Integrated Data Management System for Critical Zone Observatories

A pilot for EarthCube?

Anthony K. Aufdenkampe, Stroud Water Research Center
Kerstin Lehnert, IEDA/Columbia.
Ilya Zaslavsky, SDSC.
Jeff Horsburgh, USU.
Emilio Mayorga, UW-APL
David Tarboton, USU
David Lubinski, UC-Boulder

STROUD WATER RESEARCH CENTER

UtahState University

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UNIVERSITY of WASHINGTON

IEDA INTEGRATED EARTH DATA APPLICATIONS

News

Obama issues Executive Order in support of open data

Timothy Vollmer, May 10th, 2013

Yesterday President Barack Obama issued an Executive Order requiring federal government information to be open and machine-readable by default. This Order is the latest in a series of actions going back to 2009 in support of increasing access to and transparency of government information.

In addition to the Executive Order, the White House released a Memorandum (PDF) explaining how federal government agencies will comply with the new open data policy.

> This Memorandum requires agencies to collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine readable and open formats, data standards, and common core and extensible metadata for all new information creation and collection efforts. It also includes agencies ensuring information stewardship through the use of open licenses and review of information for privacy, confidentiality, security, or other restrictions to release.

Seal Of The Executive Office Of The President / Public Domain

It provides a forward-thinking set of guidelines for open data to be released by U.S. federal agencies:

Open data: For the purposes of this Memorandum, the term "open data" refers to publicly available data structured in a way that enables the data to be fully discoverable and usable by end users. In general, open data will be consistent with the following principles:

http://creativecommons.org/weblog/entry/38269

Integrated CZO Data: 2009-2011

$500k supplement to Boulder Creek CZO, PI: M. Williams

- Support Data Management at 1st Three CZOs
  - $240k total, $80k each for Boulder, Merced & PSU
- National CZO website & data workshops
  - $60k total, $40k to Boulder, $10k each Merced & PSU
- CZO Central Repository Pilot for Hydro Data
  - $100k to SDSC
- CZO Desktop Pilot
  - $100k to Utah State University & Idaho State Univ.

Zaslavsky et al. 2011, *Environ. Info. Management*

http://criticalzone.org/national/publications/pub/zaslavsky-et-al-2011-the-initial-design-of-data-sharing-infrastructure-for-/

CZOData: Overall Approach

- Do not reinvent the wheel! Build on
  - CUAHSI HIS, EarthChemDB, LTER, etc
- Consistent data presentation on web
  - Metadata
  - Data values
- Central data system for data discovery
  - Harvested by SDSC (pull system)

CZOData: Principles & Policies

- Each CZO will operate and be responsible for its own local data management system for collecting, organizing, quality controlling and publishing data through its web site.
  - Different philosophy than CUAHSI HIS
  - Each CZO is master of it's own data
  - We don't care what goes on under the hood (to an extent)
  - Each site uses it's own protocols, data bases, etc.
  - Allows CZO to honor site legacy data

CZOData: Principles & Policies

- Each CZO publishes its data on the web in ascii text format with sufficient metadata so that the data can be unambiguously interpreted
- Metadata follows a proscribed format
  - Data managers just need rules to follow
- Easy to harvest by central portal
- Makes it simple at the site level so scientists comply
  - Addresses the chokepoint that is getting data/ metadata from the scientists to data managers

CZO Data Publication System

CZO Data Repository & Indexing (CZO Central)

