



Linking Domain Repositories to Build Cyberinfrastructure for Interdisciplinary Critical Zone Research

Jeffery S. Horsburgh, Kerstin Lehnert, Jerad Bales



Support: 2012893,
2012748, 2012593

Goal

- Provide a robust CI for Findable, Accessible, Interoperable, and Reusable (FAIR) data from the CZ network Thematic Clusters

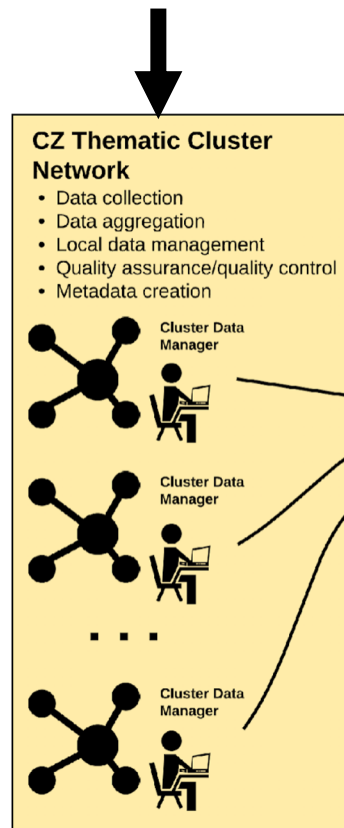
Challenges

- CZ research and data are diverse
- Existing repositories have good functionality but not coordinated
- No single repository meets all needs
- Not always clear which repository to use

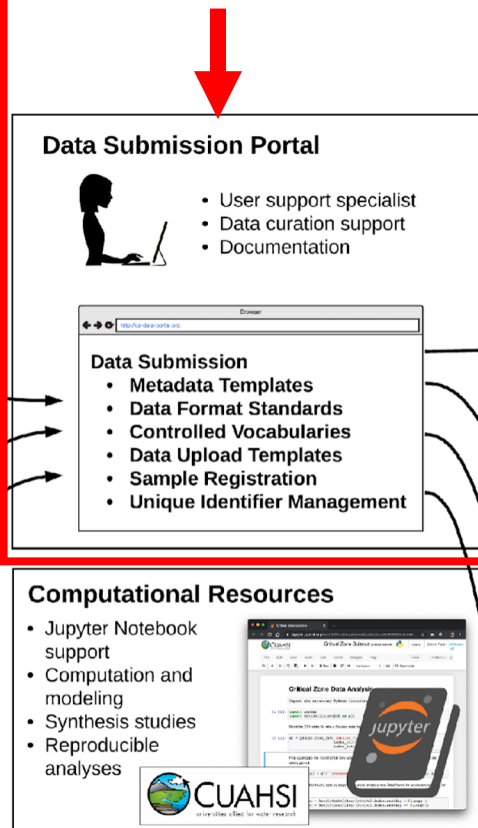
Solution

- Distributed architecture that links existing data facilities and services (e.g., HydroShare, EarthChem, SESAR, OpenTopography, etc.) via a central hub with services for easy data submission, integrated data discovery and access, and computational resources

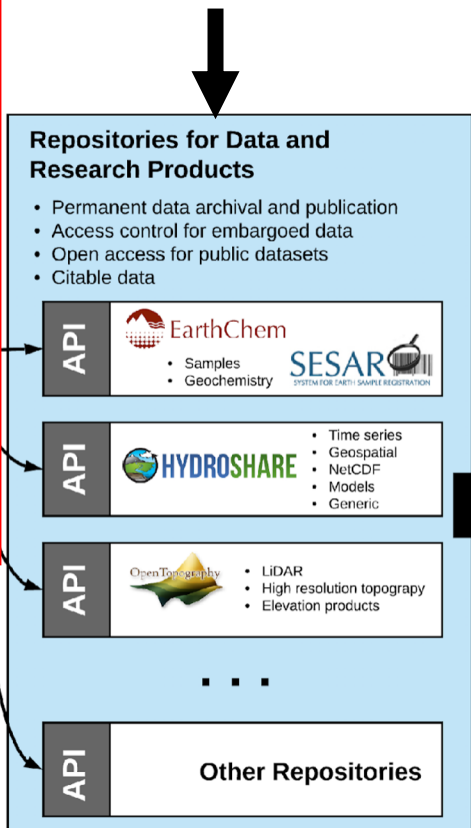
Diverse data and research products from CZ scientists



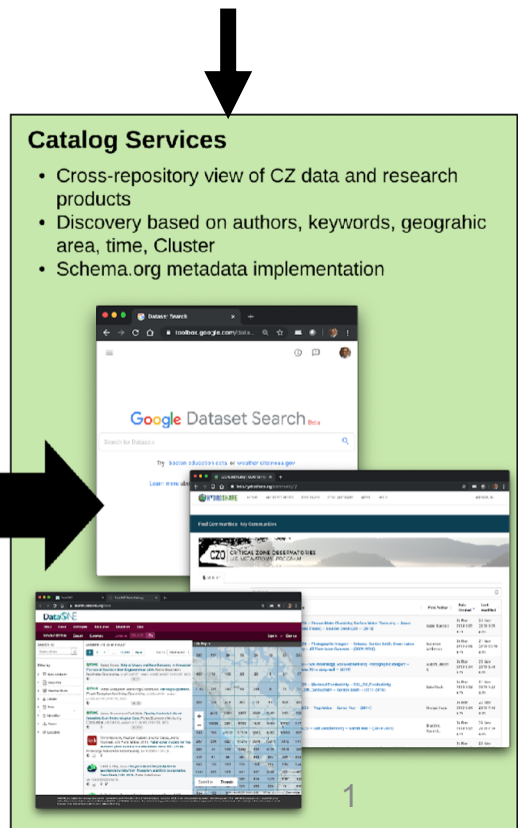
New Submission Portal promoting repository selection, metadata, templates, and formats



