Supporting Data Sharing and Discovery for the Earth's Critical Zone through Cross-Repository Interoperability

Jeffery S. Horsburgh

Utah State University

The CZ Net Hub Team

Kerstin Lehnert, Jordan Read, Chris Calloway, Scott Black, Maurier Ramirez, Lucy Profeta, Clara Cogswell, Peng Ji, David Tarboton, Martin Seul

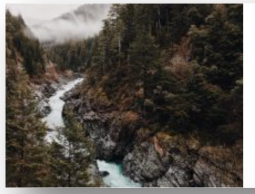
Support: 2012893, 2012748, 2012593

Critical Zone Collaborative Network

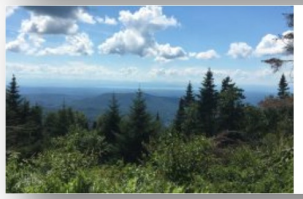
- In 2020 NSF funded a new phase of their Critical Zone research program
- Nine Thematic Cluster study areas with a wide range of geological, climatic, and land use settings working to better understand the evolution and function of the Critical Zone
- One Coordinating Hub to help coordinate activities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

Challenges

From NSF's solicitation:

*"The **Coordinating Hub** will: ensure the compatibility of the measurements across the various Clusters; lead the data management of the Network by establishing procedures for data collection, standardization, central archiving, and access by the research community"*

- 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

Then the clusters were funded

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 CZNet 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

Diverse data and research products from CZ scientists

Submission of products and/or registration of metadata

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

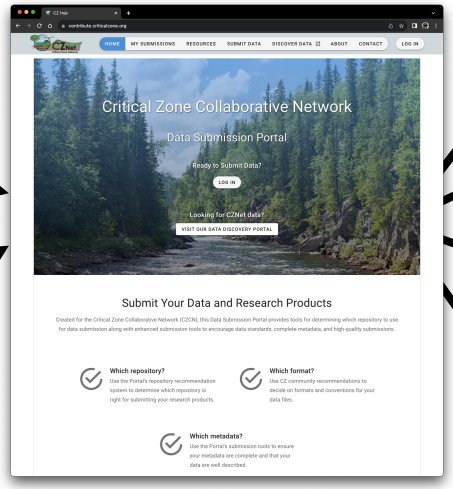
Cluster Data Manager 1

. . .

Cluster Data Manager n

Data Submission Portal

- Data/metadata submission
- Metadata templates
- Data format standards
- Data upload templates
- Unique identifiers



Repositories for Data and Research Products

- Permanent data archival and publication
- Access control for embargoed data
- Citable data

API EARTHCHEM Search. Synthesize. Contribute.

API HYDROSHARE

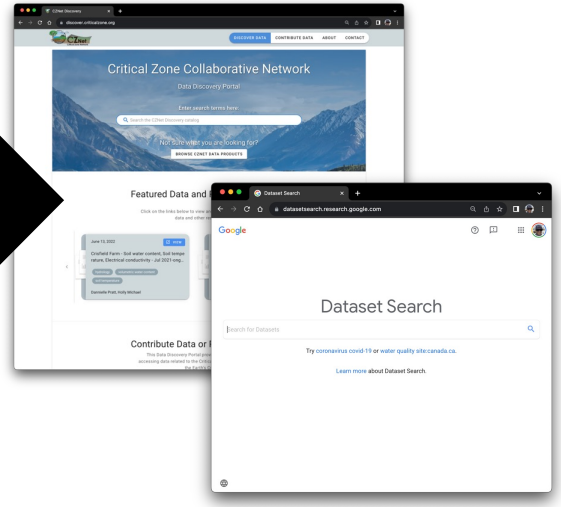
API zenodo

. . .