External cross-project registries

DataNet, NEON

Standard CZO Services

CZO Metadata | Ontology | Shared vocabularies | Archive | Harvester

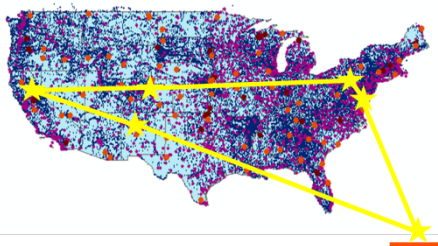
CZO Data Products

CZO Desktop Applications

CZO Desktop

- Matlab
- R
- Excel
- ArcGIS
- Modeling

CZO Web-based Data Discovery System

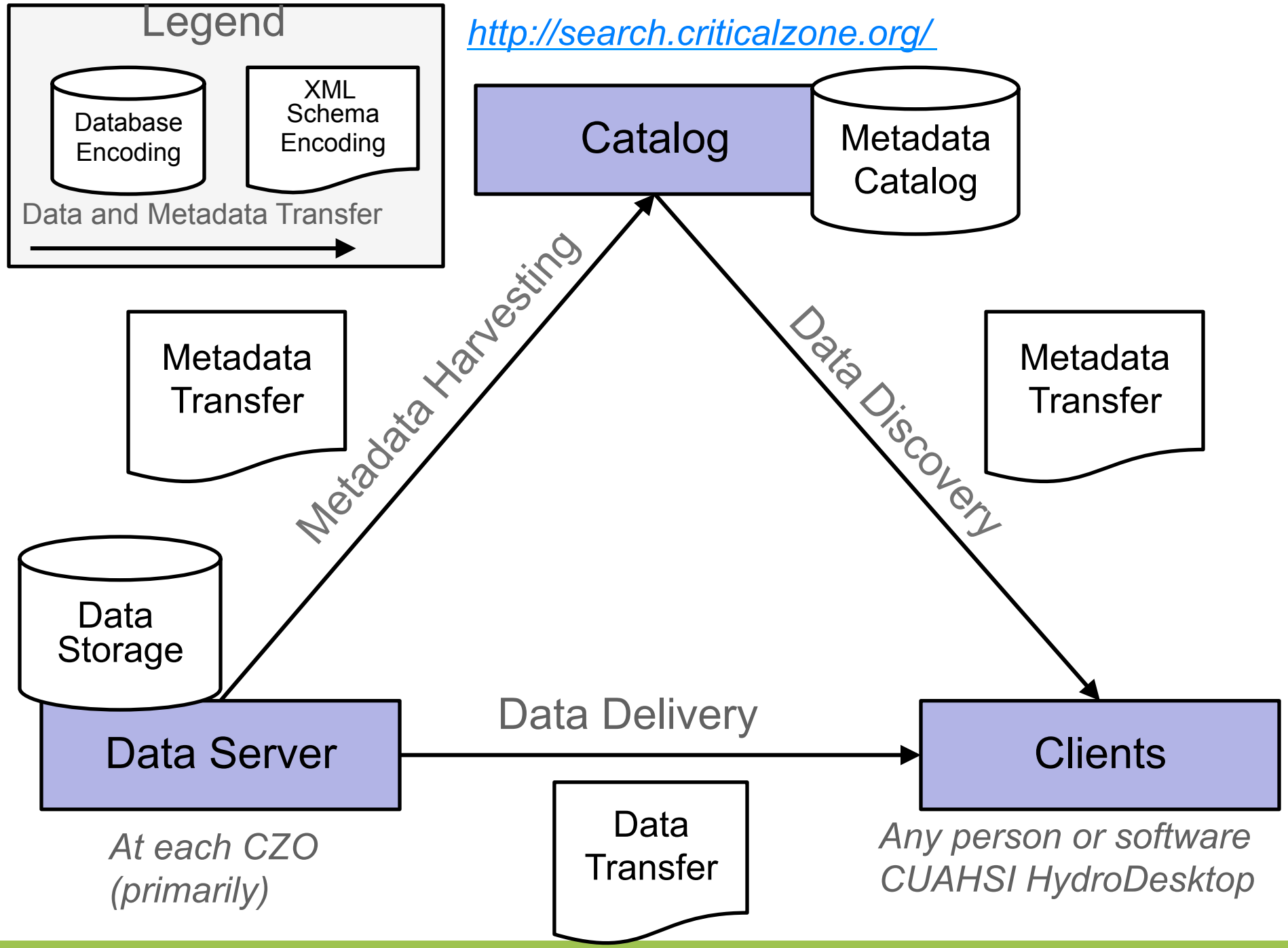


Standard CZO data display formats

Web site | Web site | Web site

Local CZO DB | Local CZO DB | Local CZO DB

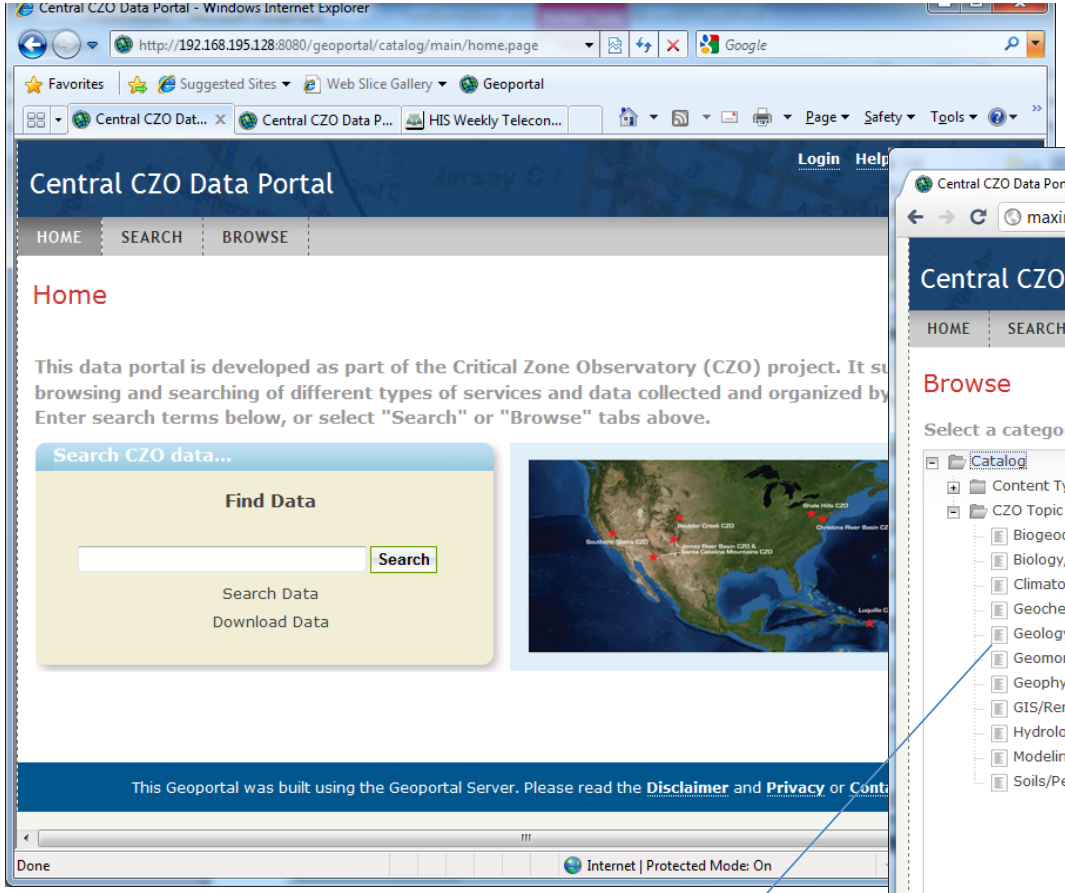
Spatial, hydrologic, geophysical, geochemical, imagery, spectral…

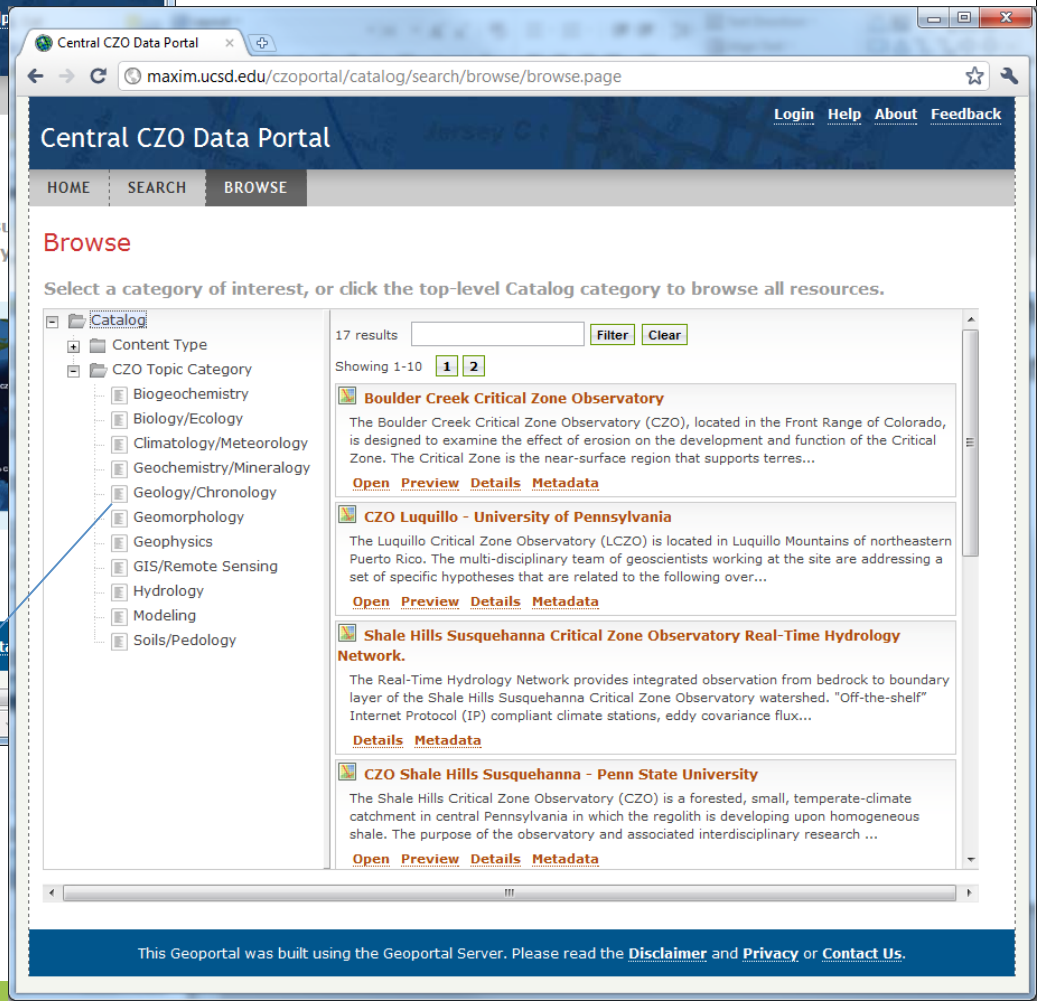

Legend
Database Encoding
XML Schema Encoding
Data and Metadata Transfer
http://search.criticalzone.org/
Catalog
Metadata Catalog
Metadata Transfer
Metadata Harvesting
Data Discovery
Metadata Transfer
Data Storage
Data Server
Data Delivery
Clients
Data Transfer
At each CZO (primarily)
Any person or software CUAHSI HydroDesktop

CZO Data Portal

At http://search.criticalzone.org/

Built on Geoportal metadata catalog software:
https://github.com/Esri/geoportal-server





