

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.



DUST²

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

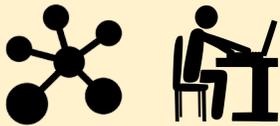
CZ Hub Approach

Diverse data and research products from CZ scientists



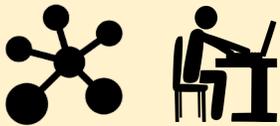
CZ Thematic Cluster Network

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



Cluster Data Manager 1

...



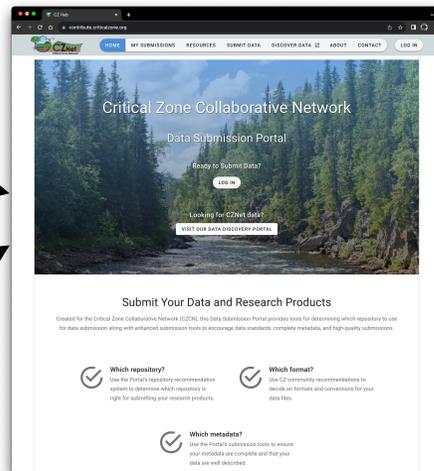
Cluster Data Manager n

Submission of products and/or registration of metadata



Data Submission Portal

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

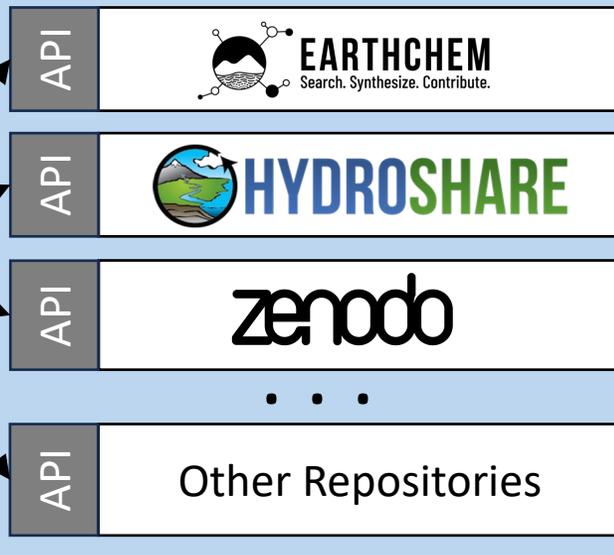


Using existing Earth science data repositories via automated submission



Repositories for Data and Research Products

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

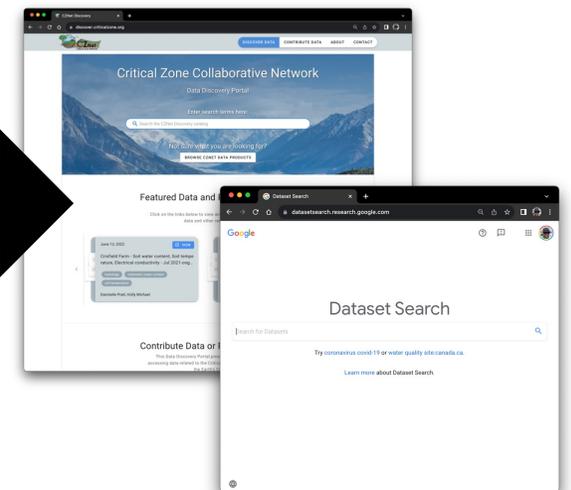


Making products Findable, Accessible, Interoperable, and Reusable (FAIR)