Using existing Earth science data repositories via automated submission

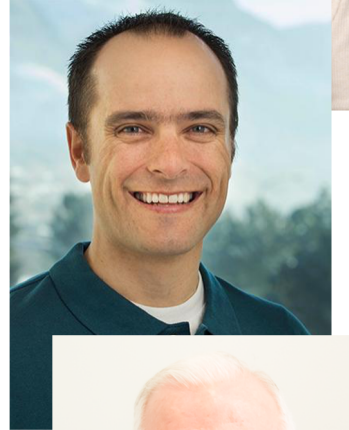


Making products Findable, Accessible, Interoperable, and Reusable (FAIR)



Our Team

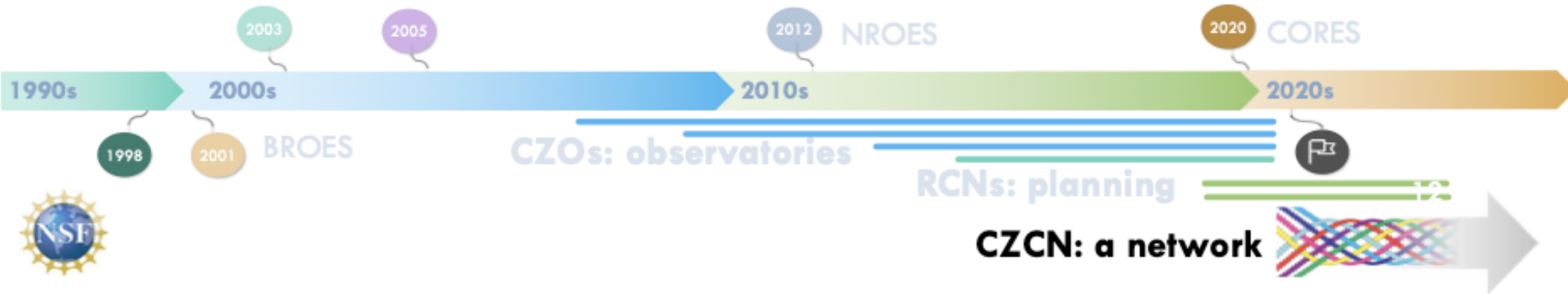
- **Kerstin Lehnert**: Lamont Doherty Earth Observatory, Columbia University
- **Jeff Horsburgh**: Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University
- **Chris Calloway**: Renaissance Computing Institute, University of North Carolina Chapel Hill
- **David Tarboton**: Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University



SHAPING CZ SCIENCE @ NSF: Critical Zone Collaborative Network

The Awards: nine clusters and one hub

- All thematic clusters include sites that extend the CZ network to nationwide (lower 48 & P.R.)
- Funding from NSF Division of Earth Sciences Critical Zone program as well as Hydrological Sciences, Geobiology & Low-T Geochemistry, Geomorphology and Land-use Dynamics, Cross-cutting activities, Education and Human Resources, in addition to Atmosphere & Geospace Sciences, Environmental Biology, Engineering Directorate and EPSCOR.



Objective of the CZ Hub CI

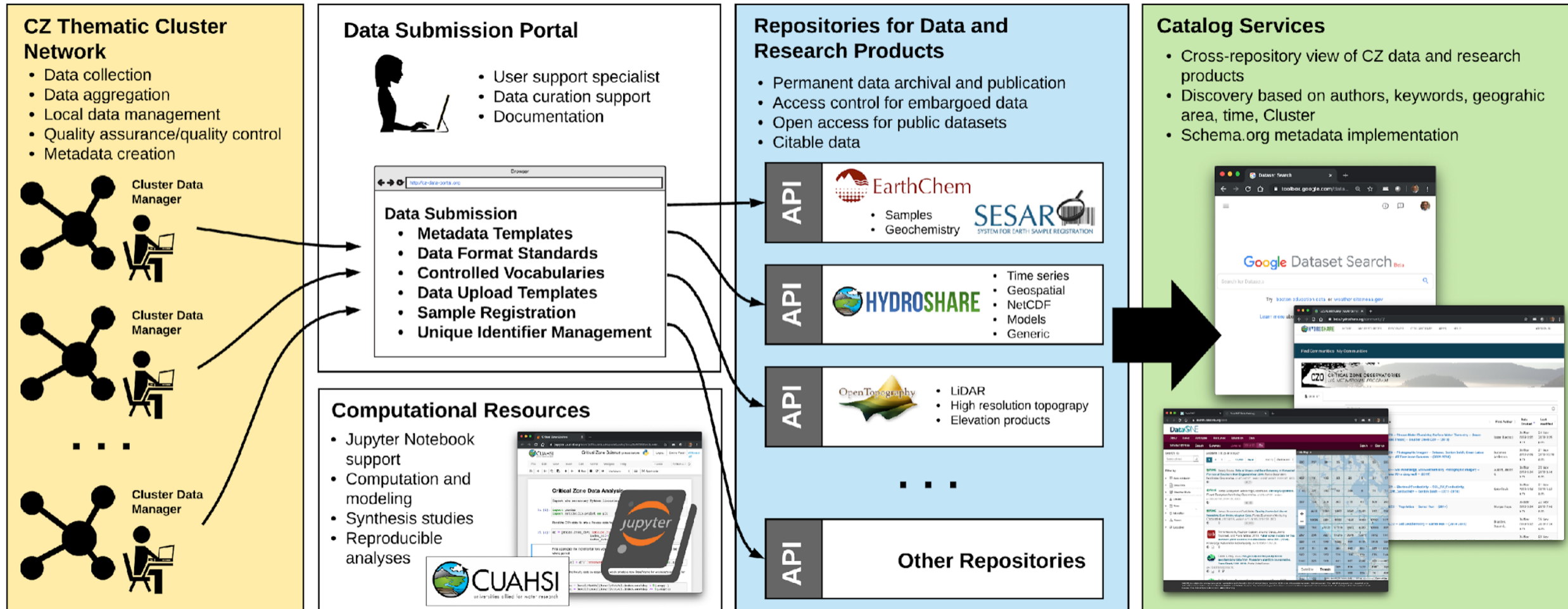
Provide a robust cyberinfrastructure for Findable, Accessible, Interoperable, and Reusable (FAIR) data from the Critical Zone Collaborative Network Thematic Cluster Projects

Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3:160018, <https://doi.org/10.1038/sdata.2016.18>.