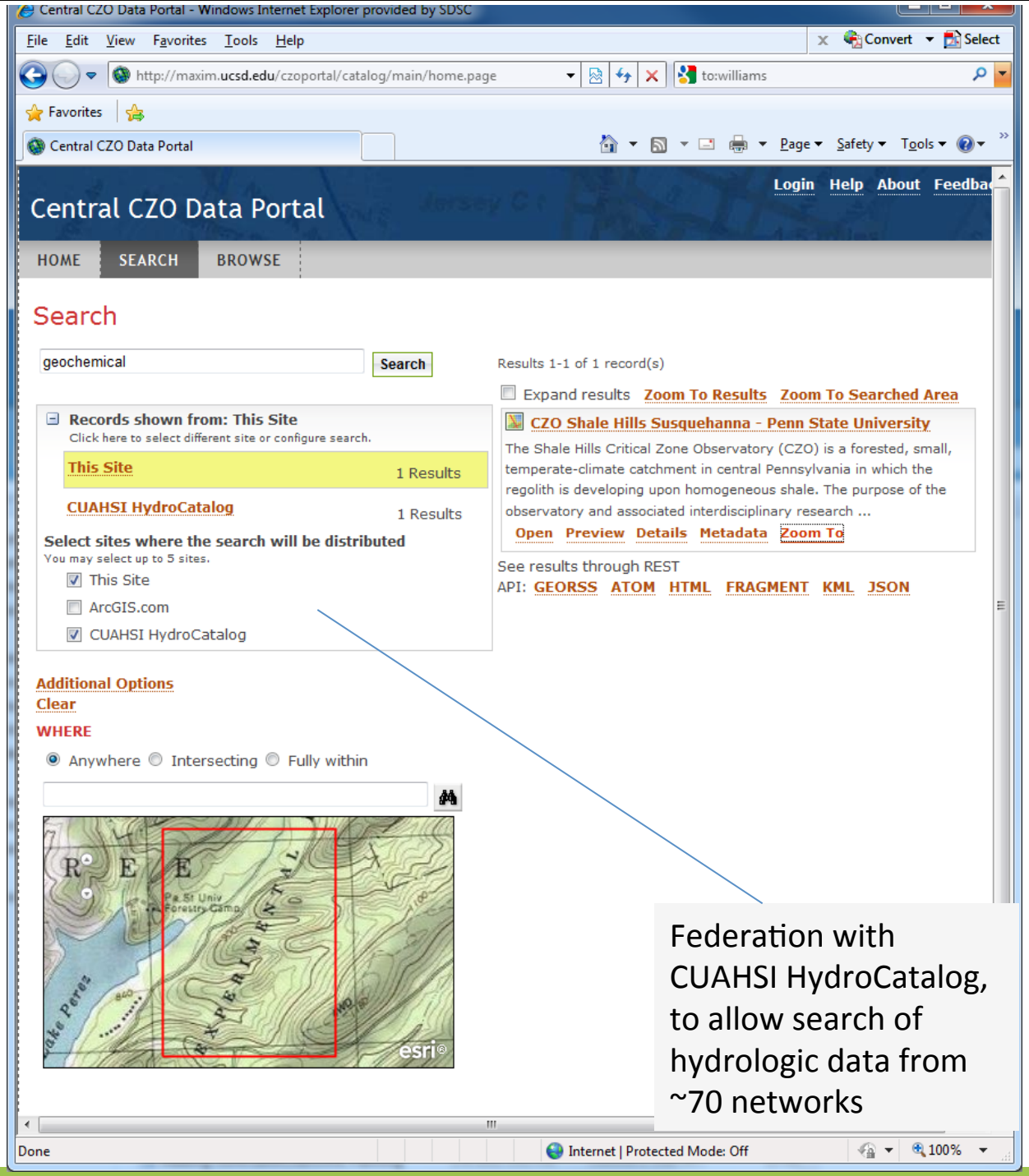
Registered data are organized by CZO thematic categories

The portal is CSW-compliant (CSW=Catalog Services for the Web): can be federated with other catalogs including *data.gov*

Supports search by location, resource type, thematic category, keywords, plus full-text abstract search

Display files from CZO web sites are registered to the data discovery portal automatically

In addition, display files of known types are expressed as data services, which are also registered in the portal



Federation with CUAHSI HydroCatalog, to allow search of hydrologic data from ~70 networks

CZO Display File (version 1)

Requirements:

- Human readable
- Machine parsable
- Conform to ODM 1.1 & CZO Shared Vocabulary

Solution:

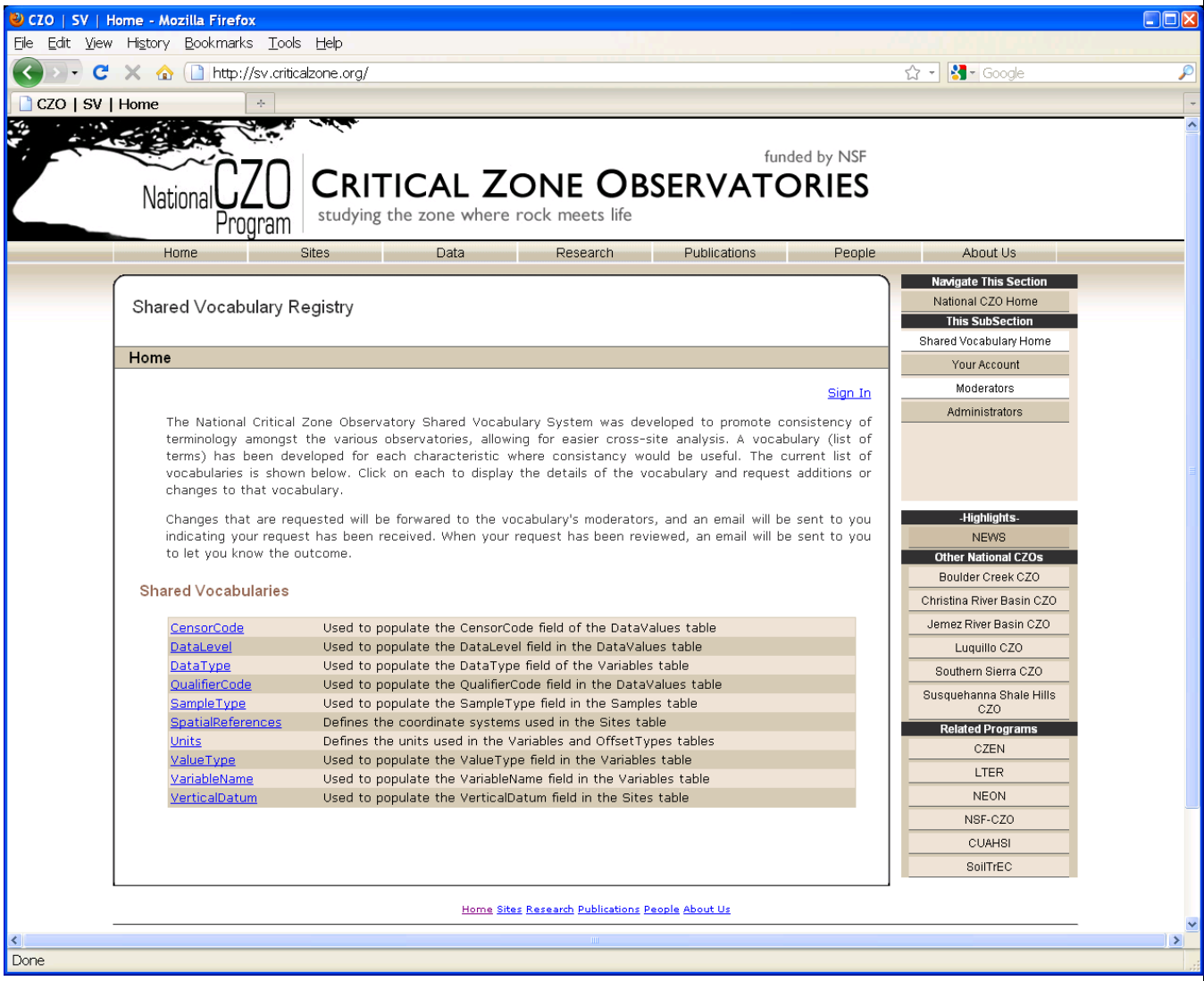
- ASCII text file, with comma-separated data
- Header with detailed, structured metadata
- Separate Sites and Methods files
- http://criticalzone.org/national/publications/pub/whitenack-et-al-2011-czo-display-file-specification/

CZO Shared Vocabulary System

http://sv.critialzone.org

Purpose:

To promote consistent use of terminology.

