Meeting the Needs of Interdisciplinary Critical Zone Scientists by Leveraging and Linking Existing Domain Repositories

Jeffery Horsburgh[1], Kerstin Lehnert[2], Christopher Calloway[3], and Jerad Bales[4]

[1]Utah State University
[2]Columbia University
[3]University of North Carolina at Chapel Hill
[4]Consortium of Universities for the Advancement of Hydrologic Science, Inc.

November 21, 2022

Abstract

Critical Zone (CZ) scientists study the system of coupled chemical, biological, physical, and geological processes operating together across all scales to support life at the Earth's surface (Brantley et al., 2007). In 2020, the U.S. National Science Foundation funded a new network of Thematic Cluster projects who are working collaboratively to answer scientific questions related to effects of urbanization on CZ processes; CZ function in semi-arid landscapes and the role of dust in sustaining these ecosystems; processes in deep bedrock and their relationship to CZ evolution; recovery of the CZ from disturbances such as fire and flooding; and changes in the coastal CZ related to rising sea level. Given the diversity of data being collected by these projects, supporting data collection, access, and archival for the larger network presents significant challenges. Leveraging existing repositories and cyberinfrastructure provides many benefits, but still poses the questions of which repositories to use and how to enable discovery of and access to data that may be deposited across different repositories. This presentation describes new cyberinfrastructure development that leverages existing, domain-specific data repositories to enable managing, curating, disseminating, and preserving data from the new network of CZ Thematic Cluster projects. A distributed architecture is under development that links existing data facilities and services, including HydroShare, EarthChem, SESAR, and eventually other systems as needed, via a CZ Hub that provides tools for simplified data submission, discovery and access, and links to computational resources for data analysis and visualization in support of CZ synthesis efforts. Our goal is to make data, samples, and software collected by the Thematic Cluster projects Findable, Accessible, Interoperable, and Reusable (FAIR), using existing domain-specific repositories. This collaboration among repositories to deliver integrated data services for an interdisciplinary science program may provide a template for future development of integrated, interdisciplinary data services. Brantley, S.L., M.B. Goldhaber, V. Ragnarsdottir (2007). Crossing disciplines and scales to understand the Critical Zone. Elements 3, 307-314, doi:10.2113/gselements.3.5.307.

Meeting the Needs of Interdisciplinary Critical Zone Scientists by Leveraging and Linking Existing Domain Repositories

Jeffery S. Horsburgh

Utah State University

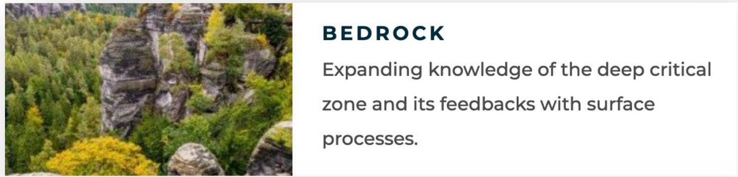
Kerstin Lehnert, Chris Calloway, Jerad Bales

CZNet

Critical Zone Network

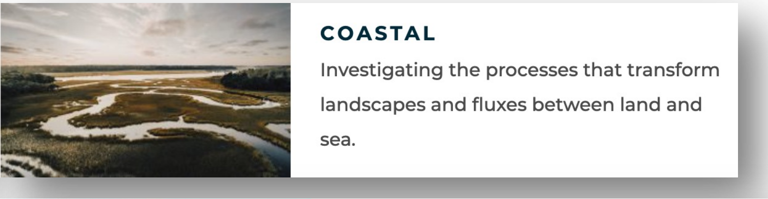
Critical Zone Collaborative Network

- In 2020 NSF funded the next phase of their Critical Zone research program
- Nine Thematic Cluster study areas with a wide range of geological, climatic, and land use settings working to better understand the evolution and function of the Critical Zone
- One Coordinating Hub to help coordinate activities across Clusters – including data management



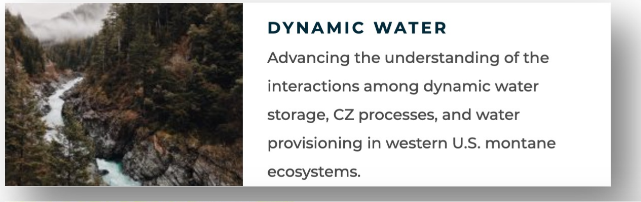
BEDROCK

Expanding knowledge of the deep critical zone and its feedbacks with surface processes.



COASTAL

Investigating the processes that transform landscapes and fluxes between land and sea.

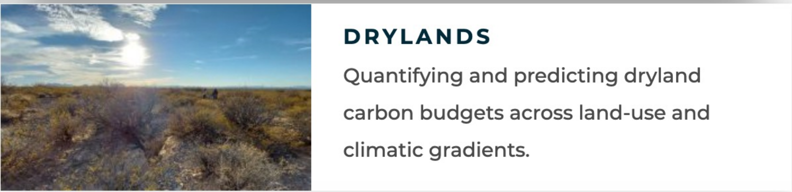


DYNAMIC WATER

Advancing the understanding of the interactions among dynamic water storage, CZ processes, and water provisioning in western U.S. montane ecosystems.

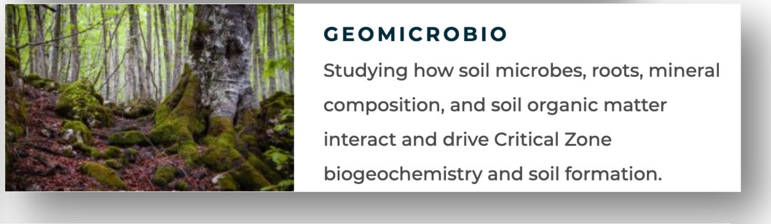
BIG DATA

Using field observations, existing data, & advanced statistical and process-based tools to investigate how the Critical Zone responds to disturbances.



DRYLANDS

Quantifying and predicting dryland carbon budgets across land-use and climatic gradients.

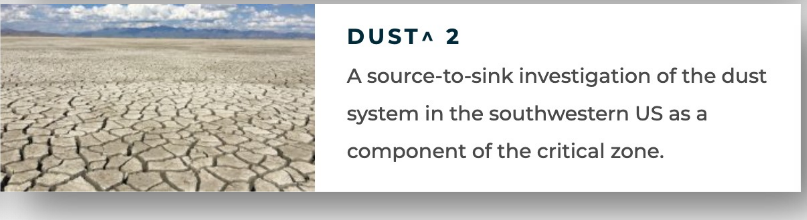


GEOMICROBIO

Studying how soil microbes, roots, mineral composition, and soil organic matter interact and drive Critical Zone biogeochemistry and soil formation.

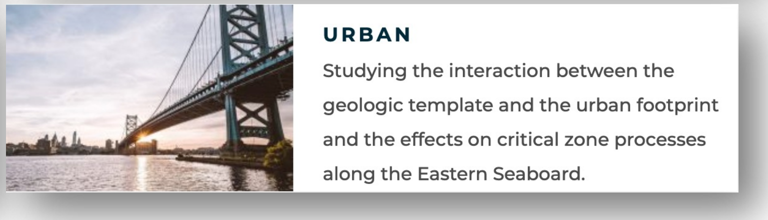
CINET

Investigating the role of critical interfaces for regulating the storage & transport of material such as water, sediment, carbon, & nutrients.



DUST^2

A source-to-sink investigation of the dust system in the southwestern US as a component of the critical zone.

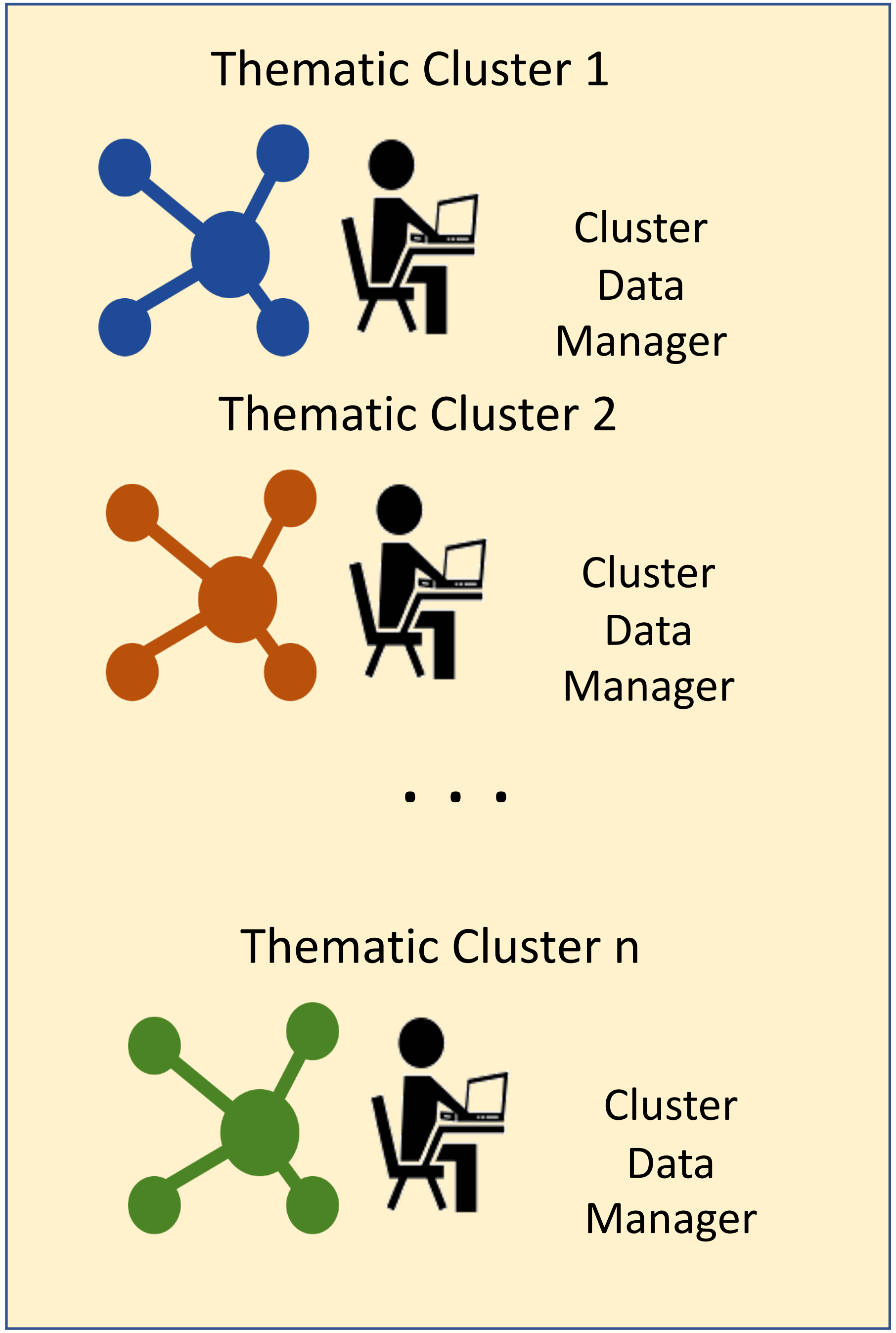


URBAN

Studying the interaction between the geologic template and the urban footprint and the effects on critical zone processes along the Eastern Seaboard.