API Other Repositories

Catalog and Discovery

- Cross-repository view of CZ data and research products
- Discovery based on authors, keywords, geographic area, time, thematic cluster

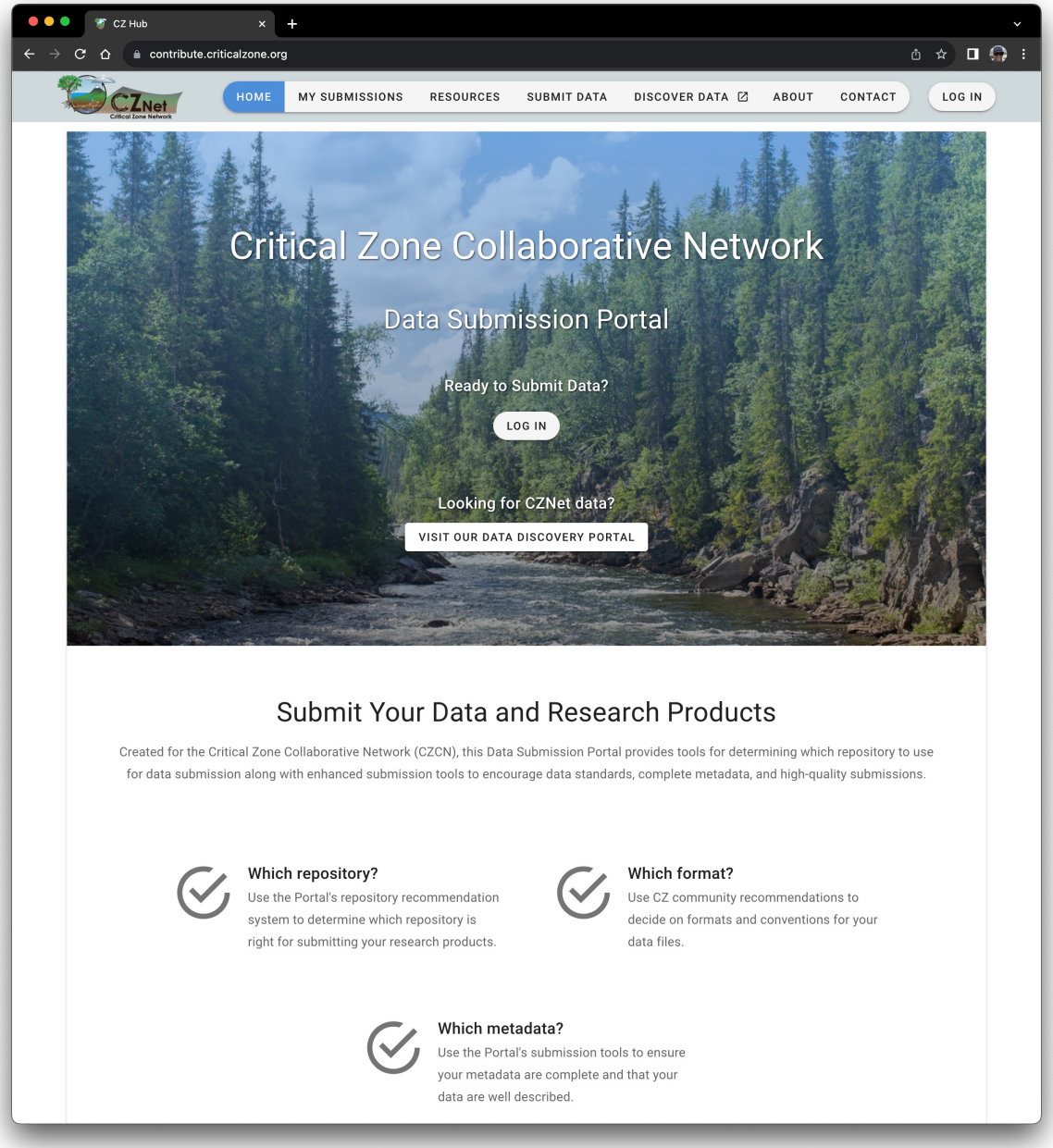


Data Submission Portal

- Web application supporting CZNet
- Enables submission to multiple geoscience data repositories through one portal
- Getting data to the right repository