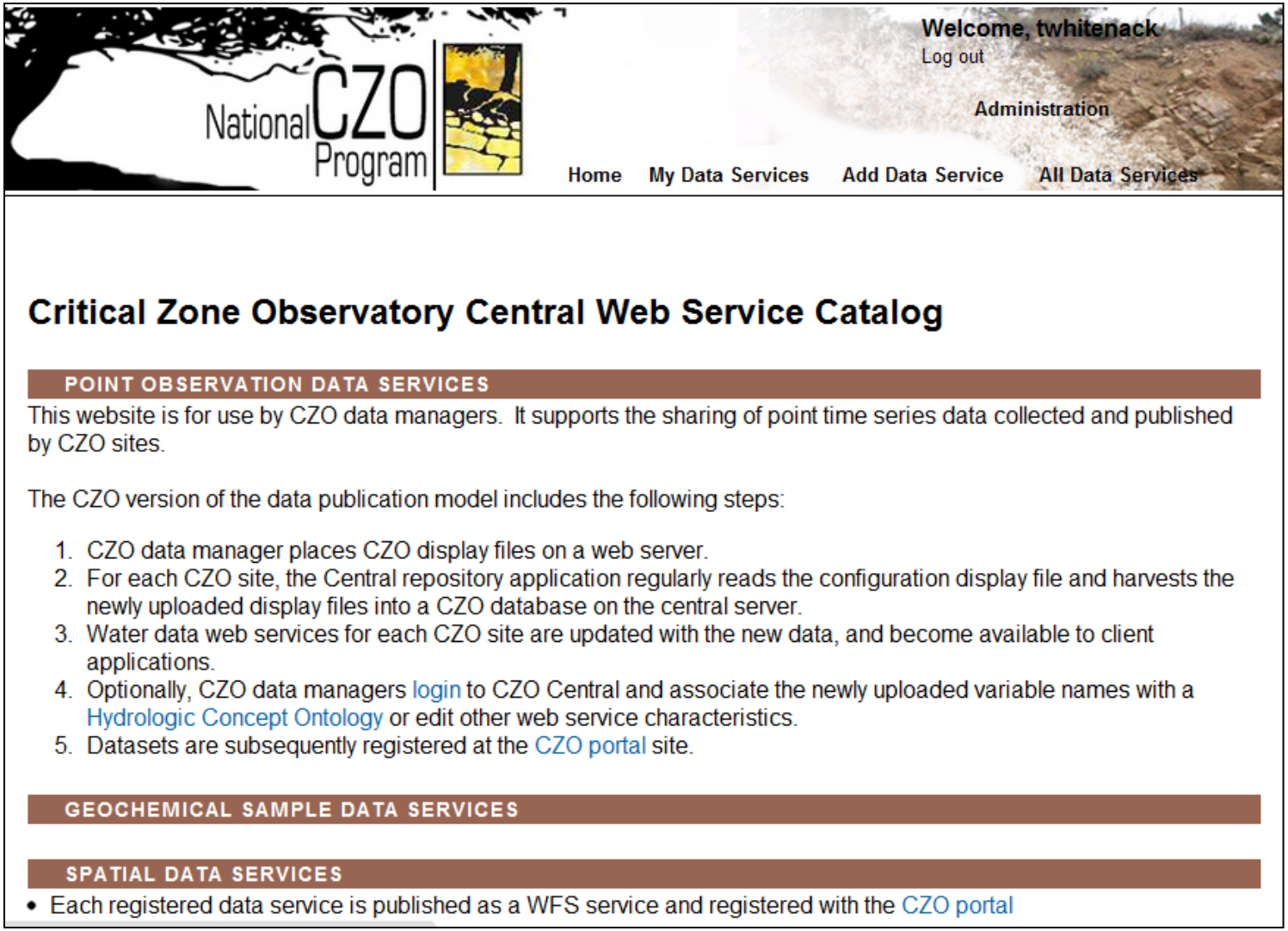


Builds on CUAHSI Master CV Registry, http://his.cuahsi.org/mastercvreg/

CZO Data Management Web Admin. Interface

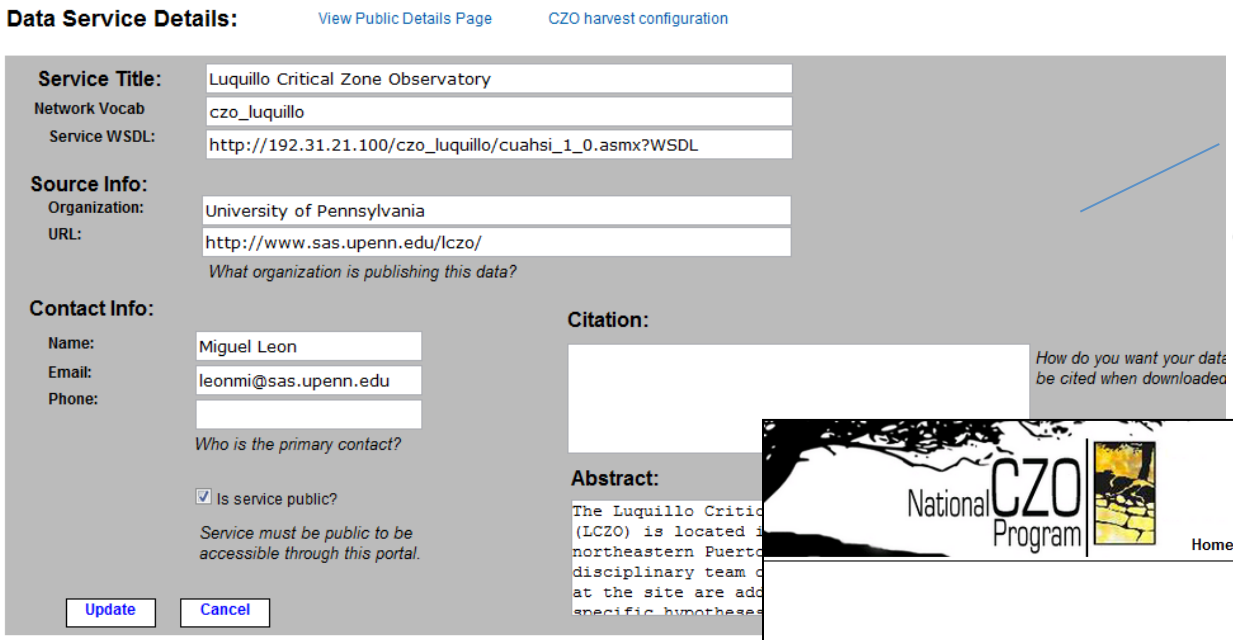
CZO data managers use this web-based system to register display files, edit service metadata, initiate data retrieval, validate the data against shared vocabularies, and update hydrologic time series services

The administration system will be extended to geochemical samples and other data

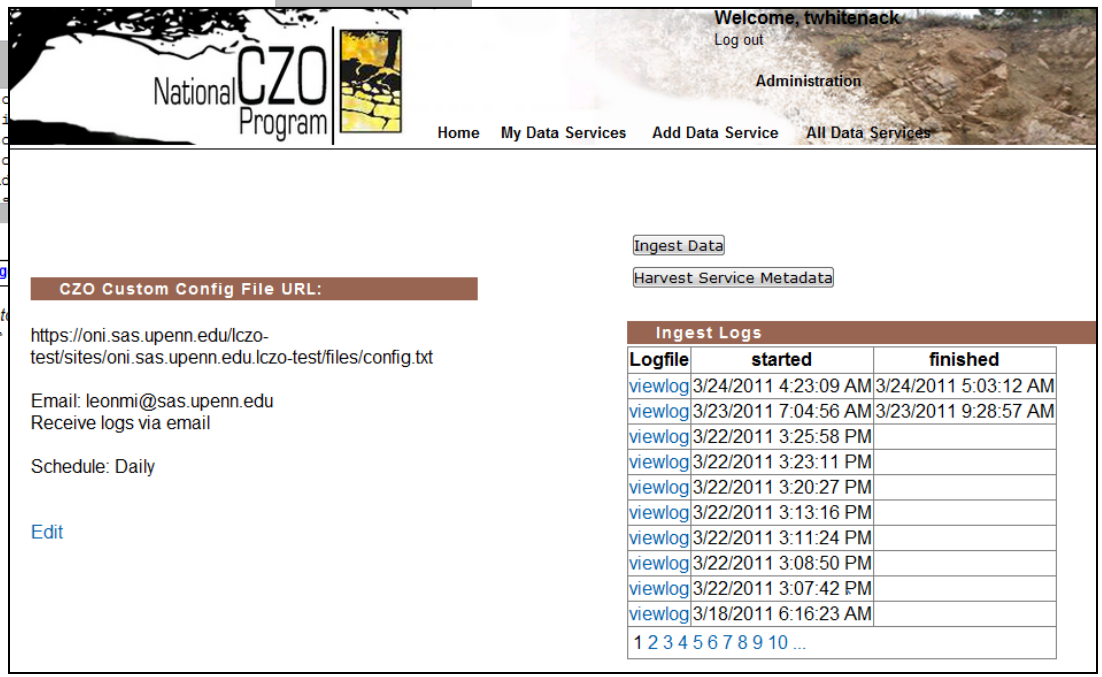


http://central.criticalzone.org

Services registered by CZO data managers



Editable service definitions and management interface for each CZO data service



Data managers control how their data is annotated.