Thematic Cluster Projects

- Activities
  - Data collection
  - Data aggregation
  - Data QA/QC
  - Defining data products
  - Metadata creation
- Data Management Plans
  - Each Thematic Cluster submitted their own Data Management Plan
  - Flexibility for local data management



Thematic Cluster 1

Cluster Data Manager

Thematic Cluster 2

Cluster Data Manager

. . .

Thematic Cluster n

Cluster Data Manager

Challenges

- Thematic Cluster teams and data are diverse
- Some are collecting new data, others are aggregating existing data, some are doing both
- No single data repository will meet the needs of interdisciplinary Critical Zone Scientists

Thematic Clusters	
1	Bedrock
2	Coastal
3	Dynamic Water
4	Big Data
5	Drylands
6	Geomicrobio
7	CINet
8	Dust^2
9	Urban

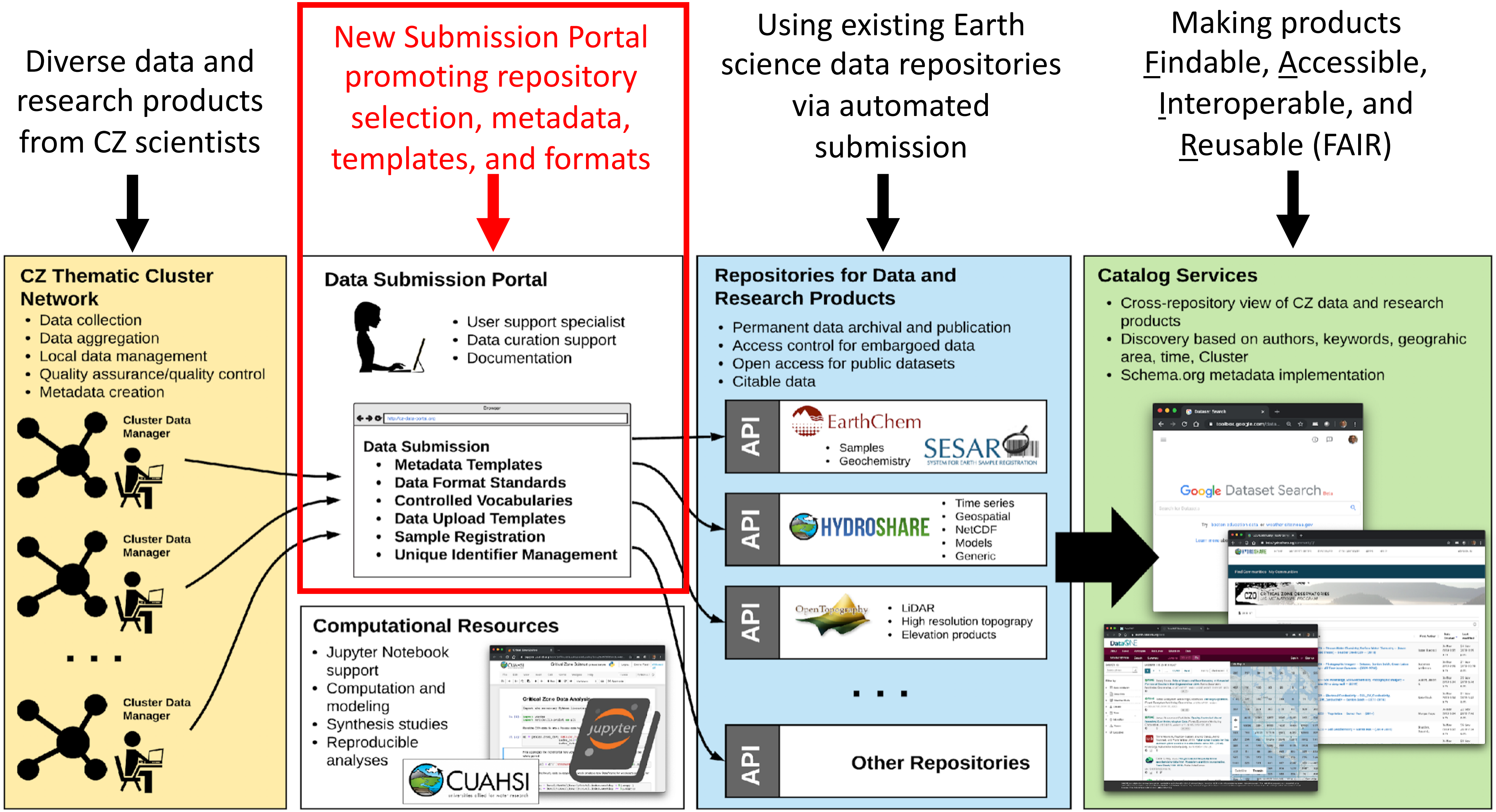
CZ Hub Objective: Provide a robust cyberinfrastructure for **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR) data from the CZ Net Thematic Clusters

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.

CZ Hub 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

CZ Hub Approach



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)

CZ Thematic Cluster Network

- Data collection
- Data aggregation
- Local data management
- Quality assurance/quality control
- Metadata creation

Data Submission Portal

- User support specialist
- Data curation support
- Documentation

Data Submission

- **Metadata Templates**
- **Data Format Standards**
- **Controlled Vocabularies**
- **Data Upload Templates**
- **Sample Registration**
- **Unique Identifier Management**

Computational Resources

- Jupyter Notebook support
- Computation and modeling
- Synthesis studies
- Reproducible analyses

Repositories for Data and Research Products

- Permanent data archival and publication
- Access control for embargoed data
- Open access for public datasets
- Citable data

- Samples
- Geochemistry

- Time series
- Geospatial
- NetCDF
- Models
- Generic

- LiDAR
- High resolution topography
- Elevation products

Other Repositories

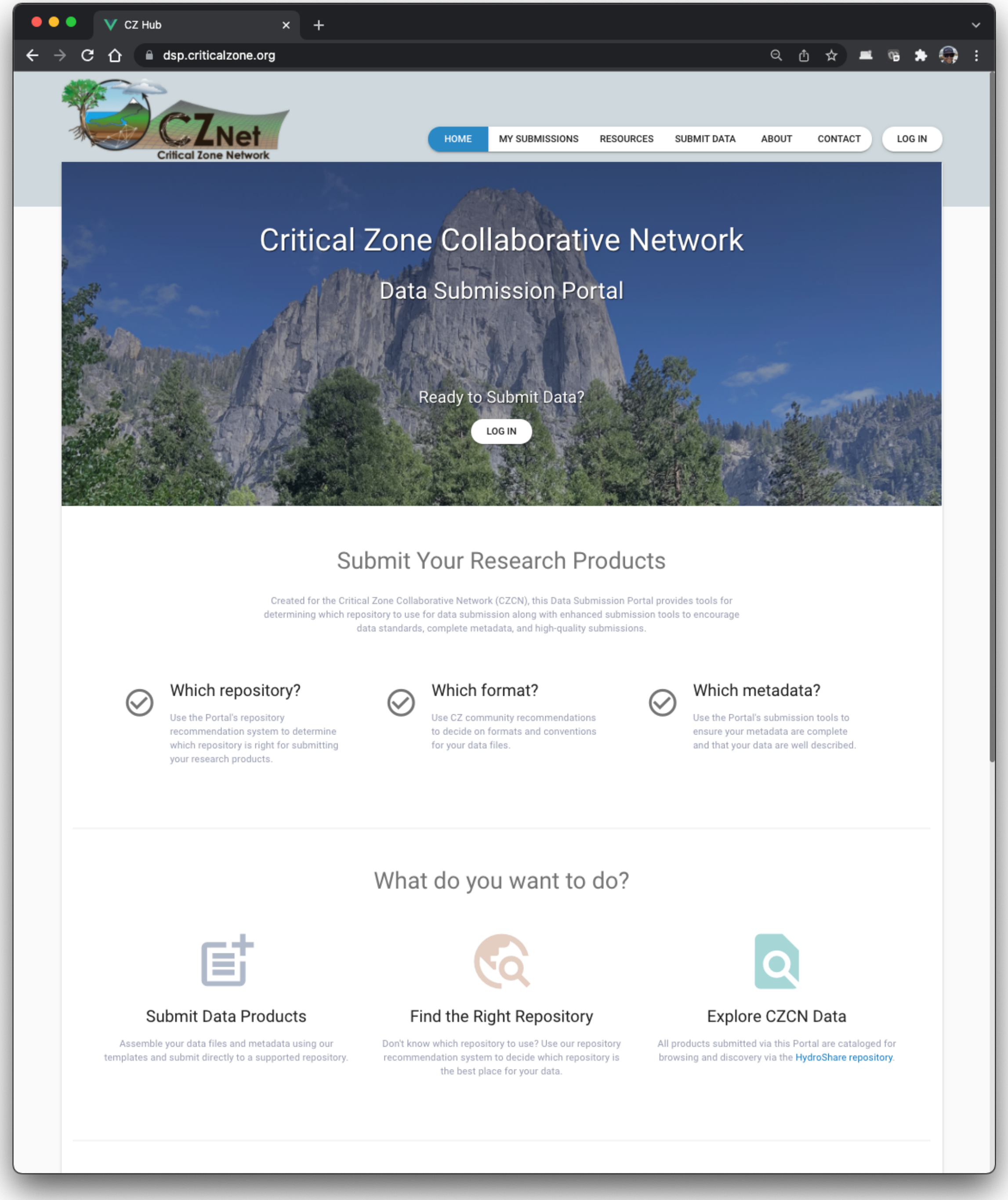
Catalog Services

- Cross-repository view of CZ data and research products
- Discovery based on authors, keywords, geographic area, time, Cluster
- Schema.org metadata implementation

Data Submission Portal

- New, web application to support CZ Net
- Enables submission to multiple geoscience data repositories through one portal
- Getting data to the right repository
- Submission directly through portal

Empower data managers to curate research products within appropriate repositories with support from our team



Critical Zone Collaborative Network

Data Submission Portal

Ready to Submit Data?

Submit Your Research Products

Created for the Critical Zone Collaborative Network (CZCN), this Data Submission Portal provides tools for determining which repository to use for data submission along with enhanced submission tools to encourage data standards, complete metadata, and high-quality submissions.

✓ **Which repository?** Use the Portal's repository recommendation system to determine which repository is right for submitting your research products.

✓ **Which format?** Use CZ community recommendations to decide on formats and conventions for your data files.

✓ **Which metadata?** Use the Portal's submission tools to ensure your metadata are complete and that your data are well described.

What do you want to do?