Catalog and Discovery

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

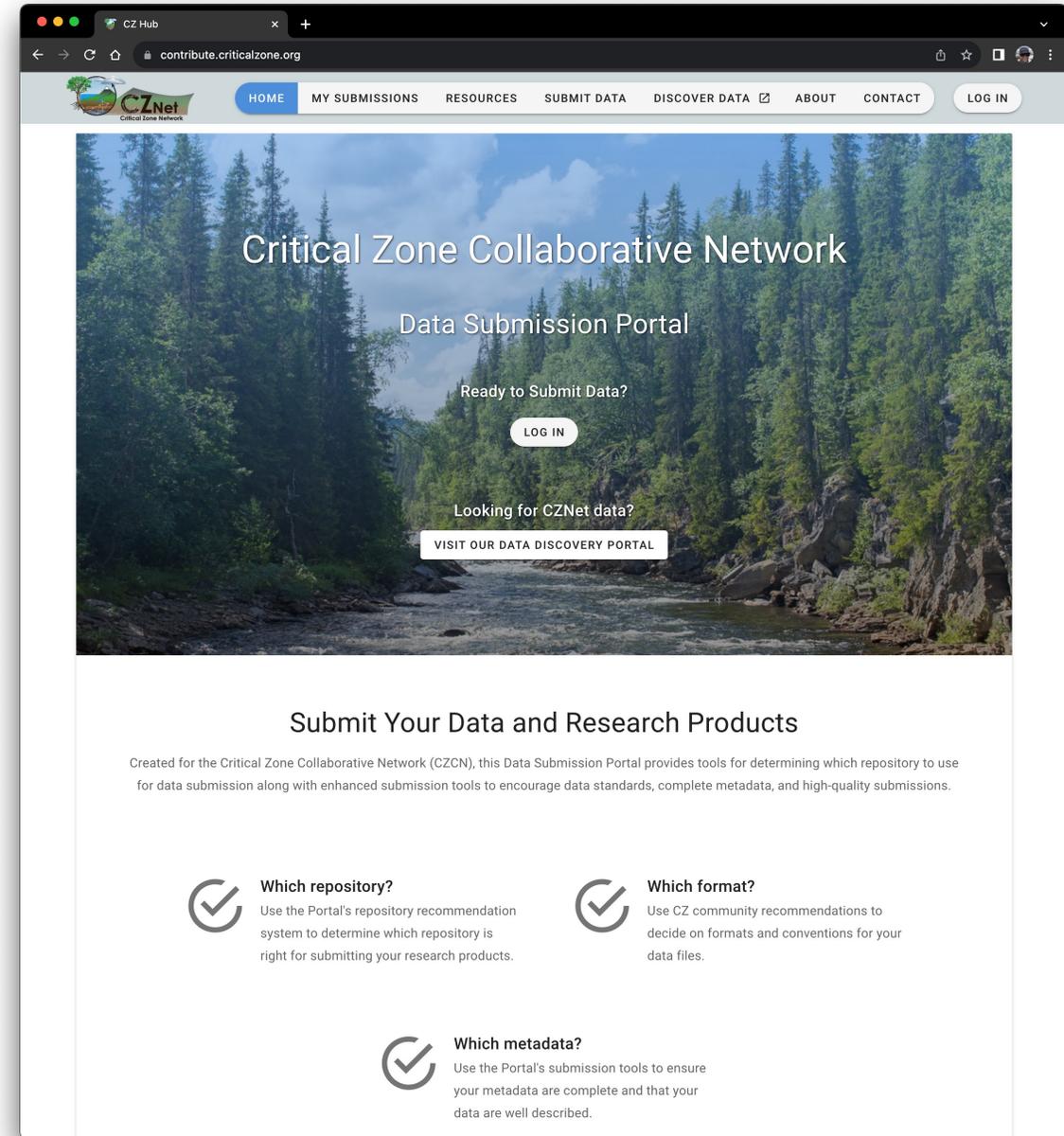


Data Submission Portal

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