Our Approach

- Link existing data facilities and services, including:
 - HydroShare
 - EarthChem
 - System for Earth Sample Registration (SESAR)
 - OpenTopography
 - Other repositories, as needed
- Develop a central CZ Hub that provides
 - Services for easy data submission
 - Integrated data discovery and access
 - Computational resources for data analysis and visualization in support of CZ synthesis efforts

Our Approach

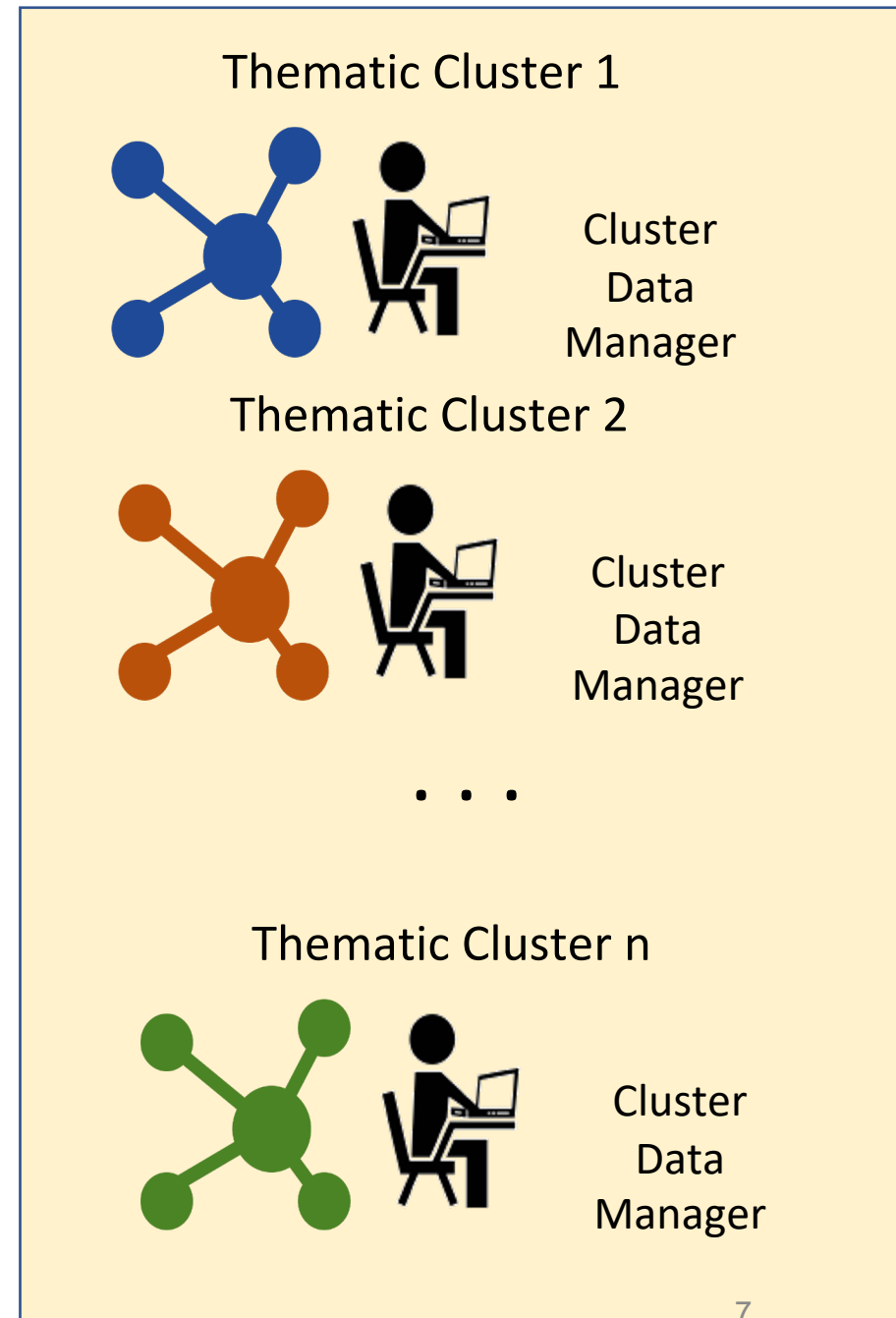


Critical Zone Collaborative Network (CZCN)

- Thematic Cluster activities built around a scientific theme (not a site)
- Thematic cluster data activities
 - Data management plans are flexible and appropriate for each theme
 - Routine and event data collection
 - Data aggregation
 - Data QA/QC
 - Metadata creation

We aim for early engagement that:

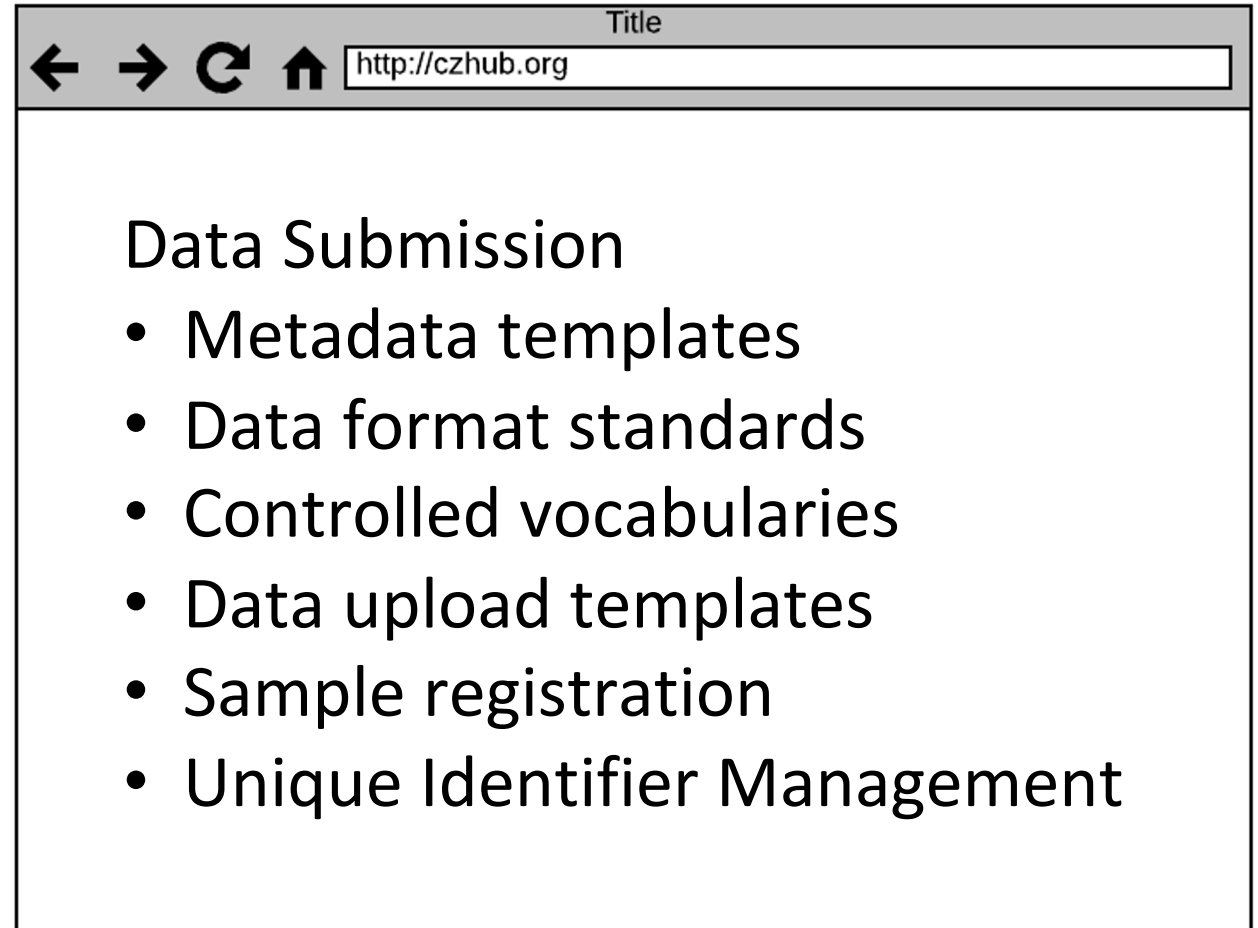
- Learns from NSF's 12-year CZO program
- Builds a CZ Data Management Community
- Builds and implements best practices for consistency



Data Submission Portal

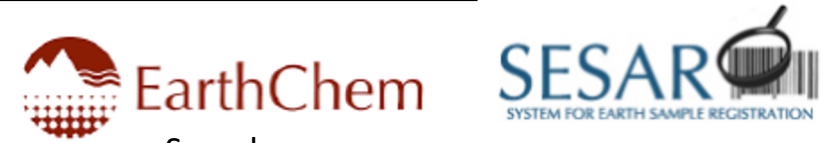
- New, web-based development to support network
- Data curation support
- Getting data to the right repository

Empower Cluster data managers to curate data products within appropriate repositories



Repositories for Data and Research Products

- No single repository will work for all CZ Data
- Operate and partner with existing repositories
 - Promote the use of FAIR principles
 - Permanent data archival and publication
 - Access control for embargoed data
 - Open access for public datasets
 - Citable data
 - Leverage existing NSF investment in CI
- User support specialists as Hub-level data managers



- Samples
- Geochemistry, mineralogy, XRD, etc.