Ingesting of Display files is triggered on the server by the Data manager.

Display file ingestion log

Data Publication Process

(for hydrologic time series)

\doc

TITLE. Streamflow data for Jemez River Basin.

ABSTRACT. 30 minute streamflow data measured and computed for
National Preserve).

INVESTIGATOR. Peter Troch, Department of Hydrology and Water
Tucson, [illegible]

VARIABLE NAMES. Streamflow.

KEYWORDS. hydrology, streamflow, Valle Caldera, New Mexico.

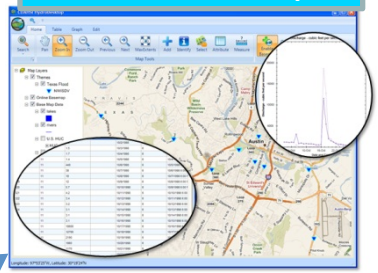
CITATION. [1] Broxton P.D., Troch P.A., and Lyon S.W. (2009)
small mountainous catchments. Water Resources Research 45: W0
[2] The following acknowledgment should accompany any publica
data were provided by the NSF-supported Jemez River Basin and

CZO Display File

ODM

WaterML Service

CZO Desktop



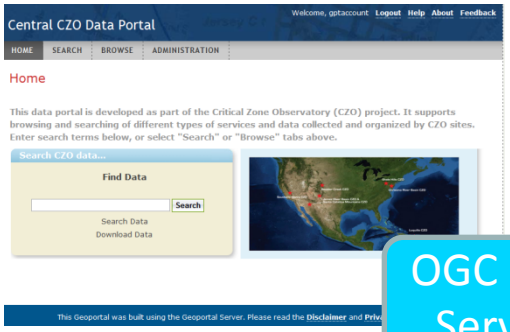
CZO Central Catalog

Catalog Search Service

Raw Display file metadata Is registered with the CZO data portal, to assure original data is discoverable and downloadable.

OGC WFS Service

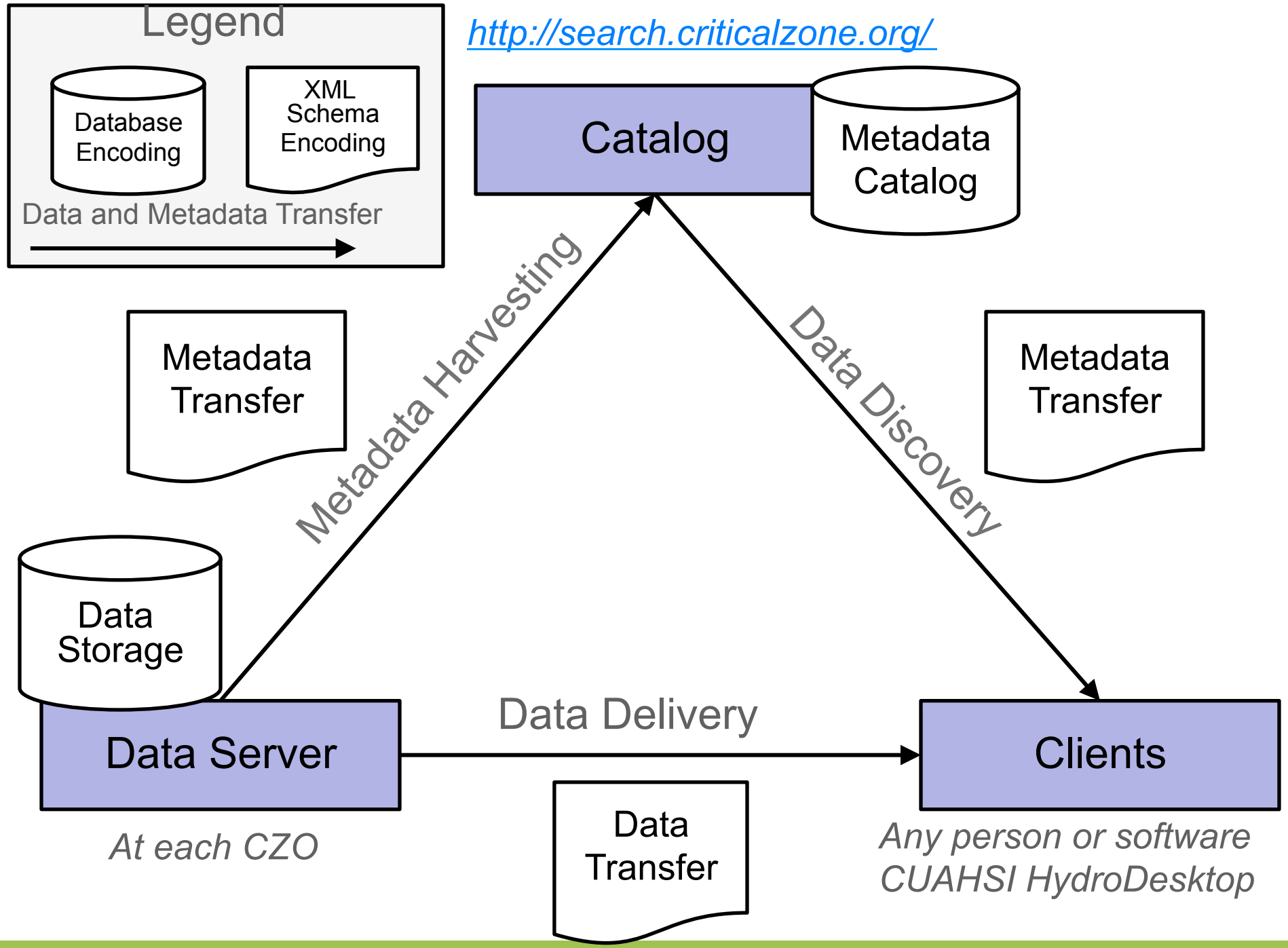
WFS Service Is registered with the CZO data portal



OGC CSW Service

CZO Portal utilizes the OGC CSW (catalog services for the web)

Broader internet community accessing data using standard protocols.