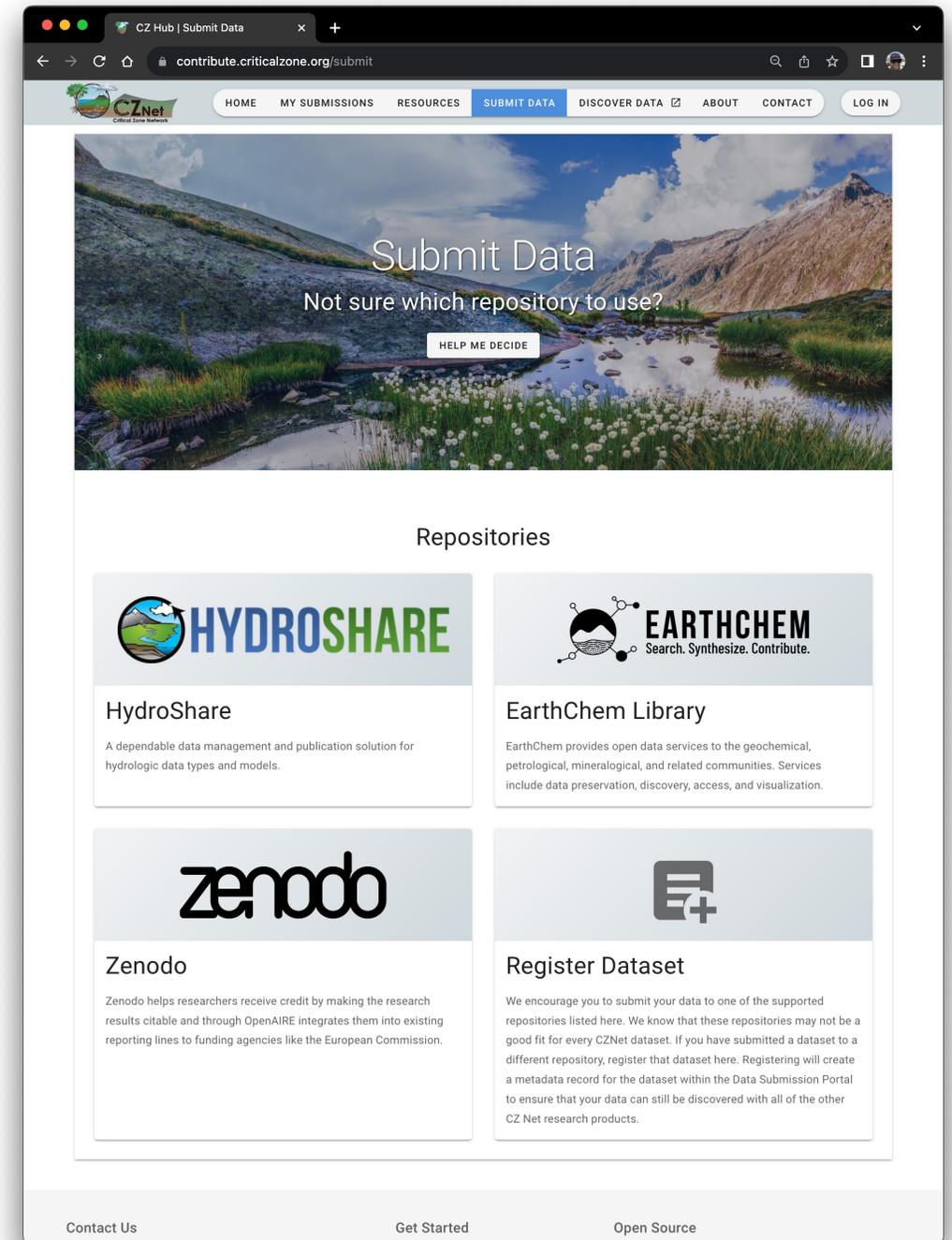
Empower data managers and investigators to curate research products within appropriate repositories with support from the CZ Hub team