- Time Series
- Geospatial
- NetCDF
- Models
- Generic



- LIDAR
- High resolution topography
- Elevation products

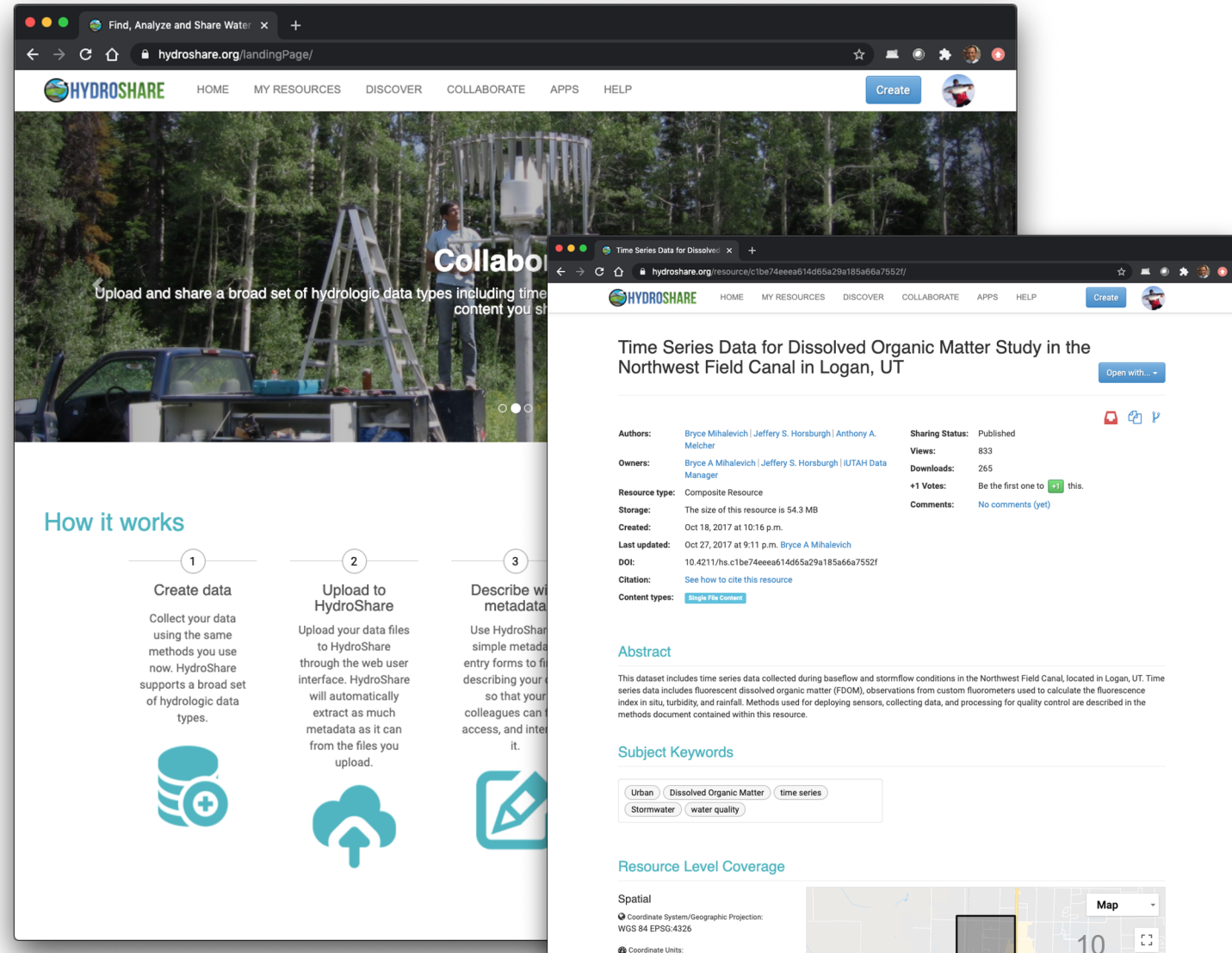
...

Other Repositories

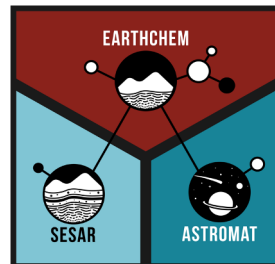
HydroShare Data and Model Repository

- Operated by CUAHSI as a community resource
- Manage data (and other products) throughout the research life cycle
- Sharing and formal publication with citable digital object identifiers (DOIs)
- Much of the legacy CZO data has been moved to HydroShare

<http://www.hydroshare.org>



EarthChem SESAR



www.earthchem.org
www.geosamples.org

- Preservation, discovery, access, and advanced mining of geochemical, petrological, geochronological, & mineralogical data and samples
- **EarthChem Library (ECL)** recommended by publishers as trusted data repository
 - DOI assignment; long-term archiving; links to publications, funding awards, & samples
 - Already hosts geochemical data from CZOs
- **EarthChem Synthesis**: extract & integrate data across thousands of datasets
- **SESAR**: Manage & share sample metadata; register samples with IGSN
- **Best Practices** for geochemical & sample data (e.g. OneGeochemistry)

ECL: Dataset Landing Page

EarthChem Library

Data DOI: 10.1594/IEDA/111144

Citation: Chorover, J., Perdrial, J., McIntosh, J., Troch, P., Amistadi, M., Losleben, M., Condon, K., Pedron, S. 2018. Jemez River Basin Soil Solution Chemistry 2014 (New Mexico, USA), Version 1.0. Interdisciplinary Earth Data Alliance (IEDA). <https://doi.org/10.1594/IEDA/111144>. Accessed 2020-08-14.

Title: Jemez River Basin Soil Solution Chemistry 2014 (New Mexico, USA)

Creator(s): Chorover, J., Perdrial, J., McIntosh, J., Troch, P., Amistadi, M., Losleben, M., Condon, K., Pedron, S.

Abstract: Soil solution samples in the Jemez River Basin field sites of the Catalina-Jemez Critical Zone Observatory (CZO) are the following two types of soil solution samplers: i) Prenart Super Quartz suction cups (www.prenart.dk). Prenart suction optimized for all chemistry analyses and were installed without addition of Si-slurry to allow for artifact-free Si analysis. Suction for each Prenart is ~ 60 kPa. ii) Custom made, fiberglass wick-based passive capillary wick samplers (PCaps) (Perdrial et al. 2012). PCaps are optimized for water flux determination and sampling for organic carbon, most (non-carbonate) metals. PCap samples should however not be used for major cations (Na, Mg, Si, K, Ca) and dissolved inorganic carbon artifacts from the fiberglass materials (see Perdrial et al (2014) for a complete list). Passive (continuous) suction was of the hanging water column, is ~3 kPa. Soil solution samplers were installed in each of six pedons in the Mixed Conifer Basin (MC-ZOB) and the fire impacted site (2011 Burned ZOB) at 3 (PCaps) and 4 (Prenarts) depths, respectively. Pits were selected to capture differences in catchment aspect (MC-ZOB SE facing: Pit 3 and 4, NW facing: Pit 1 and 6, MC-ZOB: hollow Pit 2 and 5, planar Pit 1 and 6, divergent Pit 3, convergent Pit 4), elevation and burn severity (co-located with Decagon soil moisture and Ridge fire) in June 2013 and was then renamed to 2013 Burned ZOB.

Keyword(s): Coverage Scope: Other
Geographic Location: Valles Caldera, Jemez River Basin, New Mexico

Related Publication(s): Perdrial, J.N., Perdrial, N., Harpold, A., Gao, X., LaSharr, K.M., Chorover, J. (2012) Impacts of sampling dissolved organic carbon in soil solution. *Soil Science Society of America Journal*, 76: 2019-2030, doi: 10.2136/sssaj2012.0061.