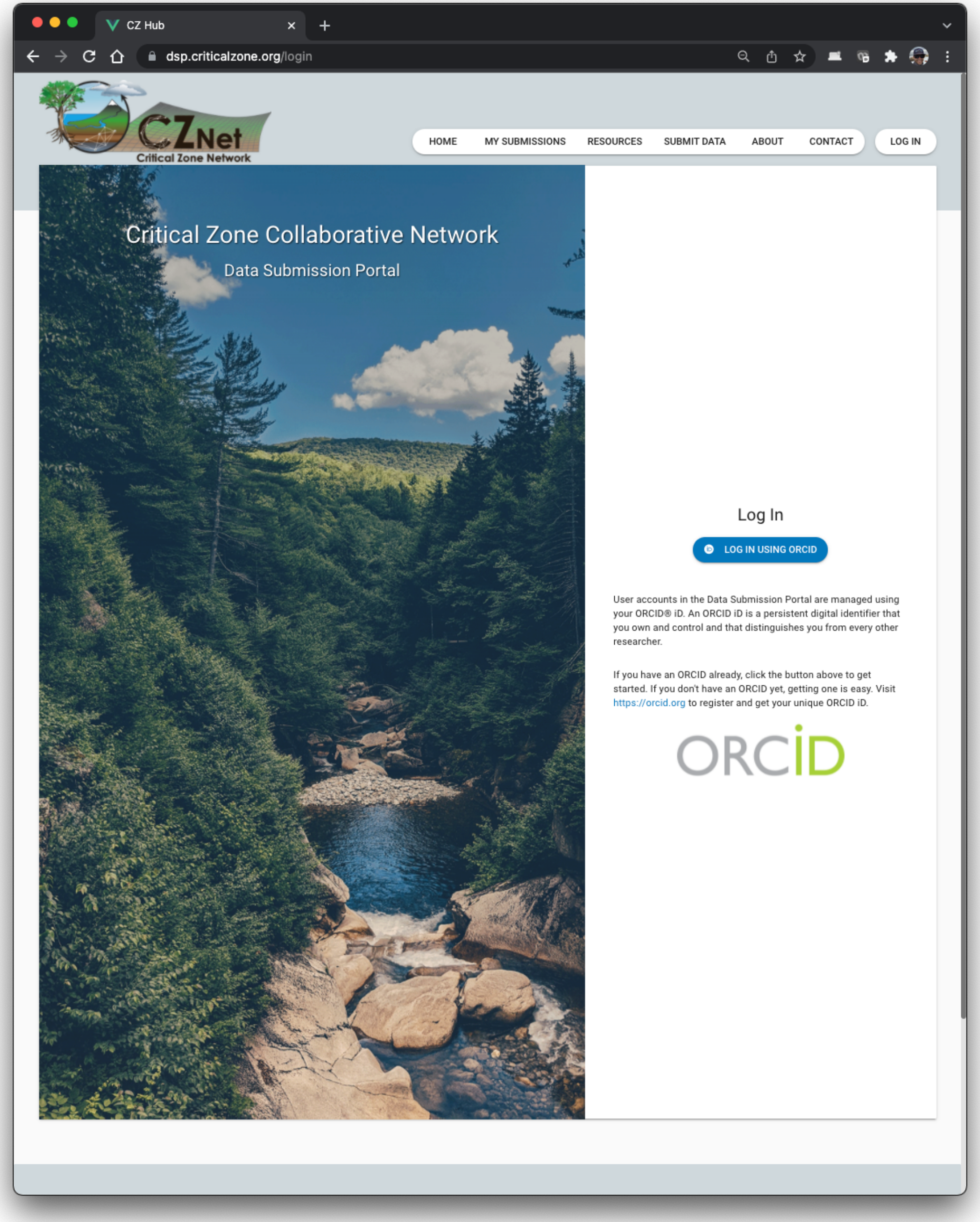
Submit Data Products
Assemble your data files and metadata using our templates and submit directly to a supported repository.

Find the Right Repository
Don't know which repository to use? Use our repository recommendation system to decide which repository is the best place for your data.

Explore CZCN Data
All products submitted via this Portal are cataloged for browsing and discovery via the HydroShare repository.

User Account Management

- Authentication and user management via ORCID
- Using an account that most people have already
- Enables authentication across multiple repositories



Critical Zone Collaborative Network

Data Submission Portal

Log In

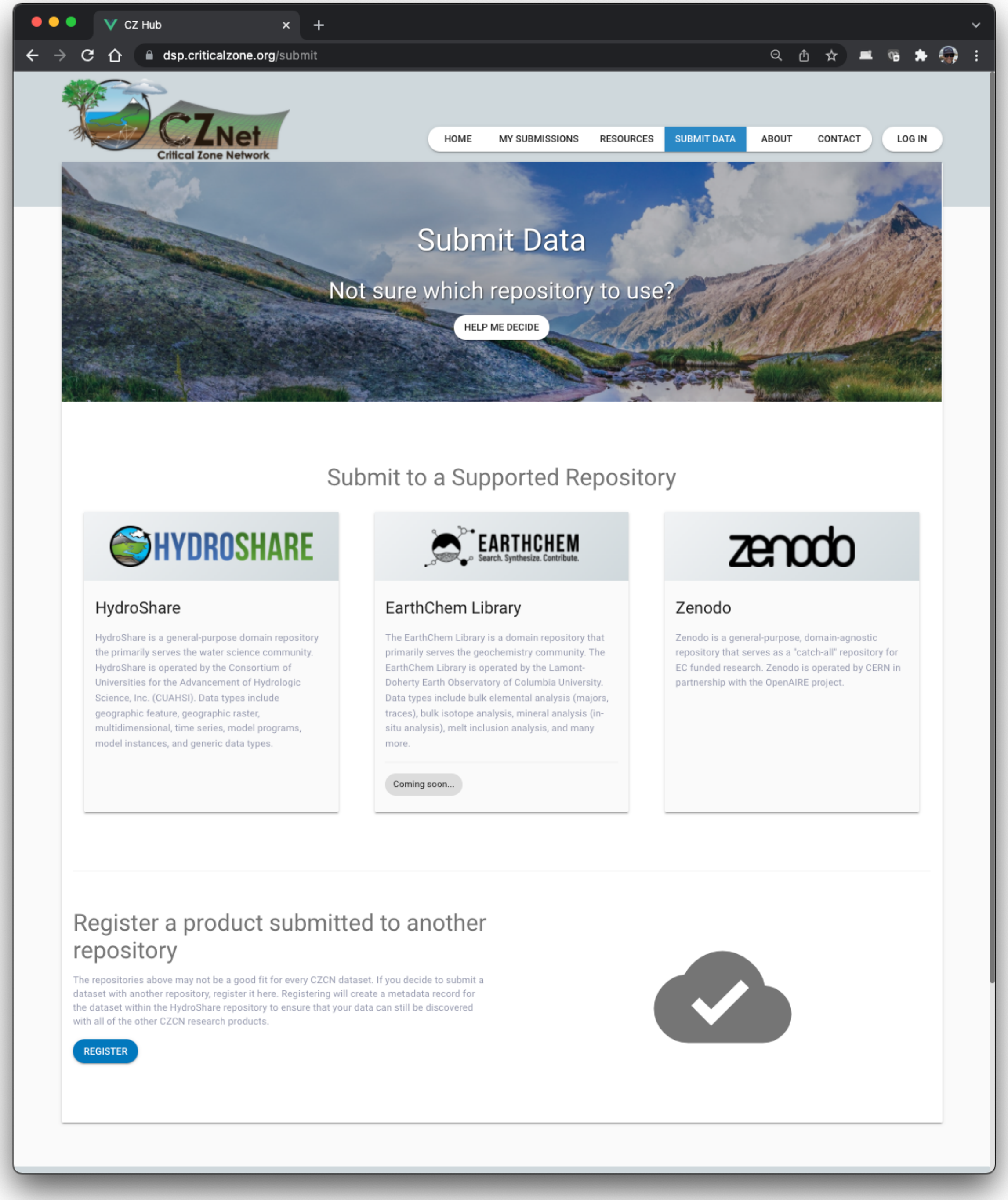
LOG IN USING ORCID

User accounts in the Data Submission Portal are managed using your ORCID® iD. An ORCID iD is a persistent digital identifier that you own and control and that distinguishes you from every other researcher.

If you have an ORCID already, click the button above to get started. If you don't have an ORCID yet, getting one is easy. Visit https://orcid.org to register and get your unique ORCID iD.

Supported Repositories

- 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



Submit Data

Not sure which repository to use?

Submit to a Supported Repository

HydroShare

HydroShare is a general-purpose domain repository the primarily serves the water science community. HydroShare is operated by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI). Data types include geographic feature, geographic raster, multidimensional, time series, model programs, model instances, and generic data types.

EarthChem Library

The EarthChem Library is a domain repository that primarily serves the geochemistry community. The EarthChem Library is operated by the Lamont-Doherty Earth Observatory of Columbia University. Data types include bulk elemental analysis (majors, traces), bulk isotope analysis, mineral analysis (in-situ analysis), melt inclusion analysis, and many more.

Zenodo

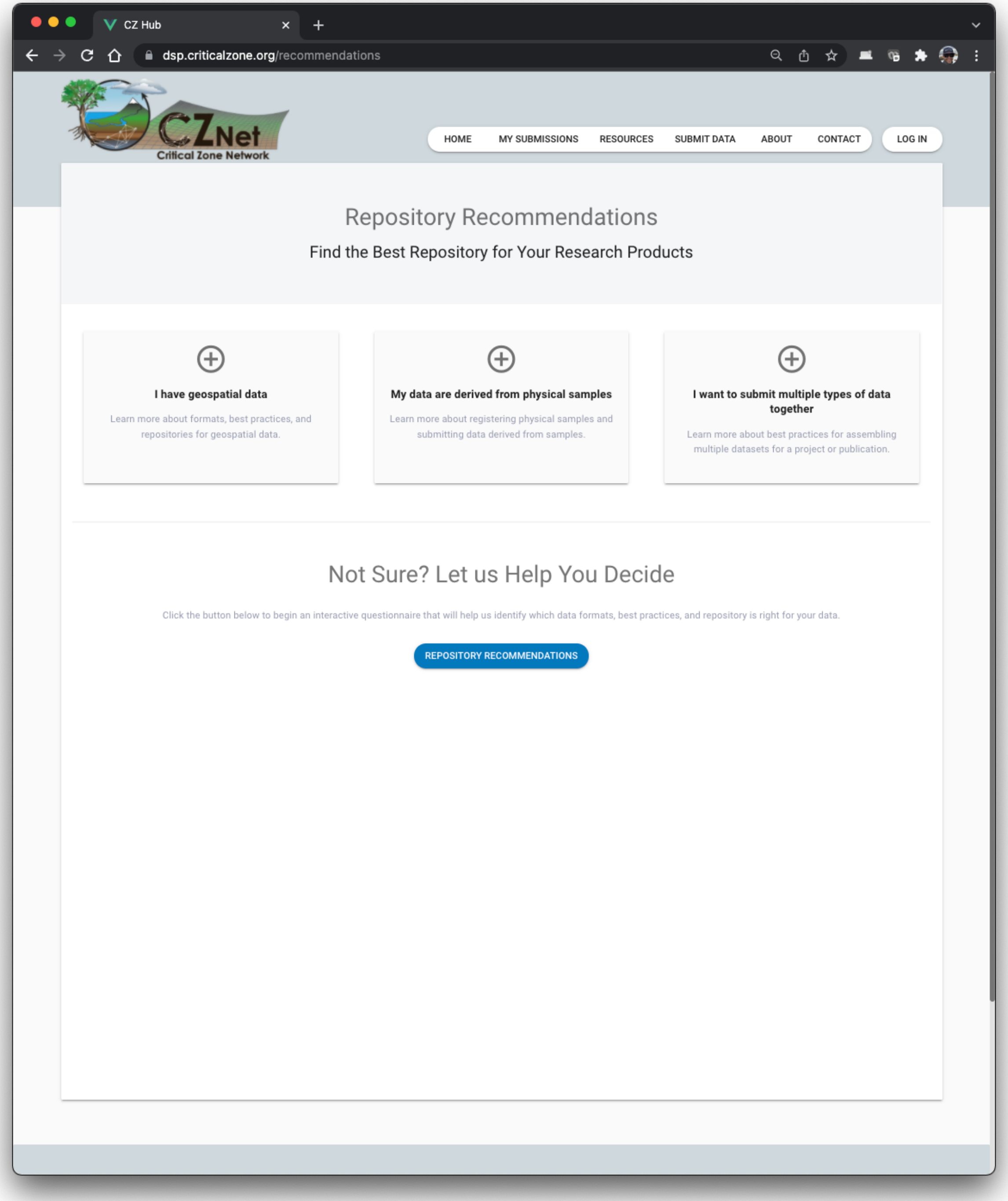
Zenodo is a general-purpose, domain-agnostic repository that serves as a "catch-all" repository for EC funded research. Zenodo is operated by CERN in partnership with the OpenAIRE project.