<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



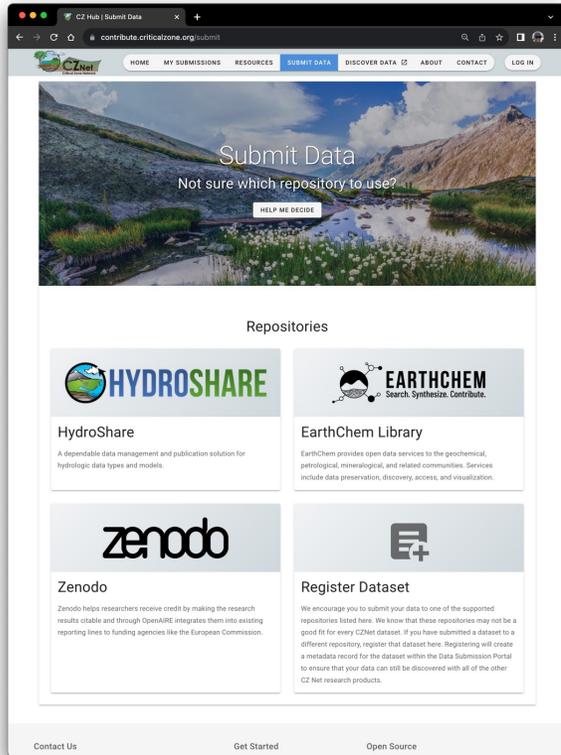
The screenshot shows the 'Submit Data' page on the CZ Hub website. The page features a navigation bar with links for HOME, MY SUBMISSIONS, RESOURCES, SUBMIT DATA, DISCOVER DATA, ABOUT, CONTACT, and LOG IN. The main content area is titled 'Submit Data' and includes a sub-header 'Not sure which repository to use?' with a 'HELP ME DECIDE' button. Below this, a section titled 'Repositories' lists four options:

- HydroShare**: A dependable data management and publication solution for hydrologic data types and models.
- EarthChem Library**: EarthChem provides open data services to the geochemical, petrological, mineralogical, and related communities. Services include data preservation, discovery, access, and visualization.
- Zenodo**: Zenodo helps researchers receive credit by making the research results citable and through OpenAIRE integrates them into existing reporting lines to funding agencies like the European Commission.
- Register Dataset**: We encourage you to submit your data to one of the supported repositories listed here. We know that these repositories may not be a good fit for every CZNet dataset. If you have submitted a dataset to a different repository, register that dataset here. Registering will create a metadata record for the dataset within the Data Submission Portal to ensure that your data can still be discovered with all of the other CZ Net research products.

At the bottom of the page, there are three links: Contact Us, Get Started, and Open Source.