Vazquez-Ortega, A., Perdrial, J., Harpold, A., Zapata-Rios, X., Rasmussen, C., McIntosh, J., Schaap, M., Geller, J. (2013) Reactive tracers of biogeochemical weathering in forested rhyolitic terrain. *Chemical Geology*, 391: 19-32, doi: 10.1016/j.chemgeo.2013.07.027.

Perdrial, J. N., Perdrial, N., Vazquez-Ortega, A., Porter, C., Leedy, J., and Chorover, J. (2014) Experimental assessment of soil solution sampling methods. *Soil Science Society of America Journal*, 78(2): 486-495, doi: 10.2136/sssaj2013.07.0279.

Vazquez-Ortega, A., Huckle, D., Perdrial, J., Amistadi, M. K., Durcik, M., Rasmussen, C., McIntosh, J., Chorover, J. (2014) Soil solution sampling methods for the Jemez River Basin. *Chemical Geology*, 428: 1-18, doi: 10.1016/j.chemgeo.2014.01.004.

Publication DOI: <https://doi.org/10.1594/IEDA/111144>

User Contributed Keyword(s): Soil solution, Sodium, Lead, Urea

Date Available: 2018-02-

Date Created: 2018-02-

Resource Type: Dataset

File Format(s): application

Funding source(s): National Science Foundation

Related Information

IsReferencedBy: DOI: [10.2136/sssaj2012.0061](https://doi.org/10.2136/sssaj2012.0061)

References:

- Sample ID: [IGSN: IECJCH800](#)
- Sample ID: [IGSN: IECJCH801](#)
- Sample ID: [IGSN: IECJCH802](#)
- Sample ID: [IGSN: IECJCH803](#)
- Sample ID: [IGSN: IECJCH804](#)
- Sample ID: [IGSN: IECJCH805](#)
- Sample ID: [IGSN: IECJCH806](#)
- Sample ID: [IGSN: IECJCH807](#)
- Sample ID: [IGSN: IECJCH808](#)
- Sample ID: [IGSN: IECJCH809](#)
- Sample ID: [IGSN: IECJCH810](#)
- Sample ID: [IGSN: IECJCH811](#)
- Sample ID: [IGSN: IECJCH812](#)
- Sample ID: [IGSN: IECJCH813](#)
- Sample ID: [IGSN: IECJCH814](#)
- Sample ID: [IGSN: IECJCH815](#)

ECL: Dataset links to samples

Return to ECL

SESAR: Sample Profile Page

IGSN: IECJCH800

IGSN: IECJCH800
Sample Name: JEMEZ_G918-Burn_Pit1-10cm
Other Name(s):
Sample Type: Individual Sample>Liquid
Parent IGSN: IECJCH008

Description

Material: Liquid>aqueous
Classification: Not Provided
Field Name: 2011 Burned ZOB Pedon 1
Description: Soil solution
Age (min): Not Provided
Age (max): Not Provided
Collection Method: Wick
Collection Method Description: Custom made, fiberglass wick-based passive capillary wick samplers (PCaps, Perdrial et al. 2012)
Size: 525 mL
Geological Age: Not Provided
Geological Unit: Not Provided
Comment: Analysis Notes: TL: Sampling Notes: evacuated until could no longer draw pressure; did not measure full volume;
Purpose: Not Provided

Geolocation

Latitude (WGS84): 35.850556
Longitude (WGS84): -106.450015
Northing (m) (UTM NAD83): 3968343
Easting (m) (UTM NAD83): 369064
Zone: 13N
Vertical Datum: NAVD88
Elevation Start: 2707.1 meters
Elevation End: Not Provided
Nav Type: GPS
Physiographic Feature: Resurgent dome
Name Of Physiographic Feature: Valles Caldera
Location Description: Not Provided
Locality: Rabbit Mountain
Locality Description: Not Provided
Country: United States
State/Province: New Mexico
County: Sandoval

Catalog Services

- Cross-repository view of CZ data and research products
- Discovery based on authors, geographic area, time, cluster
- Schema.org metadata
- Communities and Groups in HydroShare

A coordinated view and data discovery service(s) for all the data produced within the collaborative network to ensure that data are **Findable and Accessible.**