Register a product submitted to another repository

The repositories above may not be a good fit for every CZCN dataset. If you decide to submit a dataset with another repository, register it here. Registering will create a metadata record for the dataset within the HydroShare repository to ensure that your data can still be discovered with all of the other CZCN research products.

Which repository?

- Which repositories to target for different data types?
- Assist data managers in selecting an appropriate repository
  - Geospatial data
  - Data derived from physical samples
  - Hydrologic time series
  - Data and code packages
  - Models



Repository Recommendations

Find the Best Repository for Your Research Products

I have geospatial data

Learn more about formats, best practices, and repositories for geospatial data.

My data are derived from physical samples

Learn more about registering physical samples and submitting data derived from samples.

I want to submit multiple types of data together

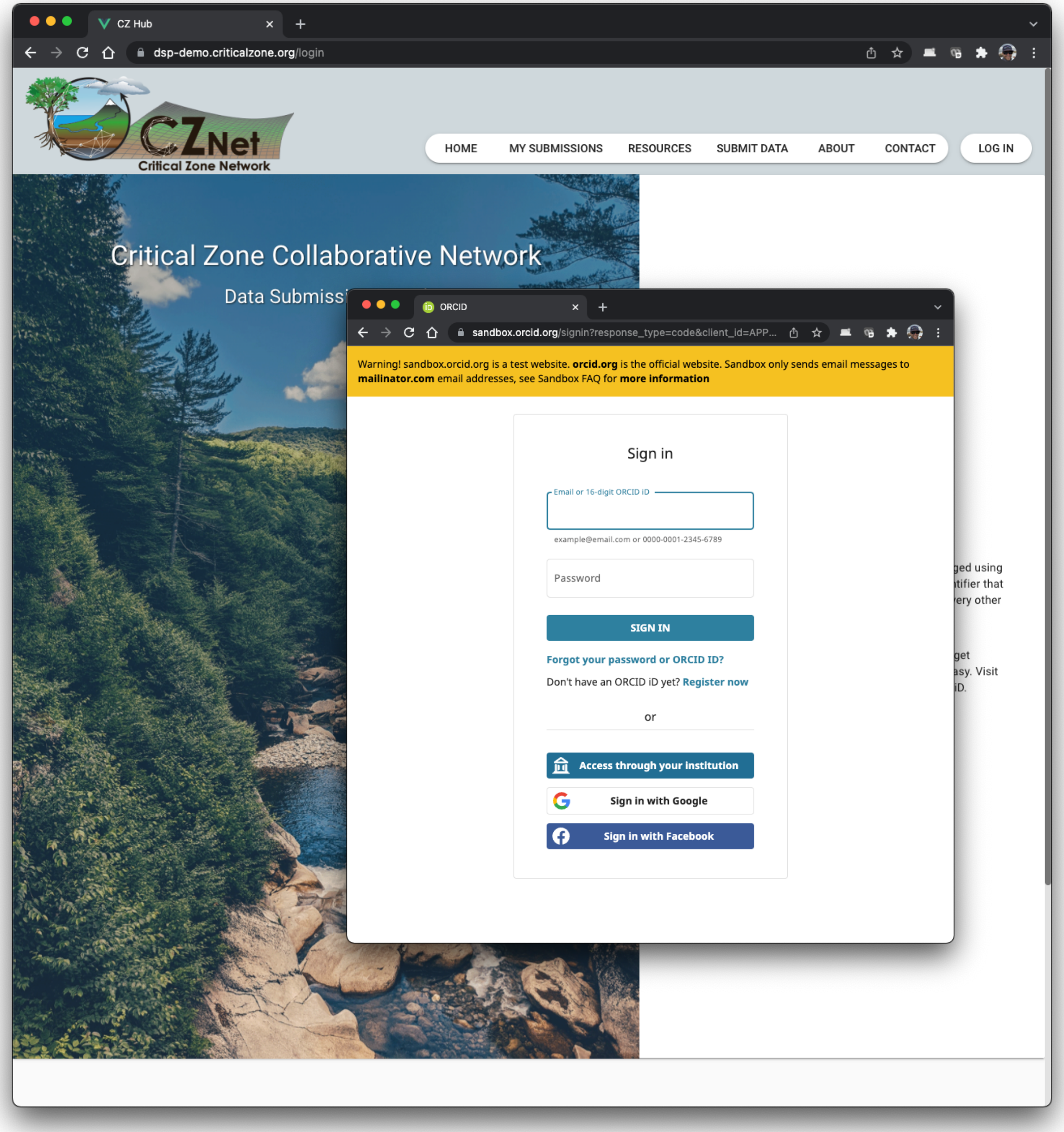
Learn more about best practices for assembling multiple datasets for a project or publication.

Not Sure? Let us Help You Decide

Click the button below to begin an interactive questionnaire that will help us identify which data formats, best practices, and repository is right for your data.

Step1: Log in

- User logs into the Portal using their ORCID
- User accounts in the Portal are associated with the ORCID

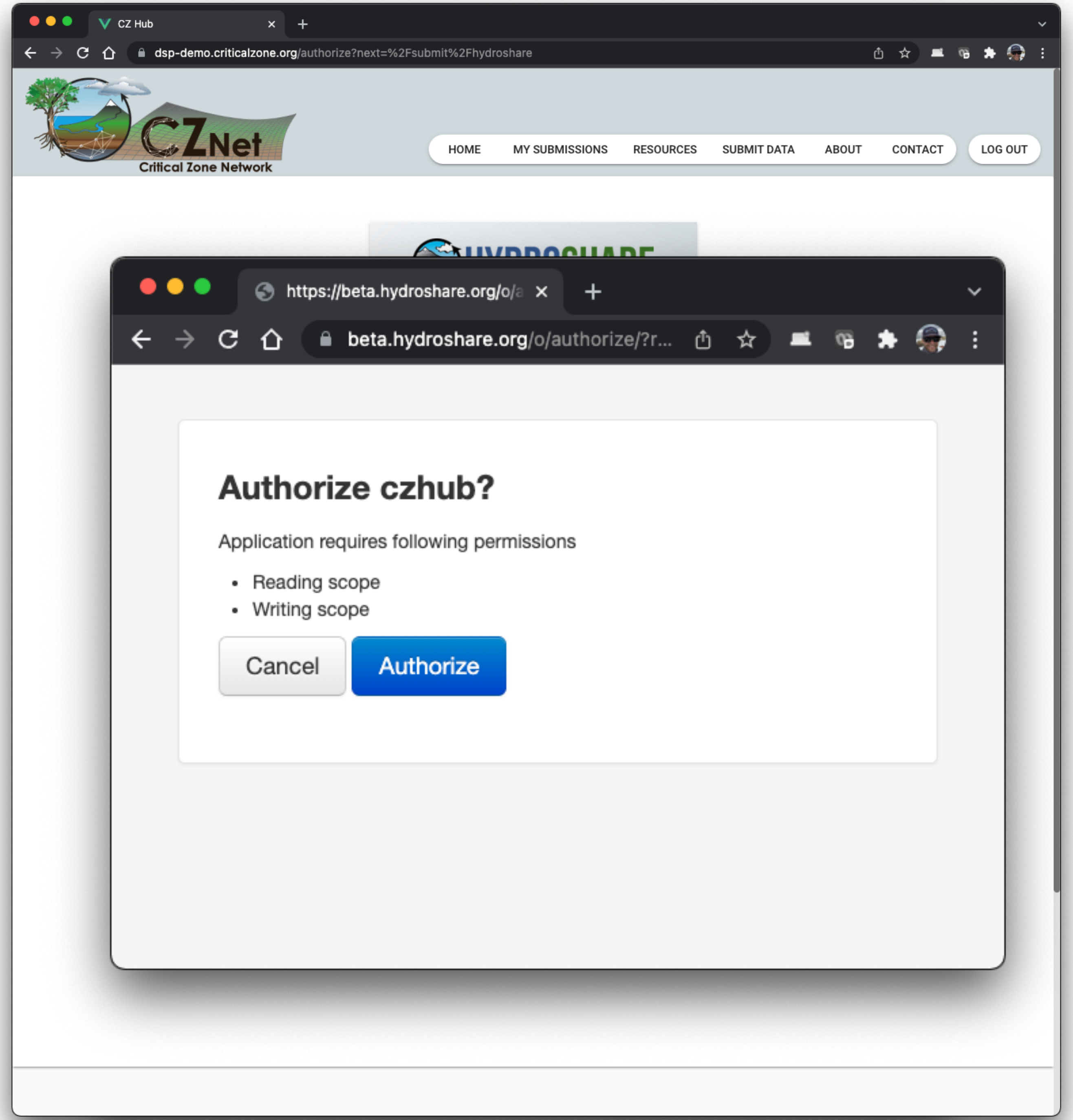


Warning! sandbox.orcid.org is a test website. **orcid.org** is the official website. Sandbox only sends email messages to **mailinator.com** email addresses, see Sandbox FAQ for **more information**

CZ Hub
dsp-demo.criticalzone.org/login
CZNet
Critical Zone Network
HOME
MY SUBMISSIONS
RESOURCES
SUBMIT DATA
ABOUT
CONTACT
LOG IN
Critical Zone Collaborative Network
Data Submissi
ORCID
sandbox.orcid.org/signin?response_type=code&client_id=APP...
Sign in
Email or 16-digit ORCID iD
example@email.com or 0000-0001-2345-6789
Password
SIGN IN
Forgot your password or ORCID ID?
Don't have an ORCID iD yet? Register now
or
Access through your institution
Sign in with Google
Sign in with Facebook
ged using
tifier that
ery other
get
asy. Visit
iD.