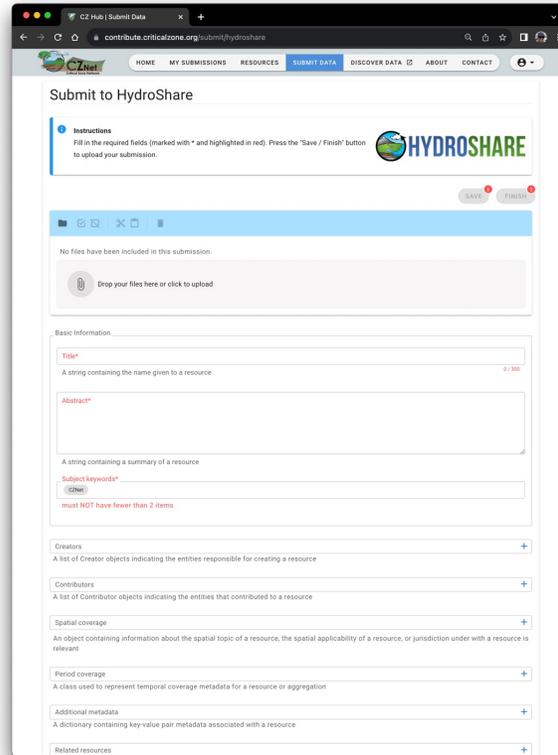
Research Product Submission Workflow

Step 1: Select a repository



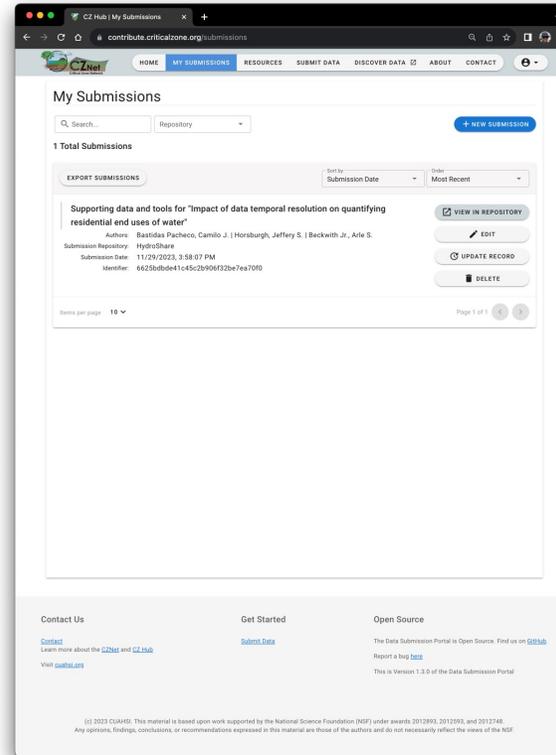
- User chooses repository
- User authorizes the Portal to submit to that repository
- Authorization managed using OAuth 2.0

Step 2: Create content



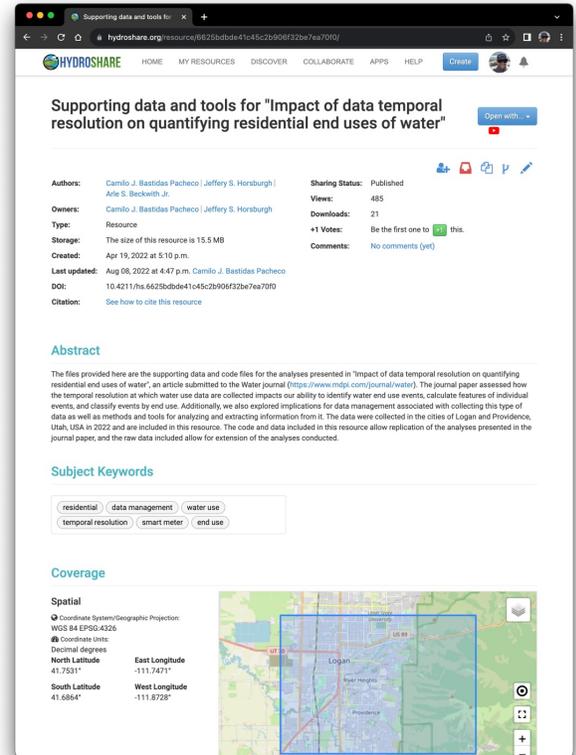
- User completes metadata form
- Metadata is repository specific
- User specifies which files belong will be uploaded

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

The image displays a web application interface for a data submission portal. On the left side, a JSON schema is shown in a dark-themed editor. The schema defines a 'Resource Metadata' object with various properties like 'title', 'description', 'abstract', 'language', 'subjects', 'creators', and 'contributors'. On the right side, a browser window shows the 'New Submission' form for 'HYDROSHARE'. The form includes a file upload area with the instruction 'Drop your files here or click to upload', and several text input fields for 'Title', 'Abstract', 'Language (required)', 'Subject keywords', 'Creators', and 'Contributors'. The browser's address bar shows the URL 'dsp-demo.criticalzone.org/submit/hydroshare'.

```
{
  "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 are expressed",
      "type": "string"
    },
    "subjects": {
      "title": "Subjects",
      "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 strings identifying the creator(s) of a resource",
      "default": [],
      "type": "array",
      "items": {
        "$ref": "#/definitions/ResourceMetadata/properties/title"
      }
    },
    "contributors": {
      "title": "Contributors",
      "description": "A list of strings identifying the contributor(s) of a resource",
      "default": [],
      "type": "array",
      "items": {
        "$ref": "#/definitions/ResourceMetadata/properties/title"
      }
    }
  }
}
```