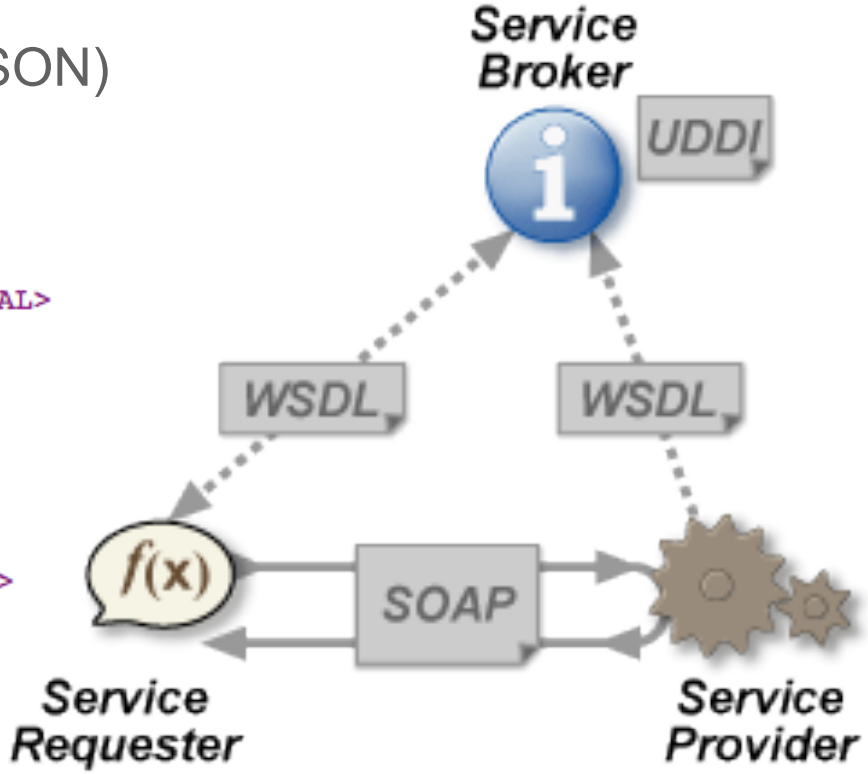

Legend
Database Encoding
XML Schema Encoding
Data and Metadata Transfer
http://search.criticalzone.org/
Catalog
Metadata Catalog
Metadata Transfer
Metadata Harvesting
Data Discovery
Metadata Transfer
Data Storage
Data Server
Data Delivery
Clients
At each CZO
Data Transfer
Any person or software
CUAHSI HydroDesktop

Web Services

a method of communication between two electronic devices over the web

- Service Requester/Consumer (clients) request information from Service Providers (servers) using standard operations like GET, POST, PUT, DELETE for HTTP
- Data is exchange via XML (or JSON)

```
<!-- Edited by XMLSpy® -->
▼<CATALOG>
 ▼<PLANT>
    <COMMON>Bloodroot</COMMON>
    <BOTANICAL>Sanguinaria canadensis</BOTANICAL>
    <ZONE>4</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$2.44</PRICE>
    <AVAILABILITY>031599</AVAILABILITY>
  </PLANT>
 ▼<PLANT>
    <COMMON>Columbine</COMMON>
    <BOTANICAL>Aquilegia canadensis</BOTANICAL>
    <ZONE>3</ZONE>
    <LIGHT>Mostly Shady</LIGHT>
    <PRICE>$9.37</PRICE>
    <AVAILABILITY>030699</AVAILABILITY>
  </PLANT>
```



WaterOneFlow and WaterML 1.1

- WaterOneFlow = web service interface for publishing hydrologic point time series

- WaterML = XML schema used by WaterOneFlow to encode data stored in an ODM database for transport over the Internet

WaterML 2.0

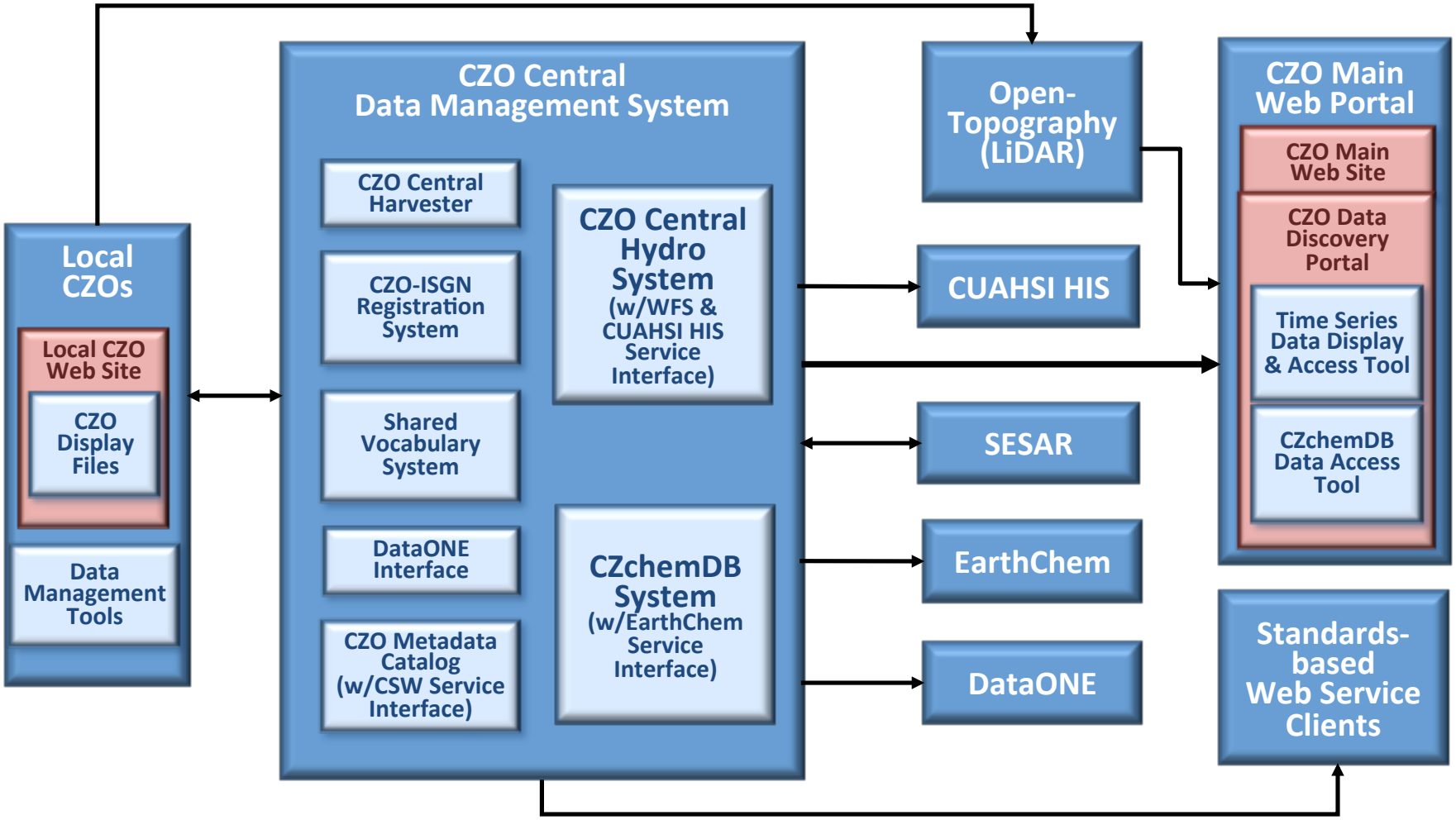
- An international metadata standard governed by the Open Geospatial Consortium (OGC)
  - To capture the semantics of hydrological observational data for data exchange
- An XML schema as a profile of OGC Observations and Measurements (O&M)
  - Part 1 – specifically focused on hydrologic time series
  - Part 2 – ratings, gaugings, and sections
  - Part 3 – water quality samples?????

Integrated CZO Data: 2012-2014

$1.5M starting Dec. 1, 2012, PI: Aufdenkampe

- **Community**: extensive and iterative interaction and feedback from CZO PIs, scientists and data managers
- **Web**: uniform web portal appearance for the CZO sites and the national CZO program
- **Metadata**: development of a consistent metadata strategy for CZO data, ensuring that the data descriptions follow consistent semantics
- **Workflows**: enhancing publication and data discovery workflows
  - collection of data submission forms and tools
  - integrating with the EarthChem system
- **Discovery**: creating a uniform data discovery portal
- **Visualization**: developing a consistent online data visualization interface for CZO time series data

CZOData II Architecture



Local CZOs | CZO Central Coordination Functions | CZO Central Data Repositories | Non-CZO Integrated Data & Discovery Sites | Clients

CZOData Challenges

1. Community Engagement

CZOData solution:

- Information Management Committee (IMC)
  - 1-2 investigators/CZO + site data managers
  - Weekly to monthly web conference meetings
  - Annual face-to-face meeting
  - Feedback to CZOData team
    - Data use scenarios
    - Meta-data requirements, shared vocabulary, etc.
- Workshops & Webinars to greater CZO community

CZO Data Challenges

1. Community Engagement
2. Web: Get small “long tail” data out of the dark!

CZOData solution:

- Leverage website to provide uniform data presentation & publication to CZO Central
  - Document repository (library)
  - Central database

Dataset

Christina River Basin - Stream Suspended Sediment (1993-2012)

TSS concentrations and elemental/isotopic composition (starting 2005) of baseflow and stormflow.