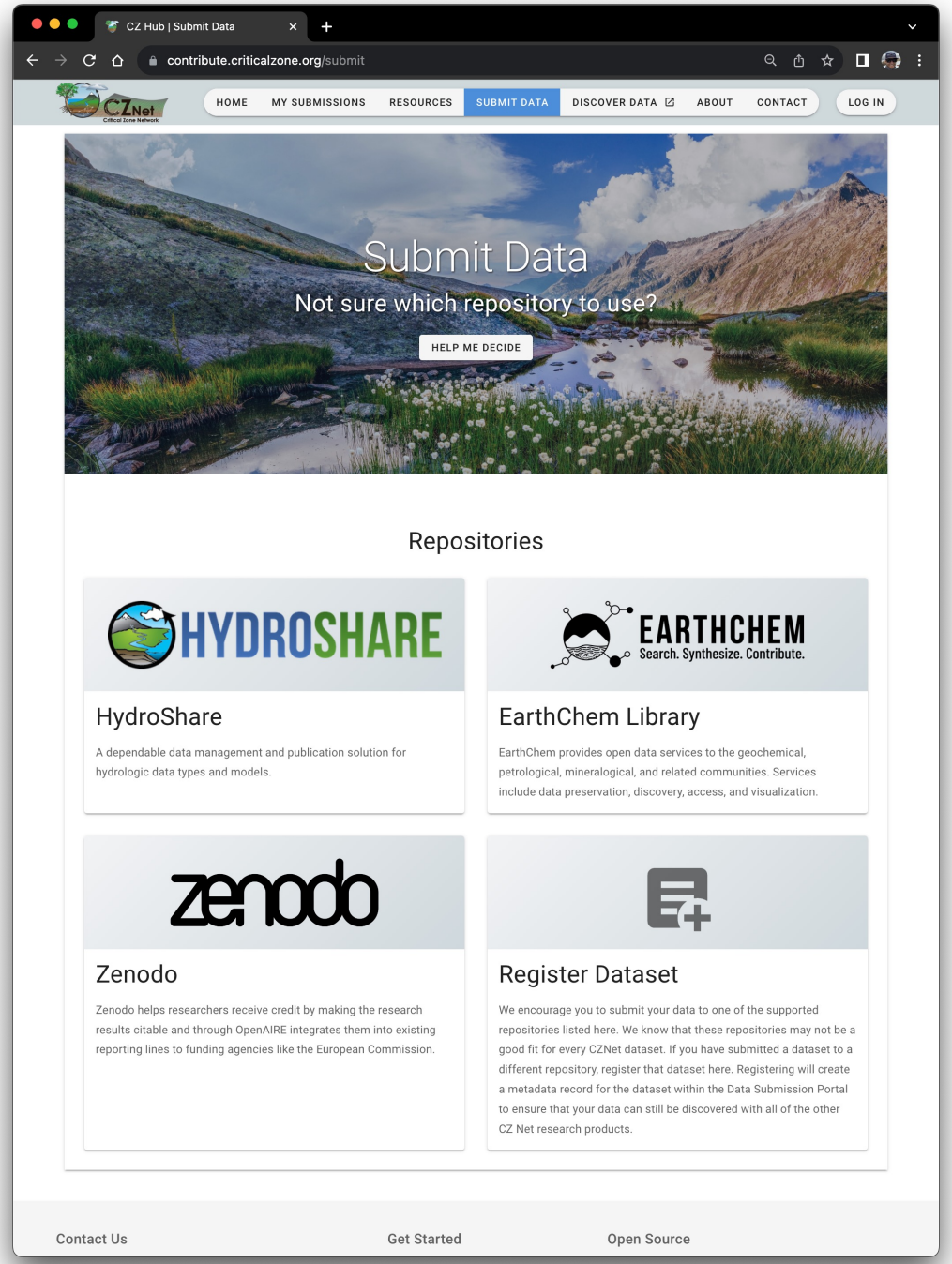
Empower data managers and investigators to curate research products within appropriate repositories with support from the CZ Hub team