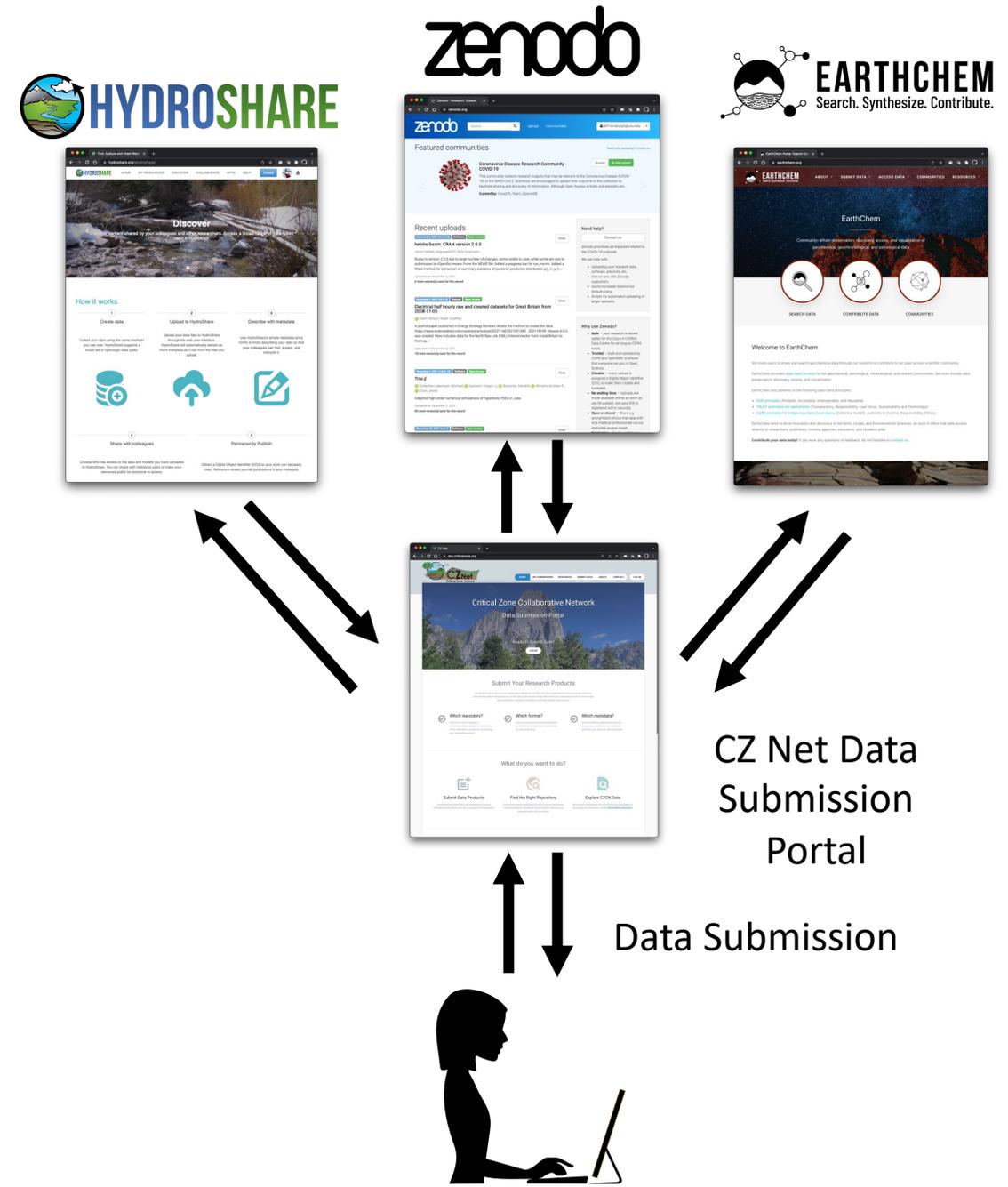
Without the Data Submission Portal

- 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

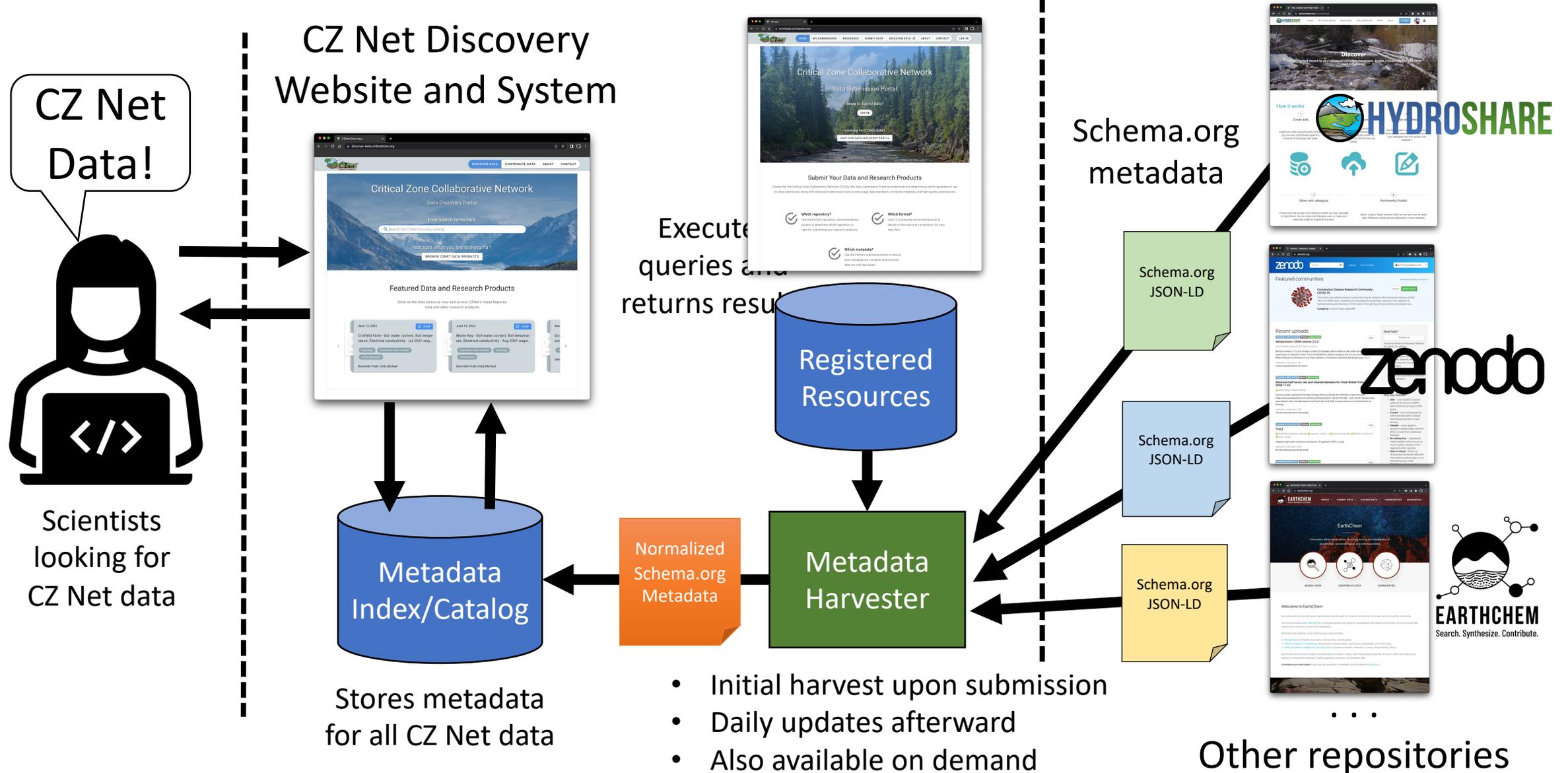
The image displays two screenshots of the CZNet Discovery Portal. The top screenshot shows the homepage with the title "Critical Zone Collaborative Network Data Discovery Portal" and a search bar. The bottom screenshot shows a search results page with filters and two data entries.