Step 2: Choose Repo

- User chooses a repository to submit to from the Submit Data page
- The user authorizes the Portal to submit to HydroShare
- The one-time authorization is stored in the user's profile


CZ Hub
dsp-demo.criticalzone.org/authorize?next=%2Fsubmit%2Fhydroshare
CZNet
Critical Zone Network
HOME
MY SUBMISSIONS
RESOURCES
SUBMIT DATA
ABOUT
CONTACT
LOG OUT
https://beta.hydroshare.org/o/a
beta.hydroshare.org/o/authorize/?r...

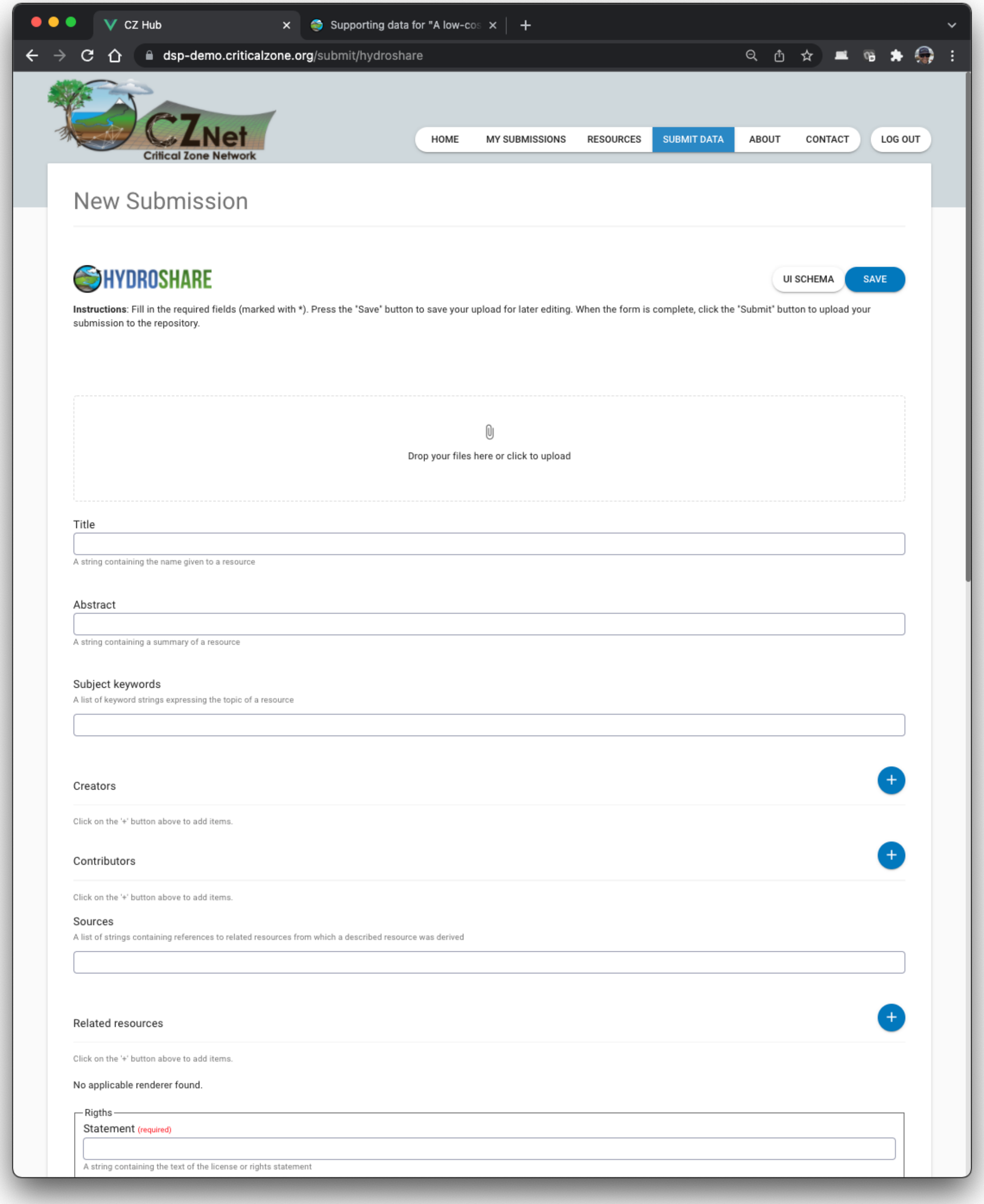
Authorize czhub?

Application requires following permissions

- Reading scope
- Writing scope

Step 3: Create Content

- User enters metadata and selects content files on the data submission form for the chosen repository
- Each repository has its own submission form
- Submission forms are built from a JSON schema that defines:
  - Required and optional metadata
  - Default values
  - Etc.



New Submission

Instructions: Fill in the required fields (marked with *). Press the "Save" button to save your upload for later editing. When the form is complete, click the "Submit" button to upload your submission to the repository.

Drop your files here or click to upload

Title

A string containing the name given to a resource

Abstract

A string containing a summary of a resource

Subject keywords

A list of keyword strings expressing the topic of a resource

Creators

Click on the '+' button above to add items.

Contributors

Click on the '+' button above to add items.

Sources

A list of strings containing references to related resources from which a described resource was derived

Related resources

Click on the '+' button above to add items.

No applicable renderer found.

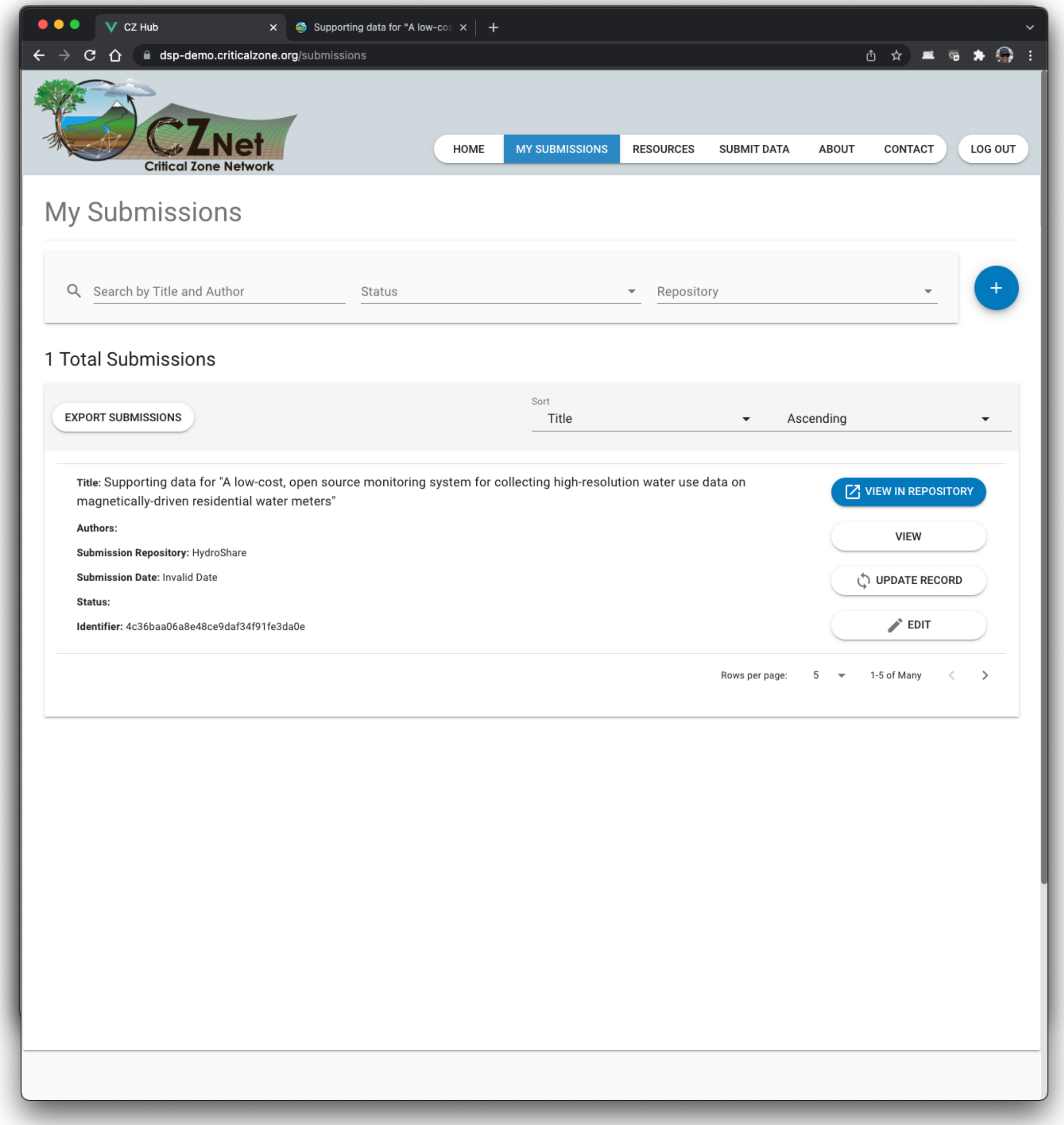
Rigths

Statement (required)

A string containing the text of the license or rights statement

Step 4: Submit Content

- Metadata and data files are sent to the repository
- A new resource is created in the repository
- A new record is created in the user's My Submissions page
  - Return later to Edit
  - Export submissions to file



My Submissions

1 Total Submissions

Sort: Title, Ascending

Title: Supporting data for "A low-cost, open source monitoring system for collecting high-resolution water use data on magnetically-driven residential water meters"

Authors:

Submission Repository: HydroShare

Submission Date: Invalid Date

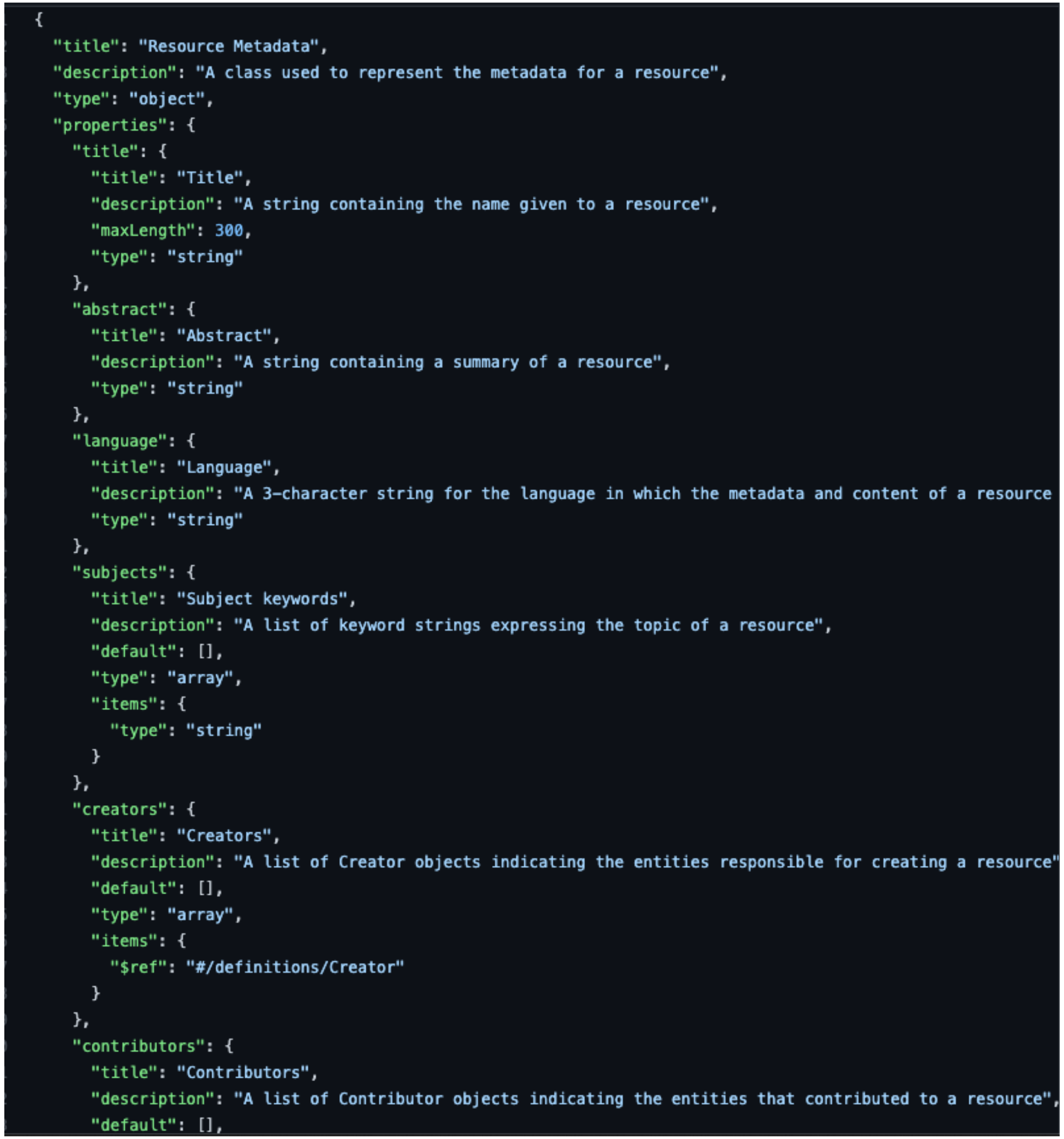
Status:

Identifier: 4c36baa06a8e48ce9daf34f91fe3da0e

Rows per page: 5 1-5 of Many

JSON Schema-based Metadata

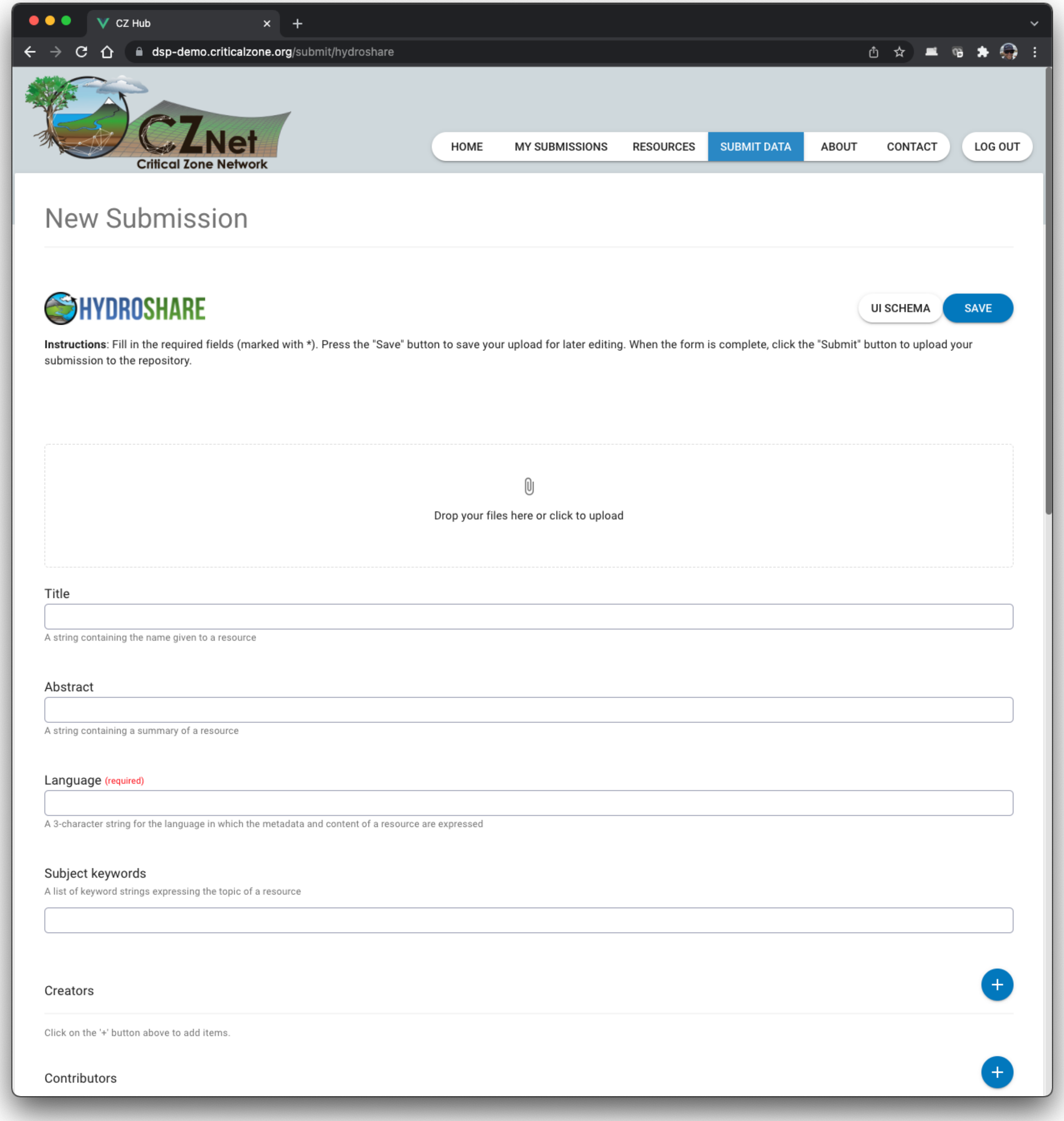
- A JSON schema defines required and optional metadata for each repository
- Submissions validated based on JSON schema
  - Data types
  - Default values
  - Required/optional
- Data submission form dynamically built from the JSON schema
- Adding a new repository to the Portal means adding a new JSON schema



```
{
  "title": "Resource Metadata",
  "description": "A class used to represent the metadata for a resource",
  "type": "object",
  "properties": {
    "title": {
      "title": "Title",
      "description": "A string containing the name given to a resource",
      "maxLength": 300,
      "type": "string"
    },
    "abstract": {
      "title": "Abstract",
      "description": "A string containing a summary of a resource",
      "type": "string"
    },
    "language": {
      "title": "Language",
      "description": "A 3-character string for the language in which the metadata and content of a resource
      "type": "string"
    },
    "subjects": {
      "title": "Subject keywords",
      "description": "A list of keyword strings expressing the topic of a resource",
      "default": [],
      "type": "array",
      "items": {
        "type": "string"
      }
    },
    "creators": {
      "title": "Creators",
      "description": "A list of Creator objects indicating the entities responsible for creating a resource"
      "default": [],
      "type": "array",
      "items": {
        "$ref": "#/definitions/Creator"
      }
    },
    "contributors": {
      "title": "Contributors",
      "description": "A list of Contributor objects indicating the entities that contributed to a resource",
      "default": [],
```

JSON Schema-based Metadata

- Data submission form dynamically built from the JSON schema
- Files and metadata sent directly to repository
- Data Submission Portal maintains a record of submission



New Submission

Instructions: Fill in the required fields (marked with *). Press the "Save" button to save your upload for later editing. When the form is complete, click the "Submit" button to upload your submission to the repository.

Drop your files here or click to upload

Title

A string containing the name given to a resource

Abstract

A string containing a summary of a resource

Language (required)

A 3-character string for the language in which the metadata and content of a resource are expressed