Variables: Solids_ total suspended, carbon to nitrogen molar ratio, carbon_ particulate organic, nitrogen_ particulate organic, nitrogen-15 stable isotope ratio delta

Date Range: (1993-2012)

Dataset Creators/Authors: Aufdenkampe, A.K.; Newbold, J.D.; Anderson, B. A.; Richardson, D.; Damiano, S.G.

Contact: Sara Geleskie Damiano, Stroud Water Research Center, 970 Spencer Road, Avondale, PA 19311, sgeleskie@stroudcenter.org

Field Area: Boulton Run | Christina River Basin | Forest Endmember: Spring Brook | White Clay Creek @ SWRC | Construction Endmember: White Clay Creek below landfill | Lower White Clay Creek | Agricultural Endmember: South Branch Doe Run

Water Chemistry
Geomorphology
Biogeochemistry
Hydrology
Geochemistry / Mineralogy

Christina

Description | Keywords | Citation | Publications | Acknowledgements

Description

Total suspended solids (TSS) and Volatile Suspended Solids (VSS) from White Clay Creek near the Stroud Water Research Center, Avondale, PA, USA. The purpose is to quantify export of inorganic and organic particulate matter from the 725-hectare watershed. Samples consist of those taken at monthly intervals, normally the first Wednesday of each month regardless of weather or flow conditions and those taken after precipitation events. The monthly samples are manual grab samples collected in 5-L polyethylene "space saver" bottles from a few centimeters below the surface and without disturbance to the stream bed. The event samples were collected in response to precipitation of 20 mm or more using an ISCO automated sampler which collected 1-L samples s in polyethylene bottles at hourly intervals through an intake approximately 20 cm above the bed. Each of approximately four events per year are represented by approximately 10 samples selected from the hourly series to characterize the rise, peak, and falling limb of the hydrograph. Additional events are represented by the three samples nearest peak flow.

Data

Christina River Basin - Stream Suspended Sediment 1993

(.csv) Data Level 1, Metadata

CZO Field Areas

Boulton Run

Christina River Basin

Forest Endmember: Spring Brook

CZO Data Challenges

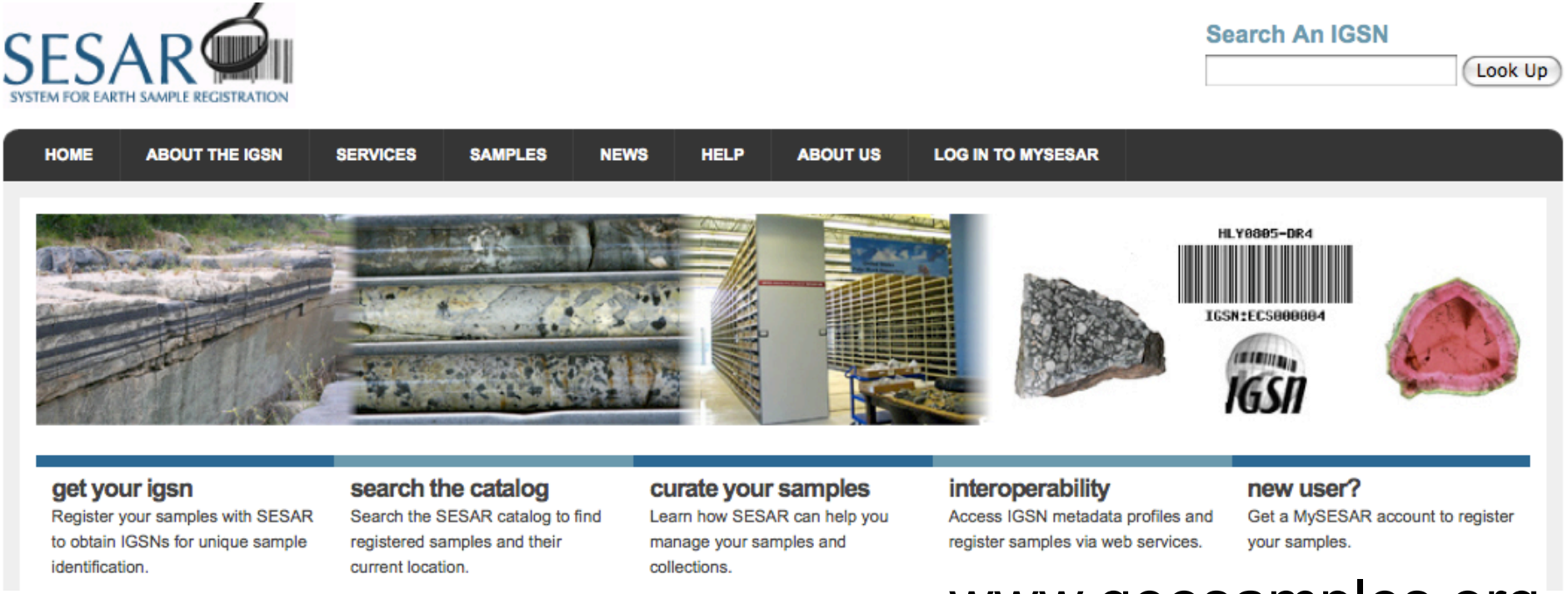
1. Community Engagement
2. Web: Get small “long tail” data out of the dark!
3. Metadata: Integrate data from all CZ disciplines

CZOData solutions:

- Strong integration with IEDA and SESAR
- Observations Data Model 2.0 (ODM2)
  - Separate NSF Geoinformatics project, PI: J. Horsburgh