Search Results:

- Filter by:** Publication year (1900 to 2022), Data temporal coverage (1900 to 2022), Author / Creator name, CZ project, Repository.
- Sort results by:** RELEVANCE, DATE, TITLE.
- Entry 1:** Crisfield Farm - DTW, Water Temperature, Specific conductance - May 2021-ongoing. Dannielle Pratt, Holly Michael. January 27, 2022. This dataset includes depth to water (m, from ground surface), water temperature (degrees C) and specific conductance (uS/cm) measurements from four monitoring wells at the Maryland Agricultural site (Crisfield Farm) for the CZNet Coastal Cluster. Pressure, temperature and electrical conductivity were measured continuously in the field at 15-minute intervals... <https://www.hydroshare.org/resource/24ae4ba3490346e088c5fca5a7450f069>
- Entry 2:** Monie Bay - DTW, Water Temperature, Specific conductance - Mar 2021-ongoing. Dannielle Pratt, Holly Michael. December 1, 2021. This dataset includes depth to water (m, from ground surface), water temperature (degrees C) and specific conductance (uS/cm) measurements from four monitoring wells at the Maryland Forested site (Monie Bay) for the CZNet Coastal Cluster. Pressure, temperature and electrical conductivity were measured continuously in the field at 15-minute intervals... <https://www.hydroshare.org/resource/1c69a4c9140f4893a644b5187b0e16fd>

<https://discover.criticalzone.org/>

How does it work?



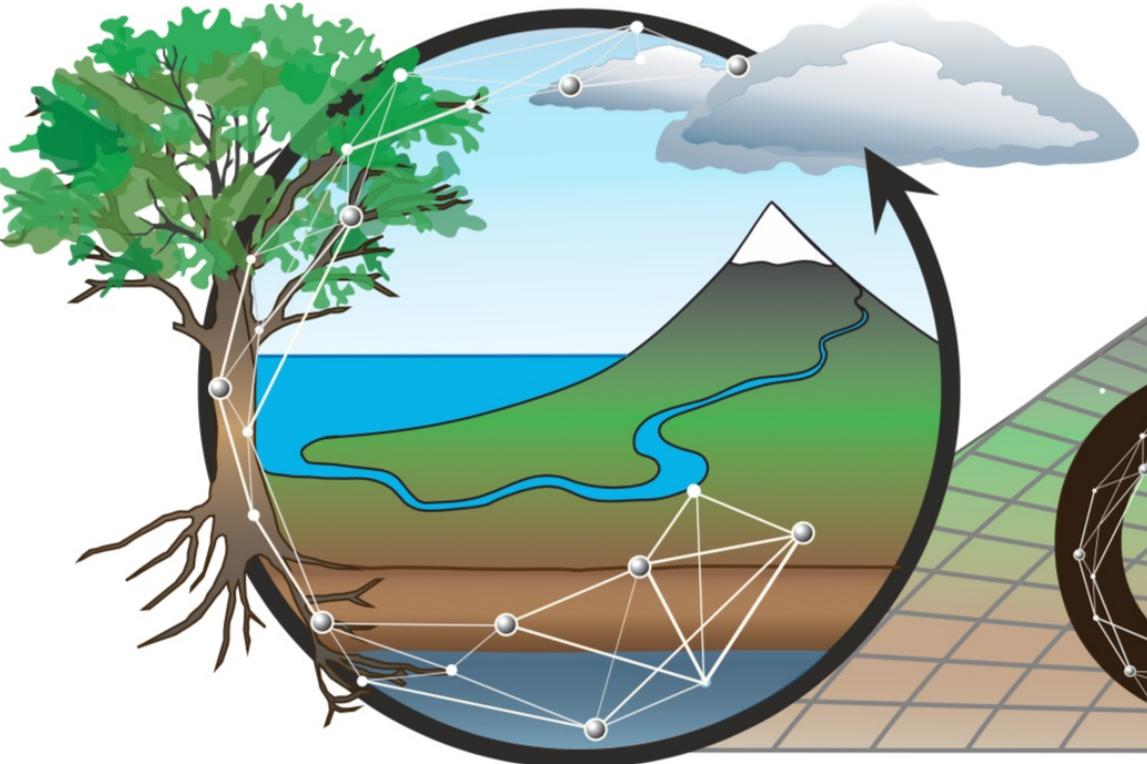
Conclusions

- 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



GitHub

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



CZNet

Critical Zone Network

<https://www.criticalzone.org/>



Support: 2012893,
2012748, 2012593

Questions?

Jeffery S. Horsburgh
jeff.horsburgh@usu.edu