Subject keywords

A list of keyword strings expressing the topic of a resource

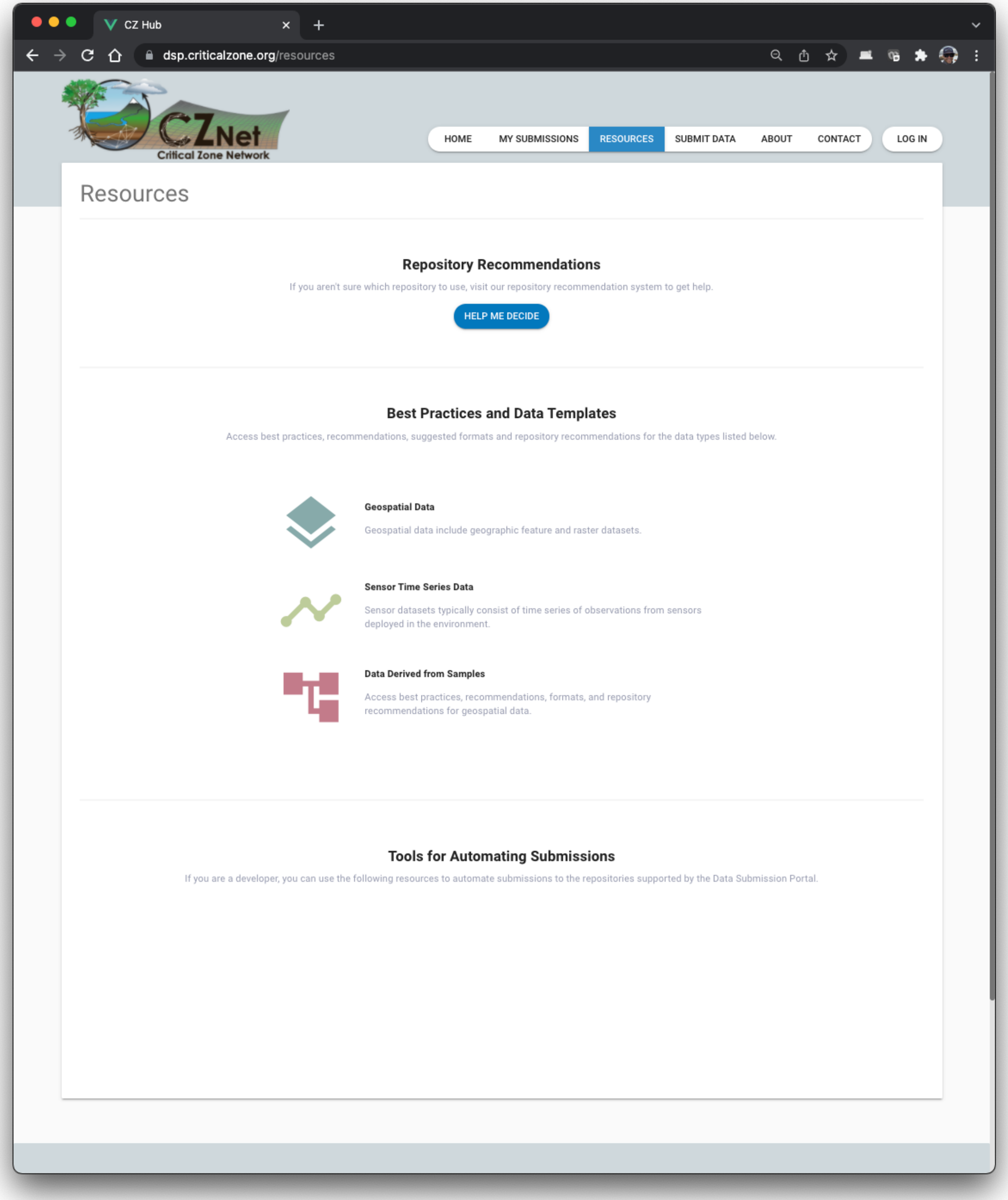
Creators

Click on the '+' button above to add items.

Contributors

Promoting best practices

- Repository functionality is not specific to a community of users
- CZ Net may want to use:
  - Community standard formats
  - Templates
  - Best practices
- We can promote those through the Data Submission Portal



Resources

Repository Recommendations

If you aren't sure which repository to use, visit our repository recommendation system to get help.

Best Practices and Data Templates

Access best practices, recommendations, suggested formats and repository recommendations for the data types listed below.

Geospatial Data

Geospatial data include geographic feature and raster datasets.

Sensor Time Series Data

Sensor datasets typically consist of time series of observations from sensors deployed in the environment.

Data Derived from Samples

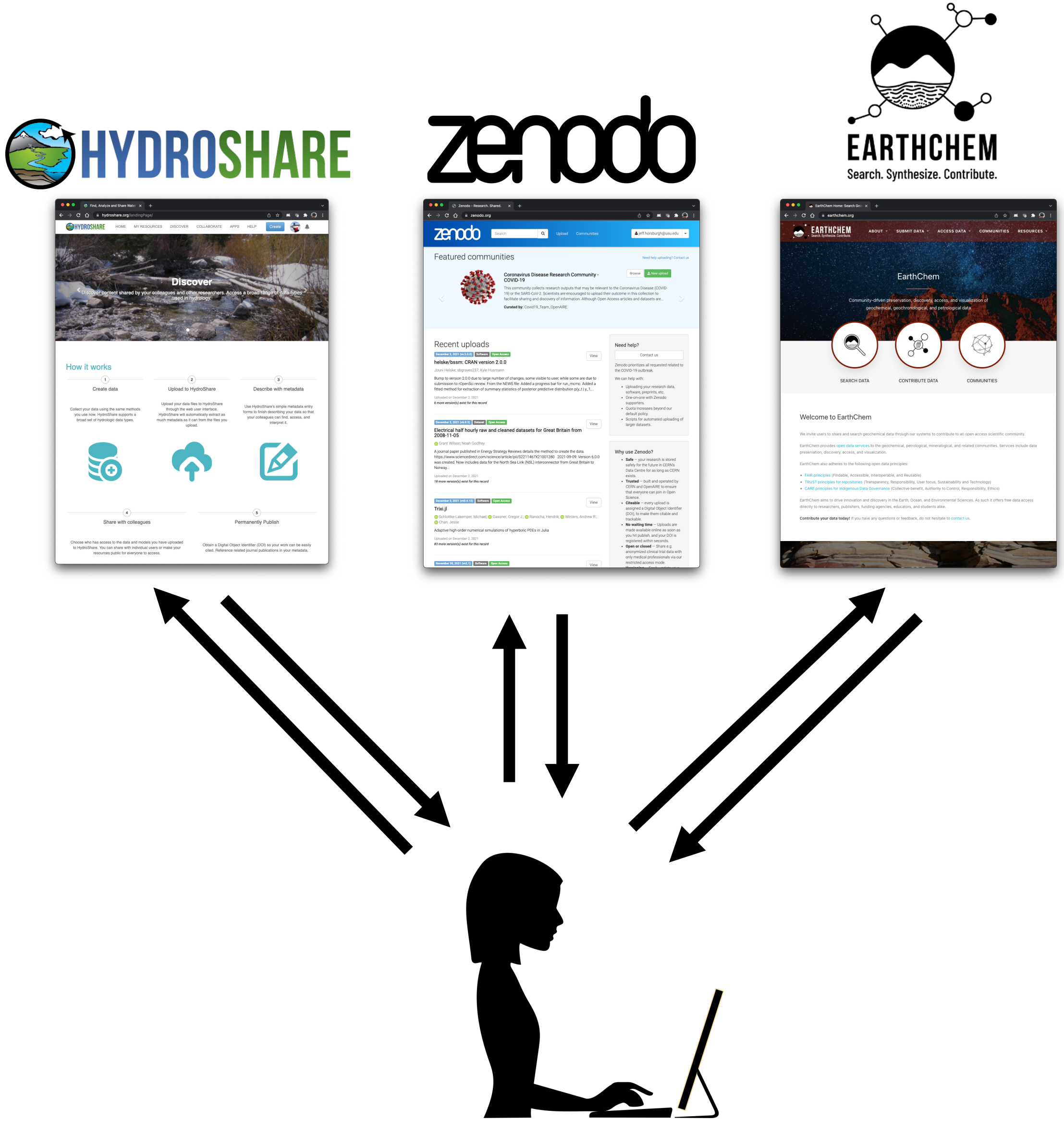
Access best practices, recommendations, formats, and repository recommendations for geospatial data.

Tools for Automating Submissions

If you are a developer, you can use the following resources to automate submissions to the repositories supported by the Data Submission Portal.

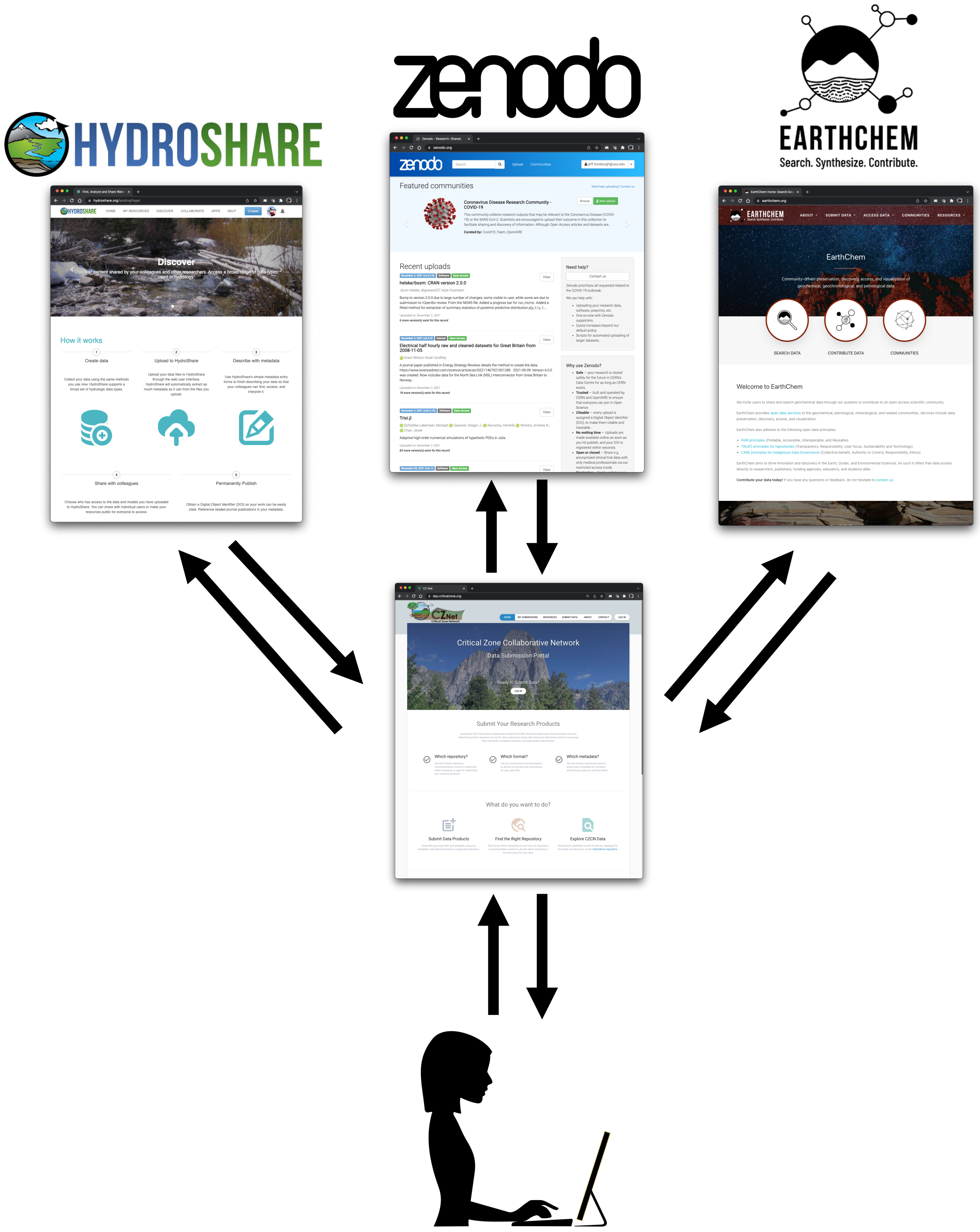
Without the DSP

- Data managers must navigate user interfaces of multiple systems
- Must keep track of what has been submitted to each one
- Difficult for CZ Hub Team to track what has been submitted



With the DSP

- Data managers need only interact through one user interface
- Submissions to all repositories tracked in one place
- Submissions automatically registered for cataloging/discovery