https://contribute.criticalzone.org



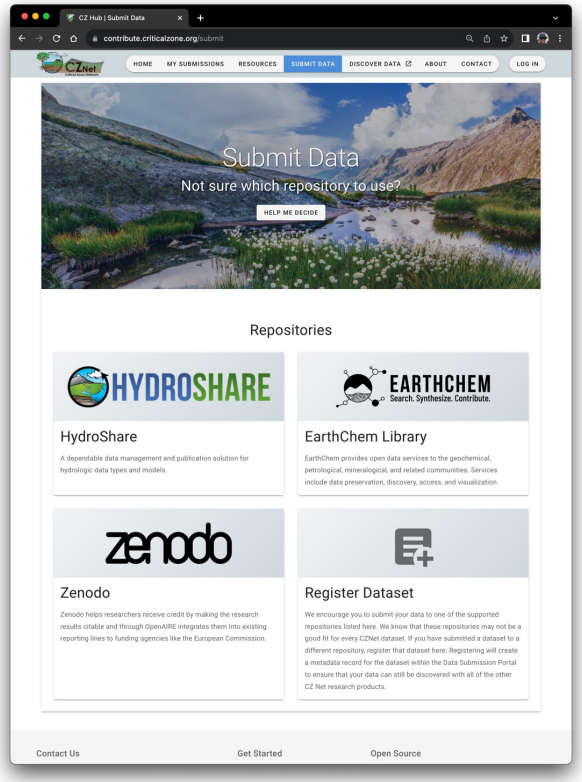
Supported Repositories

- Operate and partner with existing repositories
  - Commonly used by CZ scientists
  - 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



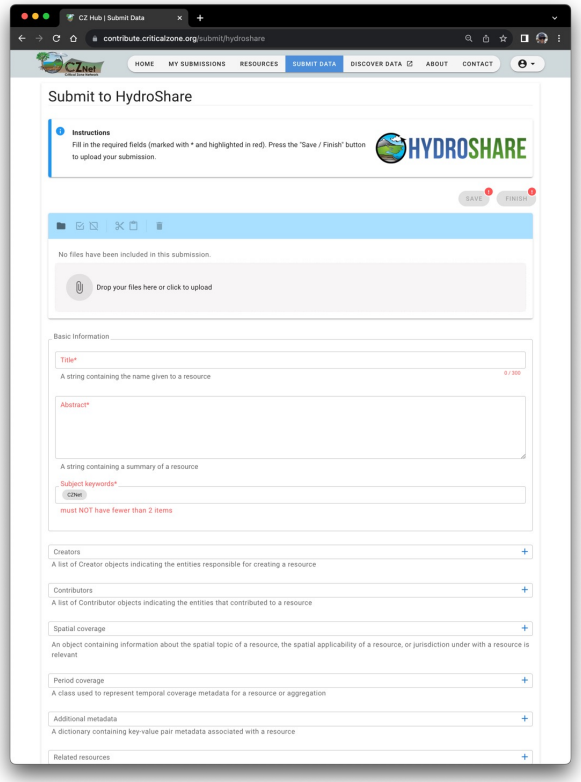
Research Product Submission Workflow

Step 1: Select a repository



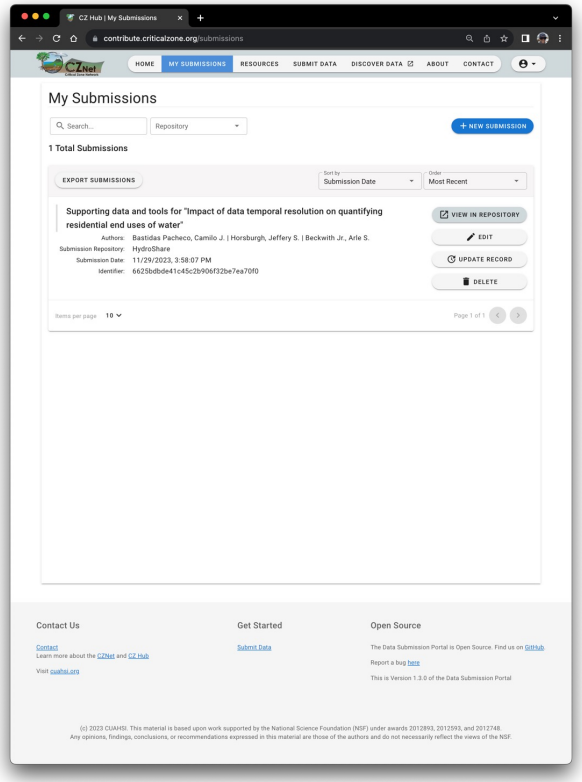
- User chooses repository
- User authorizes the Portal to submit to that repository
- Authorization managed using OAuth 2.0