HydroShare datasets discoverable via Google Dataset Search

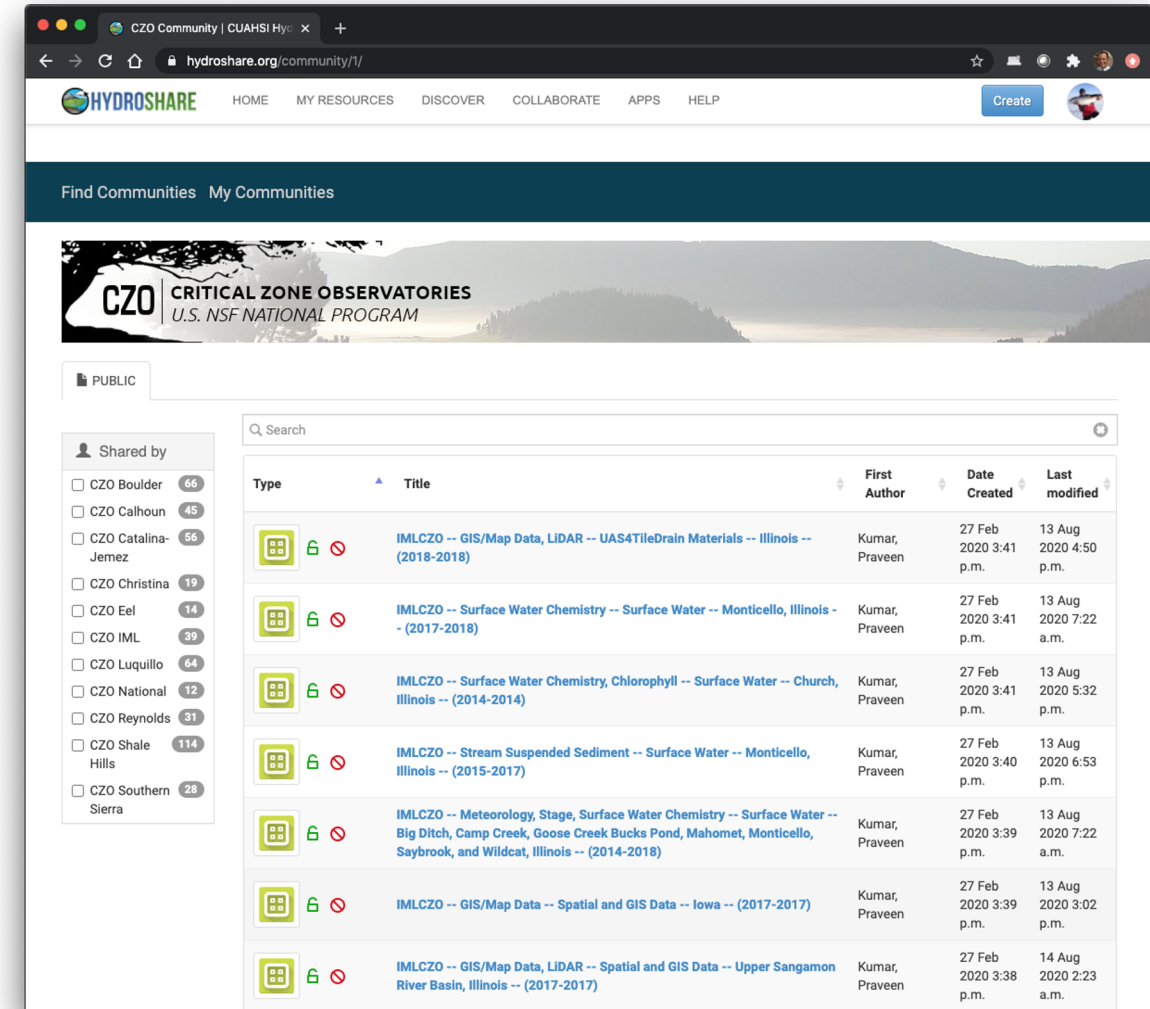
The screenshot shows a Google Dataset Search results page for the query 'Logan River'. The search bar at the top is circled in red. Below the search bar, there are filters for 'Last updated', 'Download format', 'Usage rights', 'Topic', and 'Free'. The results list shows 100+ datasets found. The first result is 'SURVEY-LOGAN RIVER, CACHE COUNTY, UT, USA' from catalog.data.gov. The second result, 'Logan River Observatory: Logan River near Franklin Basin Aquatic Site (LR_FB_BA) Quality Controlled Data', is highlighted with a red circle around its title and a red arrow pointing to the 'Explore at HydroShare' button. The third result is 'Discharge Rating Curve at Blacksmith Fork above...' from www.hydroshare.org. The fourth result is 'Albert and Logan River, South East Queensland Peri-Urban SuperSit...' from researchdata.edu.au. The fifth result is 'Logan River 23 M and 135 M tidal barrage sites seismic refraction...' from researchdata.edu.au. The sixth result is 'Logan River Continuous Water Quality Station' from CSIRO. On the right side, there is a detailed view of the 'Logan River Observatory' dataset, showing its update date (Aug 5, 2020), provider (HydroShare), license (Attribution 4.0 (CC BY 4.0)), area covered, dataset funding (National Science Foundation), and a description of the data.

Catalog Services

- Cross-repository view of CZ data and research products
- Discovery based on authors, geographic area, time, cluster
- Schema.org metadata
- Communities and Groups in HydroShare

A coordinated view and data discovery service(s) for all the data produced within the collaborative network to ensure that data are **Findable and Accessible.**

CZO “Community” in HydroShare with individual “Groups” for each observatory



The screenshot shows the HydroShare web interface for the CZO (Critical Zone Observatories) community. The page has a header with the HydroShare logo and navigation links (HOME, MY RESOURCES, DISCOVER, COLLABORATE, APPS, HELP). Below the header is a banner for CZO with the text "CRITICAL ZONE OBSERVATORIES U.S. NSF NATIONAL PROGRAM". A sidebar on the left lists various CZO groups with their member counts, such as "CZO Boulder" (66), "CZO Calhoun" (45), and "CZO Shale Hills" (114). The main content area displays a table of data groups, each with a "Type" icon, a "Title", "First Author", "Date Created", and "Last modified" date. The table lists several data groups, including "IMLCZO -- GIS/Map Data, LIDAR -- UAS4TileDrain Materials -- Illinois -- (2018-2018)" and "IMLCZO -- Surface Water Chemistry -- Surface Water -- Monticello, Illinois -- (2017-2018)".

Type	Title	First Author	Date Created	Last modified
IMLCZO	IMLCZO -- GIS/Map Data, LIDAR -- UAS4TileDrain Materials -- Illinois -- (2018-2018)	Kumar, Praveen	27 Feb 2020 3:41 p.m.	13 Aug 2020 4:50 p.m.
IMLCZO	IMLCZO -- Surface Water Chemistry -- Surface Water -- Monticello, Illinois -- (2017-2018)	Kumar, Praveen	27 Feb 2020 3:41 p.m.	13 Aug 2020 7:22 a.m.
IMLCZO	IMLCZO -- Surface Water Chemistry, Chlorophyll -- Surface Water -- Church, Illinois -- (2014-2014)	Kumar, Praveen	27 Feb 2020 3:41 p.m.	13 Aug 2020 5:32 p.m.
IMLCZO	IMLCZO -- Stream Suspended Sediment -- Surface Water -- Monticello, Illinois -- (2015-2017)	Kumar, Praveen	27 Feb 2020 3:40 p.m.	13 Aug 2020 6:53 p.m.
IMLCZO	IMLCZO -- Meteorology, Stage, Surface Water Chemistry -- Surface Water -- Big Ditch, Camp Creek, Goose Creek Bucks Pond, Mahomet, Monticello, Saybrook, and Wildcat, Illinois -- (2014-2018)	Kumar, Praveen	27 Feb 2020 3:39 p.m.	13 Aug 2020 7:22 a.m.
IMLCZO	IMLCZO -- GIS/Map Data -- Spatial and GIS Data -- Iowa -- (2017-2017)	Kumar, Praveen	27 Feb 2020 3:39 p.m.	13 Aug 2020 3:02 p.m.
IMLCZO	IMLCZO -- GIS/Map Data, LIDAR -- Spatial and GIS Data -- Upper Sangamon River Basin, Illinois -- (2017-2017)	Kumar, Praveen	27 Feb 2020 3:38 p.m.	14 Aug 2020 2:23 a.m.