Data Submission Portal Advantages

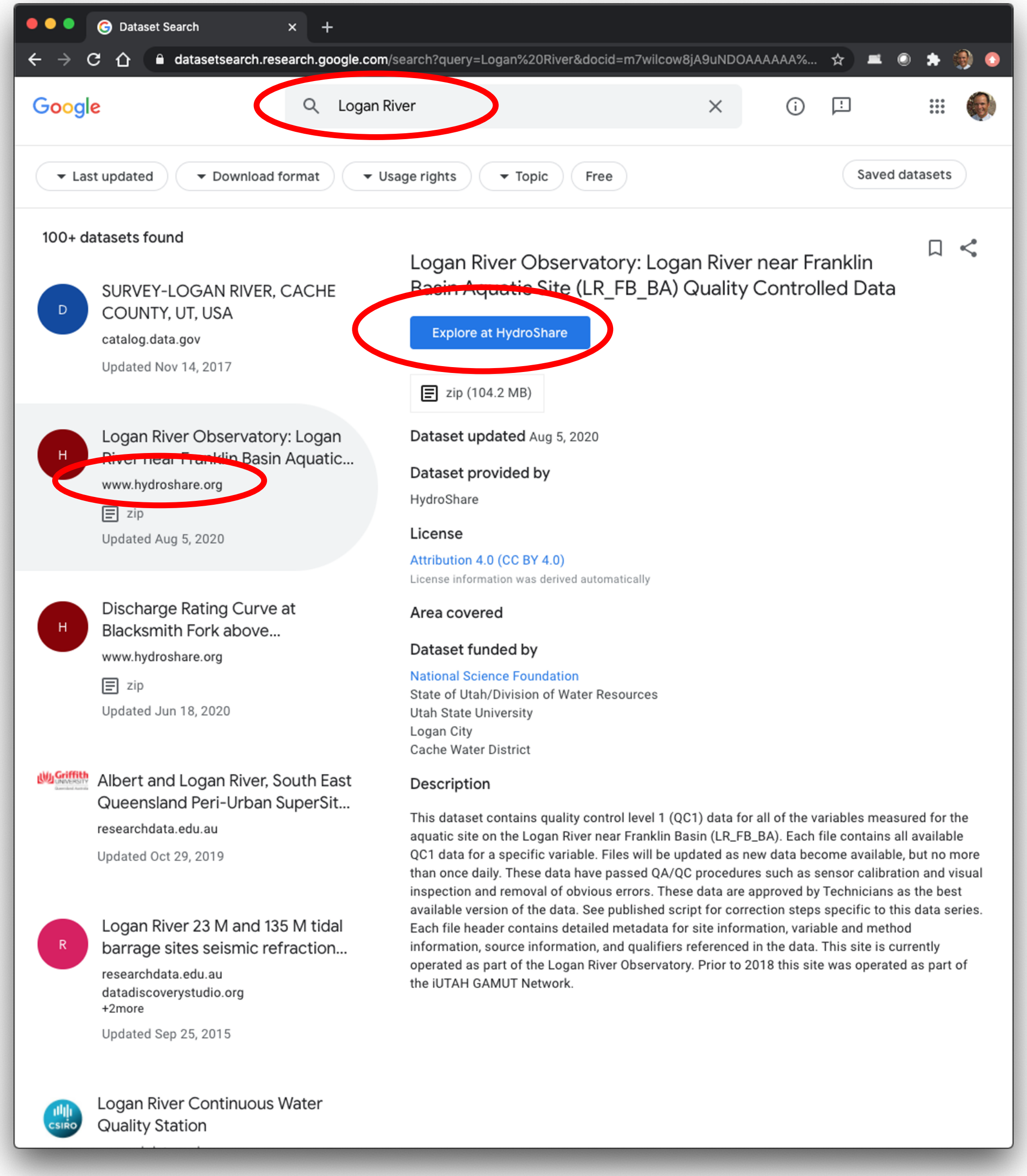
- Through validation, promote consistency in CZ Net data products across repositories
- Ensure data products end up in an appropriate, trusted repository
- Enforce minimum metadata requirements
  - Consistent keywords
  - Funding agency/grant information
- Enable use of controlled vocabularies where needed
- Promote templates, common formats, and best practices
- Enable data managers to use a single interface/tool to submit data
- Enable simple and consistent registration of CZ Net datasets with a metadata index for discovery
- Helping Thematic Clusters keep track of what has been submitted

CZ Net Catalog Services

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

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



Dataset Search

datasetsearch.research.google.com/search?query=Logan%20River&docid=m7wilcow8jA9uNDOAAAAAA%...

Google

Logan River

Last updated · Download format · Usage rights · Topic · Free · Saved datasets

100+ datasets found

D — SURVEY-LOGAN RIVER, CACHE COUNTY, UT, USA
catalog.data.gov
Updated Nov 14, 2017

H — Logan River Observatory: Logan River near Franklin Basin Aquatic...
www.hydroshare.org
zip
Updated Aug 5, 2020

H — Discharge Rating Curve at Blacksmith Fork above...
www.hydroshare.org
zip
Updated Jun 18, 2020

Albert and Logan River, South East Queensland Peri-Urban SuperSit...
researchdata.edu.au
Updated Oct 29, 2019

R — Logan River 23 M and 135 M tidal barrage sites seismic refraction...
researchdata.edu.au
datadiscoverystudio.org
+2more
Updated Sep 25, 2015

Logan River Continuous Water Quality Station

Logan River Observatory: Logan River near Franklin Basin Aquatic Site (LR_FB_BA) Quality Controlled Data

Explore at HydroShare

zip (104.2 MB)

Dataset updated Aug 5, 2020

Dataset provided by
HydroShare

License
Attribution 4.0 (CC BY 4.0)
License information was derived automatically

Area covered

Dataset funded by
National Science Foundation
State of Utah/Division of Water Resources
Utah State University
Logan City
Cache Water District

Description

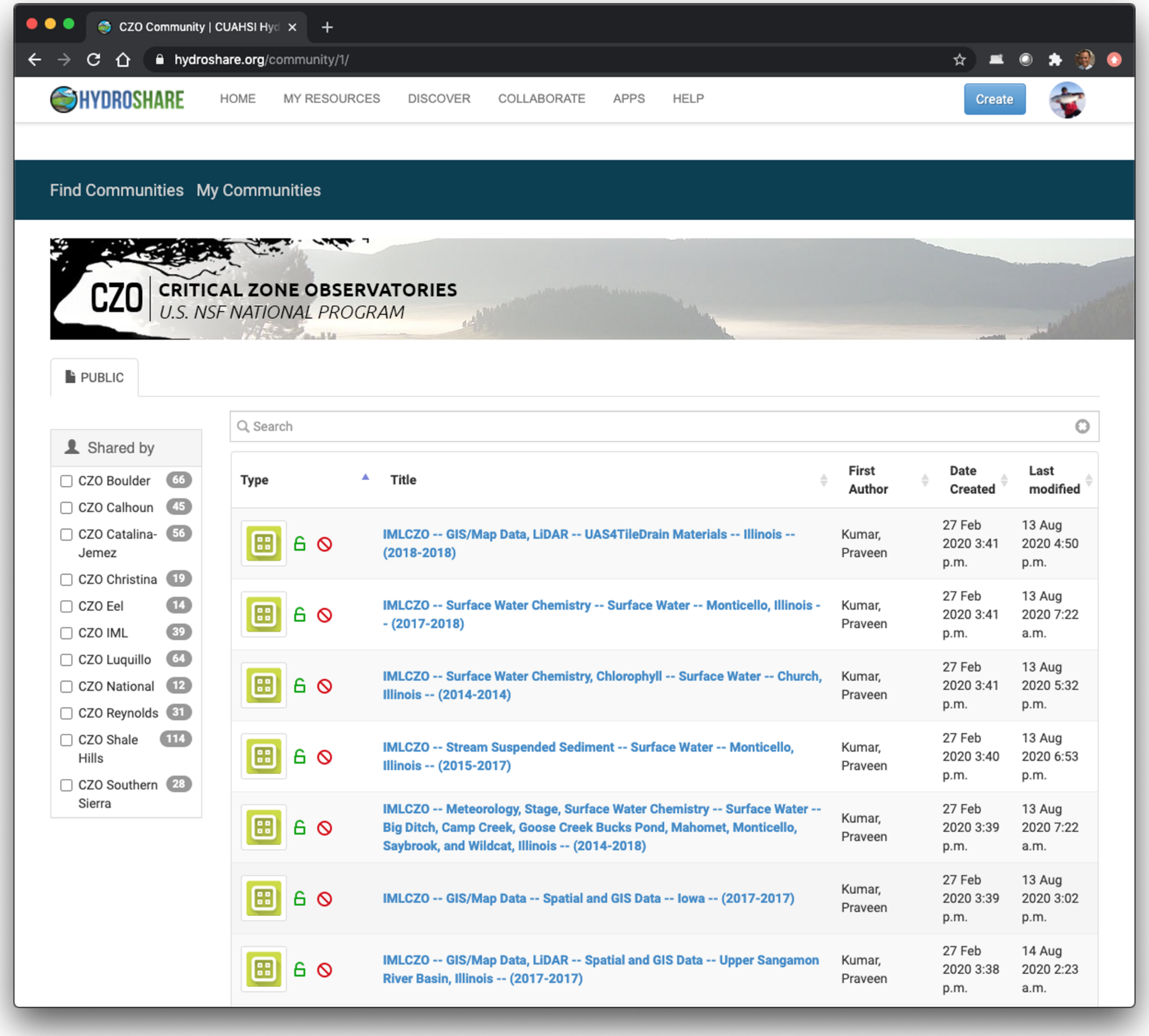
This dataset contains quality control level 1 (QC1) data for all of the variables measured for the aquatic site on the Logan River near Franklin Basin (LR_FB_BA). Each file contains all available QC1 data for a specific variable. Files will be updated as new data become available, but no more than once daily. These data have passed QA/QC procedures such as sensor calibration and visual inspection and removal of obvious errors. These data are approved by Technicians as the best available version of the data. See published script for correction steps specific to this data series. Each file header contains detailed metadata for site information, variable and method information, source information, and qualifiers referenced in the data. This site is currently operated as part of the Logan River Observatory. Prior to 2018 this site was operated as part of the iUTAH GAMUT Network.

CZ Net Catalog Services

- Cross-repository view of CZ Net 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


CZO Community | CUAHSI Hyd
hydroshare.org/community/1/
HYDROSHARE
HOME
MY RESOURCES
DISCOVER
COLLABORATE
APPS
HELP
Create
Find Communities
My Communities
CZO
CRITICAL ZONE OBSERVATORIES
U.S. NSF NATIONAL PROGRAM
PUBLIC
Search

Shared by

☐ CZO Boulder	66
☐ CZO Calhoun	45
☐ CZO Catalina-Jemez	56
☐ CZO Christina	19
☐ CZO Eel	14
☐ CZO IML	39
☐ CZO Luquillo	64
☐ CZO National	12
☐ CZO Reynolds	31
☐ CZO Shale Hills	114
☐ CZO Southern Sierra	28

Type	Title	First Author	Date Created	Last modified
	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 -- 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 -- 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 -- 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 -- 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 -- 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 -- 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.

GitHub

The Data Submission Portal is on GitHub
https://github.com/cznethub

CZNet
Critical Zone Network

Questions?

Jeffery S. Horsburgh

jeff.horsburgh@usu.edu

Support: 2012893, 2012748, 2012593