Step 2: Create content



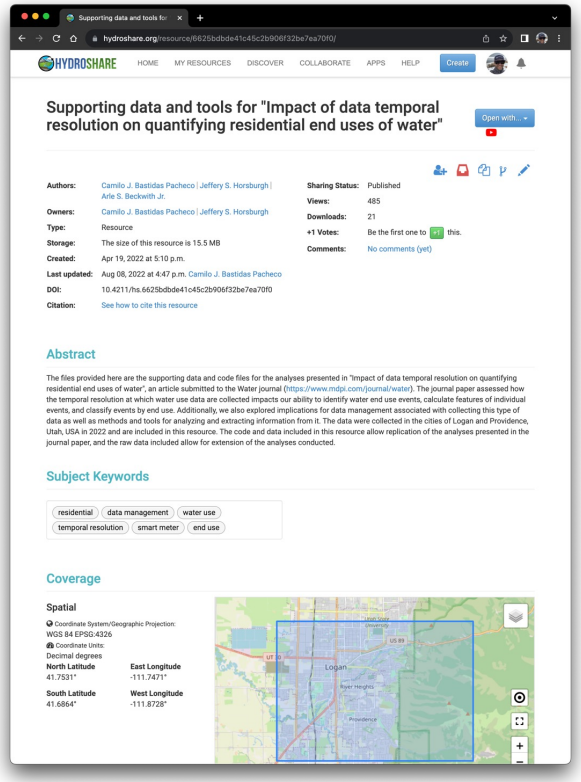
- User completes metadata form
- Metadata is repository specific
- User specifies which files belong will be uploaded

Step 3: Submit content



- Content sent to repository via API
- New record created in repository
- New submission created in user's My Submissions page
- Can view and edit later

