed by our team will be open source and hosted within open GitHub repositories.

Plans for archiving data: EarthChem and HydroShare will serve as our two primary archival and discovery platforms. EarthChem and the System for Earth Sample Registration (SESAR) are data systems for the solid Earth sciences that provide open access to user-submitted data in the trusted EarthChem Library data repository, and to over 30 million geochemical, petrological, and environmental observations in curated synthesis databases. SESAR operates a registry and metadata catalog for physical samples and tools for researchers and repository curators to share, manage, preserve, and access information about samples and uniquely identify them by registration with IGSN (International Geo Sample Number). SESAR manages sample metadata and related data such as images and scanned notebooks. The proposed activities include efforts to maintain, grow, and improve data collections and the software applications used to ingest and curate data and metadata and make the data accessible to the public online. EarthChem and SESAR follow established procedures for data quality assurance and use software applications for data and metadata assembly and ingestion. HydroShare is a collaborative platform for data sharing, management, and publication operated by CUAHSI. Described in detail in the proposal, HydroShare supports data types commonly used within the CZ community (evidenced by the large amount of Legacy CZO data already in HydroShare), but also accepts files of any format. HydroShare can be used to manage collaborative data access so that Clusters can keep their data private until they are ready to publish it. HydroShare also provides the capacity for formally publishing data sets with a digital object identifier (DOI). HydroShare also has the capacity to capture workflows so that data management or modeling workflows are reproducible. HydroShare and EarthChem are operated as NSF-supported data facilities, and all of their attendant systems are hosted on fault-tolerant, enterprise-class servers housed in managed, climate controlled, UPS-backed Information Technology facilities, ensuring their reliability.

It is likely that not all CZ data are appropriate for EarthChem and HydroShare. For example, if topographic data are collected, those data are most appropriately archived in OpenTopography. We will develop a Memorandum of Understanding with OpenTopography (see Letter of Support) to facilitate archival of CZ topographic data. Should other repositories be needed, we will make the necessary connections that facilitate use of the repositories by the Clusters.