Computational Resources

- Repository linkages
- Jupyter Notebook support
- Computation and modeling
- Synthesis studies
- Reproducible analyses – sharing data, results, and the executable code used to generate/recreate them
- Enhanced trust in research through transparency, replicability and reproducibility

HydroShare is a web based gateway to computing

The screenshot shows the HydroShare website with the title 'Introduction to TauDEM'. On the right, a dropdown menu titled 'Open with...' is open, listing several applications: HydroShare GIS, CUAHSI JupyterHub, OPeNDAP, CyberGIS-Jupyter for Water, and MATLAB Online. The menu is highlighted with a red rectangular box. The page also displays metadata for the resource, including authors (David Tarboton), owners (David Tarboton), resource type (Composite Resource), storage size (54.2 MB), creation date (Dec 08, 2019), and sharing status (Public).

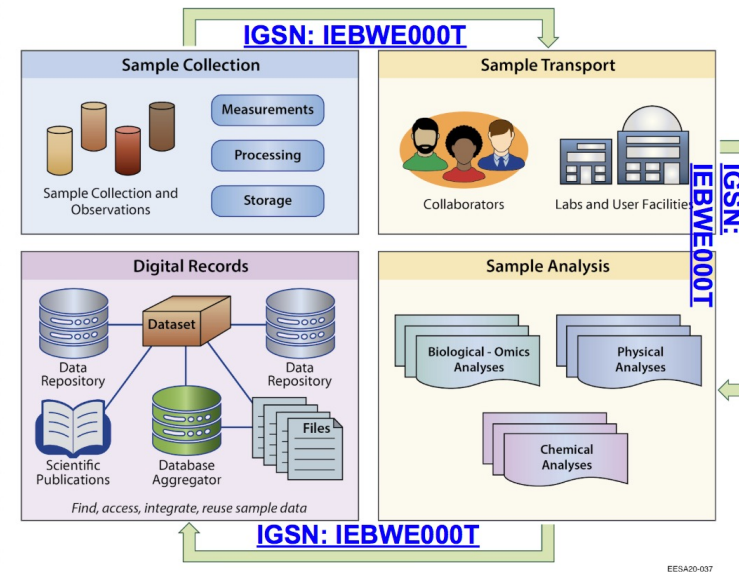
The screenshot shows the JupyterLab file browser interface. The file 'TauDEM.ipynb' is highlighted with a red rectangular box. The interface shows a directory structure with 'Downloads' and 'data' folders. The file 'TauDEM.ipynb' is listed with a status of 'Running' and a size of 17.8 kB. Other files like 'logan.tif' and 'logan.vrt' are also visible.

The screenshot shows the Jupyter Notebook interface for the 'Hydrologic Terrain Analysis Using TauDEM' notebook. The notebook title is 'Hydrologic Terrain Analysis Using TauDEM'. The introductory text states: 'The purpose of this notebook is to introduce **Terrain Analysis Using Digital Elevation Models (TauDEM)** software for Hydrologic Terrain Analysis in Jupyter. TauDEM is a free and open source set of Digital Elevation Model (DEM) tools for the extraction and analysis of hydrologic information from topography as represented by DEM. This software is developed at Utah State University (USU) for hydrologic digital elevation model analysis and watershed delineation.'

Resources for Samples

- SESAR: Best practices & tools for FAIR samples
 - Personal or institutional workspaces for sample management
 - Metadata submission via template upload, web form, or web services
 - IGSN minting
 - Sample landing pages
- IGSN: Persistent, globally unique identifier for samples
 - Web architecture for sample discovery & access
 - International samples community

Community Need for Sample Tracking



Challenge

Lack of a practical, standardized sample tracking system



Solution

International Geo/General Sample Numbers (IGSNs)
Standardized core sample metadata and templates



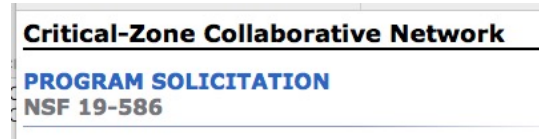
Credit: Joan Damerow et al. (2021): "Sample Identifiers and Metadata to Support Data Management and Reuse in Multidisciplinary Ecosystem Sciences" *Data Science Journal*, 20(1). DOI: <http://doi.org/10.5334/dsj-2021-011>



www.igsn.org
www.geosamples.org



Expectations: The Role of the Clusters



- Share cluster data.
- Ensure compatibility of data & sample collection and management with Network-wide protocols.
- Follow best practices for data citation to give credit to your colleagues.
- Collaborate with the Hub
 - Share Data Management Plans
 - Contribute to establishing the network-wide protocols, e.g. share protocols and use cases, review proposed protocols as they evolve.
 - Provide ongoing feedback about functionality, usability, and utility 👍👎
 - Align Cluster's data & samples acquisition and data management workflows with agreed upon Best Practices & Standards.

Contact Information

- CZHub CI contacts:
 - Jeff Horsburgh – jeff.Horsburgh@usu.edu
 - Kerstin Lehnert – lehnert@Ideo.columbia.edu
 - Chris Calloway – cbc@renci.org