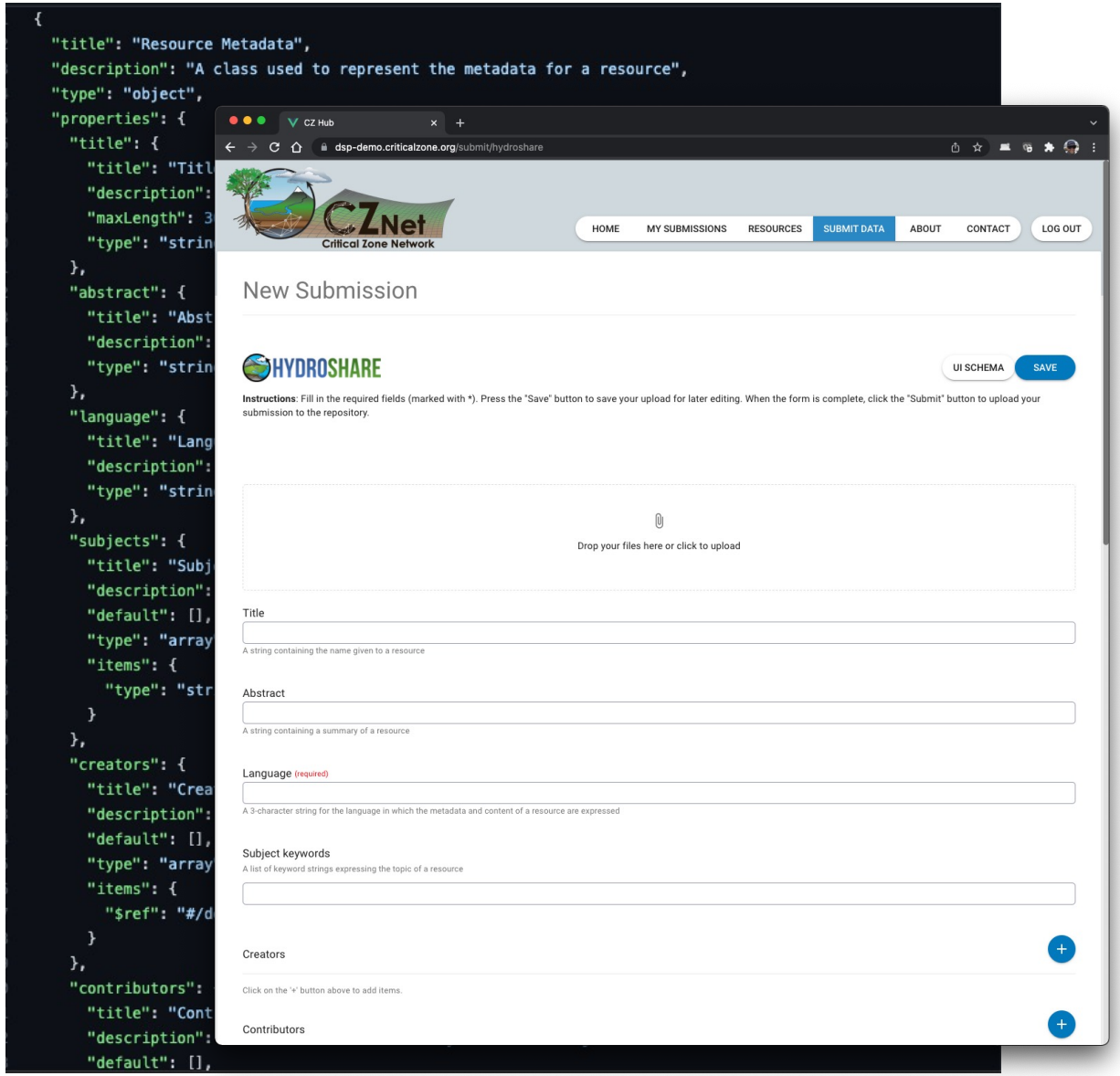
Result in target repository



- Data archived in repository
- Data registered for cataloging and discovery

Interoperability Through JSON Schema-based Metadata

- A JSON schema defines required and optional metadata for each repository
- Enables validation of metadata
  - Data types
  - Default values
  - Required/optional elements
- Data submission form dynamically built from the JSON schema
- Adding a new repository to the Portal means adding a new JSON schema
- Files and metadata sent directly to repository via API
- Data Submission Portal maintains a record of submission



Without the Date Submission Portal

- Data managers must know which repository to use
- 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
- No coordination of metadata to facilitate discovery of CZ Net products

With the Data Submission Portal

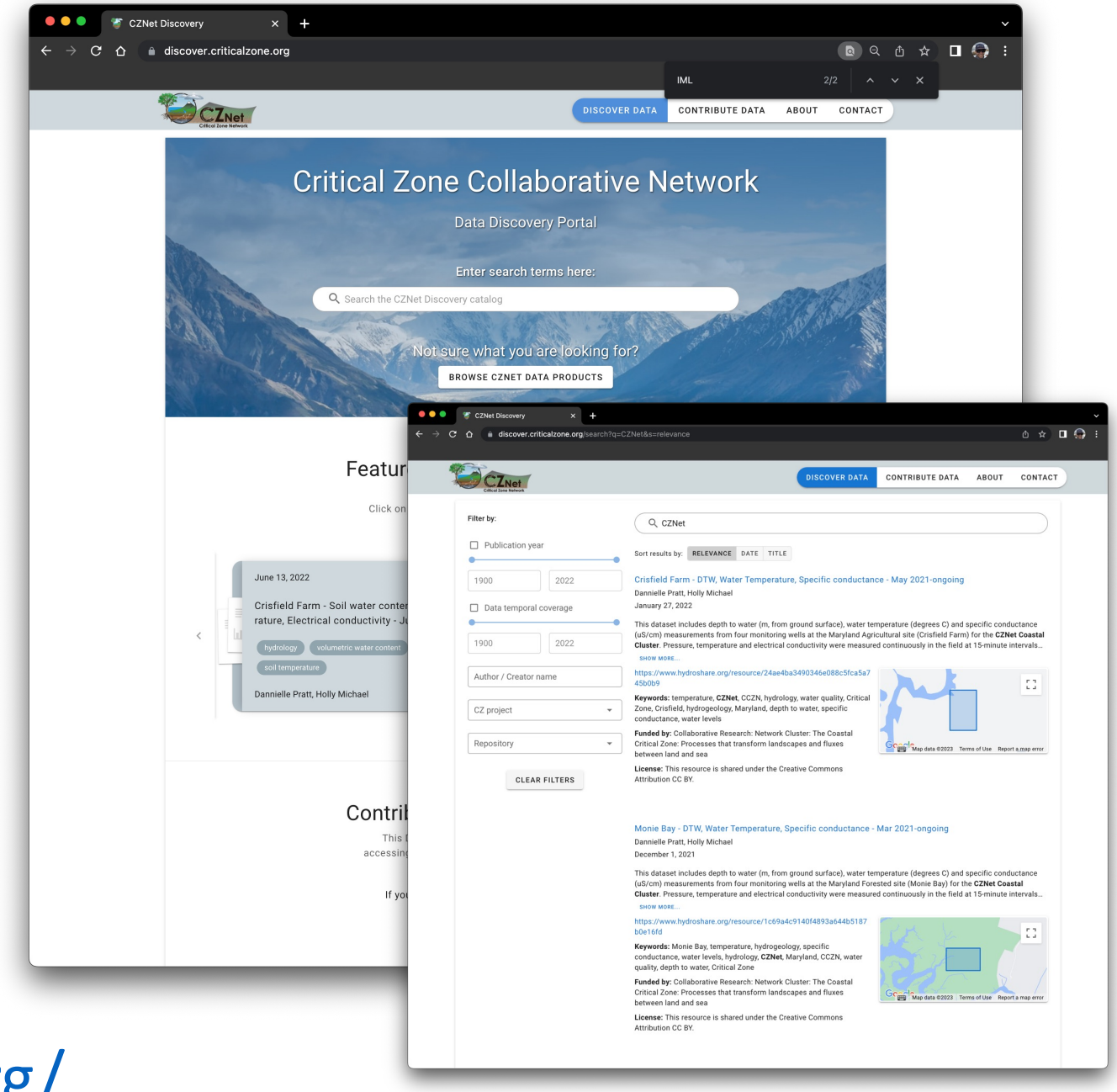
- Ensure data products end up in an appropriate, trusted repository
  - Single interface for direct submission of datasets to HydroShare, EarthChem, and Zenodo
  - User registration of datasets submitted to other repositories
- Validation to promote consistency in CZNet data products across repositories
  - Enforce minimum metadata requirements – e.g., keywords, funding
- Promote templates, common formats, and best practices
- Helping Thematic Clusters track what has been submitted
- Enables simple, consistent, and automated registration of CZNet datasets with a metadata index for discovery

CZ Net Data Discovery

How do prospective data users Find and Access CZ Net datasets spread across multiple repositories?

CZ Net Catalog and Discovery Portal

- Cross-repository view of CZNet data and research products
- Discovery based on keywords, authors, geographic area, time, CZ project, repository
- Interoperability through Schema.org metadata



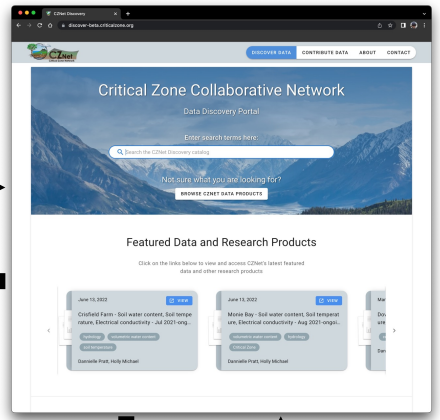
https://discover.criticalzone.org/

How does it work?

CZ Net Data!

Scientists looking for CZ Net data

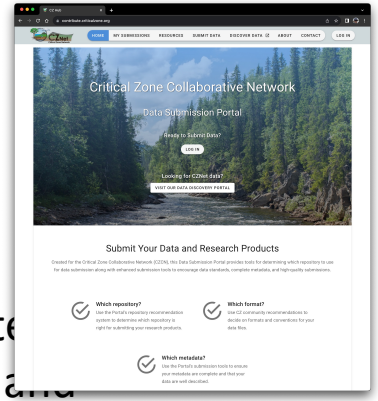
CZ Net Discovery Website and System



Metadata Index/Catalog

Stores metadata for all CZ Net data

Data Submission Portal



Execute queries and returns results

Registered Resources

Normalized Schema.org Metadata

Metadata Harvester

- Initial harvest upon submission
- Daily updates afterward
- Also available on demand

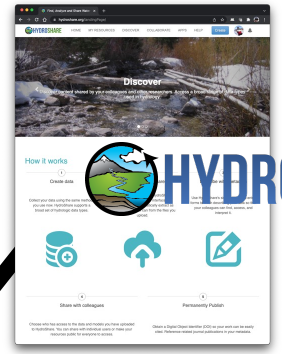
Datasets in Multiple Data Repositories

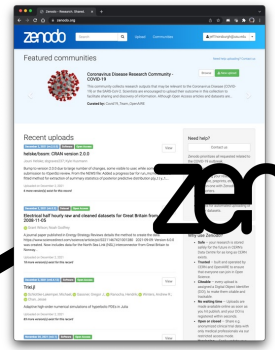
Schema.org metadata

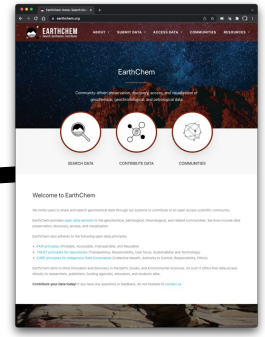
Schema.org JSON-LD

Schema.org JSON-LD

Schema.org JSON-LD





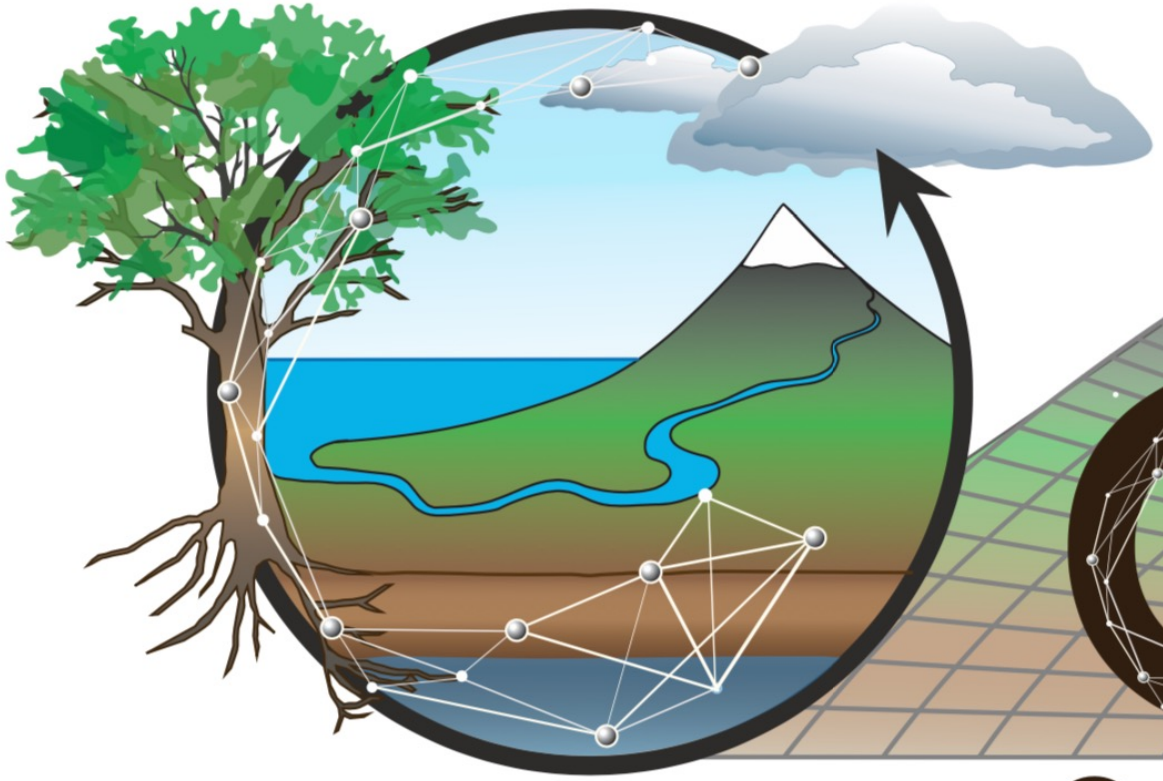


. . .

Other repositories

Conclusions

- The CZ Net approach encourages sharing of data in existing trusted repositories
- The Data Submission Portal facilitates submission and registration – makes it easier for the CZ Net Hub to track submissions and create a coordinated discovery view
- Schema-based metadata has enabled us to build interoperability across repositories for submission and discovery – a model for adding new repositories
- Repository APIs make this possible – but when they change it can break things
- Schema.org metadata make interoperability for consistent cataloging and discovery easier



CZNet

Critical Zone Network

GitHub

CZNet Hub is on GitHub
https://github.com/cznethub

https://www.criticalzone.org/

Questions?

Jeffery S. Horsburgh
jeff.horsburgh@usu.edu

Support: 2012893, 2012748, 2012593