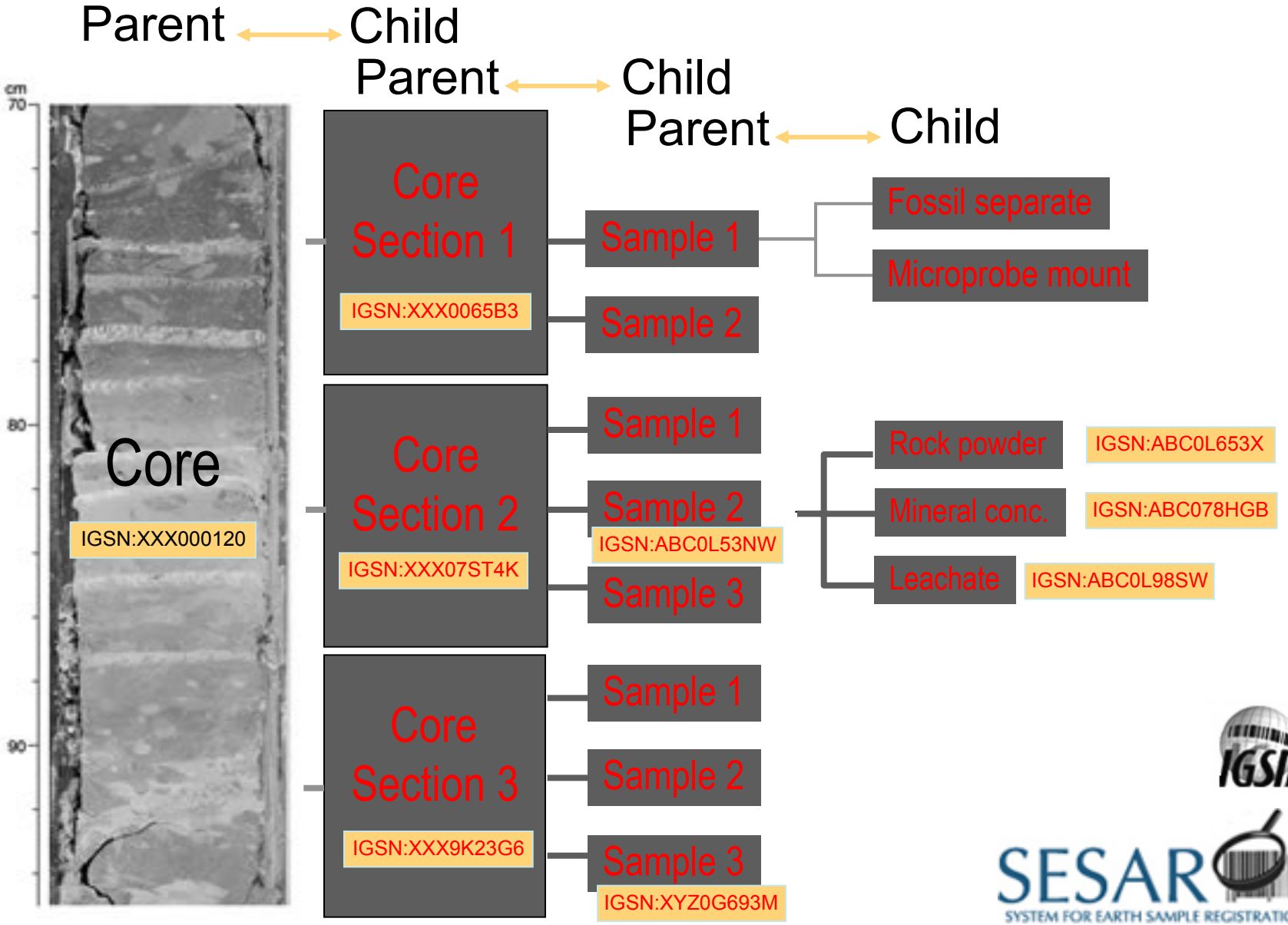
System for Earth Sample Registration

- Facilitate discovery of samples
- Ensure unique identification with International GeoSample Number (IGSN)
- Preserve sample metadata



www.geosamples.org

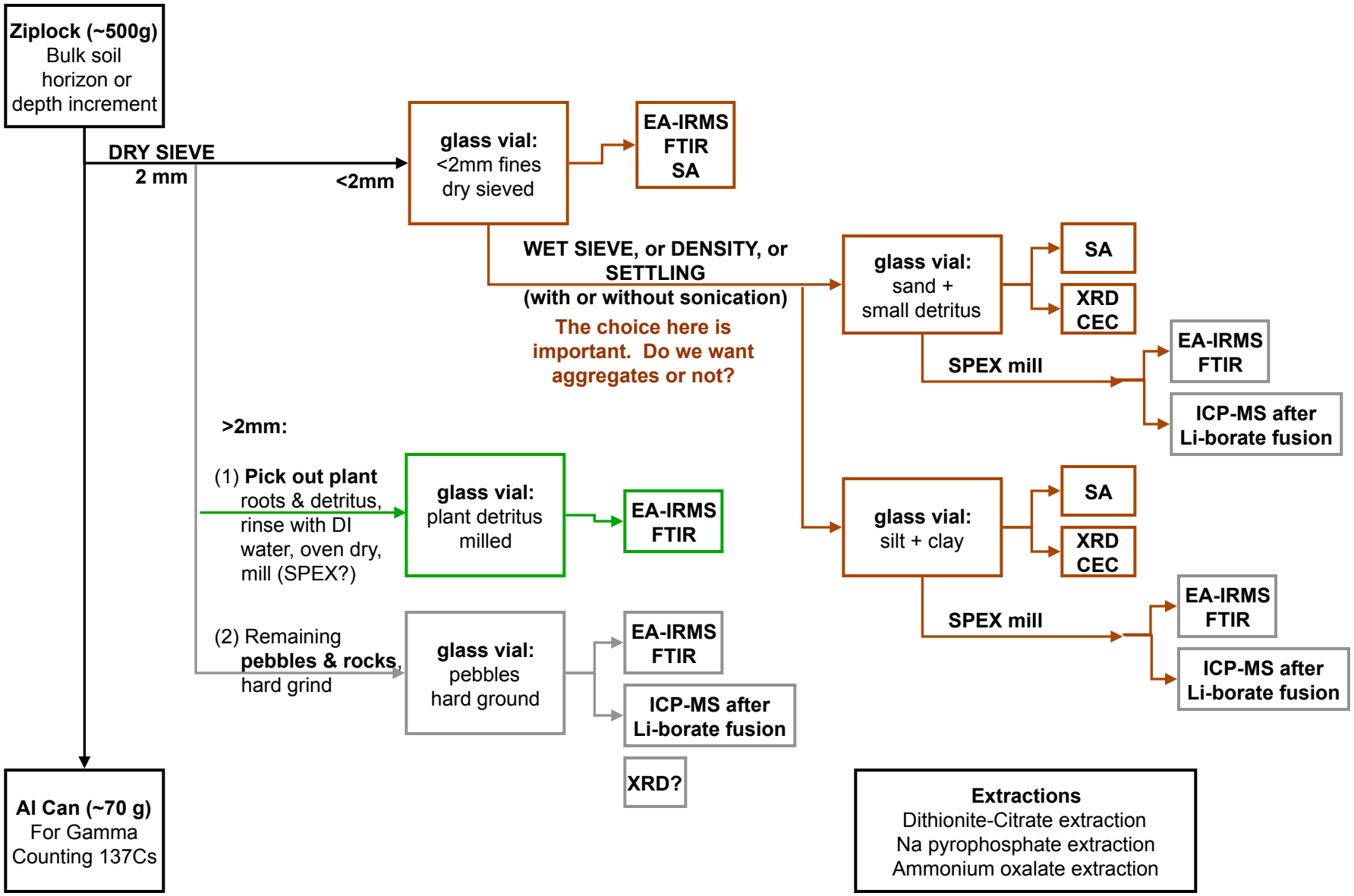
Sample Management for Geochemistry



IGSN

SESAR
SYSTEM FOR EARTH SAMPLE REGISTRATION

Sample Fractions for Soil Geochemistry



Observation Data Model 2.0

- NSF funded project: PI. Jeff Horsburgh
  - "Developing a Community Information Model and Supporting Software to Extend Interoperability of Sensor and Sample Based Earth Observations"

  - To achieve interoperability between IEDA, EarthCHEM, CUAHSI HIS, and other data systems
  - Better support for samples and unique identifiers (IGSN/SESAR)
  - Extensibility to table attributes
  - Better annotation and provenance
  - Enable integrated web service based publication of a broader class of CZO data

Data Management Concepts

- Flat files → relational databases
  - Linking tables
  - Reducing redundancy
  - Why a “data model” matters
- Meta-data “standards” via
  - Controlled Vocabulary (or Shared Vocab)
  - Choice of fields in data model
- Web Services
  - Clients-Server exchanges via XML
- Data discovery

Flat files → relational databases

Field,
Attribute

Record →

DateTime	Station	Latitude	Longitude	Sensor	StreamStage	Units
1/23/12 12:00	WCC-PumpHouse	39.86066	-75.783855	Solinist	23	cm
1/23/12 12:15	WCC-PumpHouse	39.86066	-75.783855	Solinist	24	cm
1/23/12 12:30	WCC-PumpHouse	39.86066	-75.783855	Solinist	25	cm
1/23/12 12:45	WCC-PumpHouse	39.86066	-75.783855	Solinist	26	cm
1/23/12 13:00	WCC-PumpHouse	39.86066	-75.783855	Solinist	27	cm
1/23/12 13:15	WCC-PumpHouse	39.86066	-75.783855	Solinist	28	cm
1/23/12 13:30	WCC-PumpHouse	39.86066	-75.783855	Solinist	29	cm
1/23/12 13:45	WCC-PumpHouse	39.86066	-75.783855	Solinist	30	cm
1/23/12 14:00	WCC-PumpHouse	39.86066	-75.783855	Solinist	31	cm

Flat files → relational databases

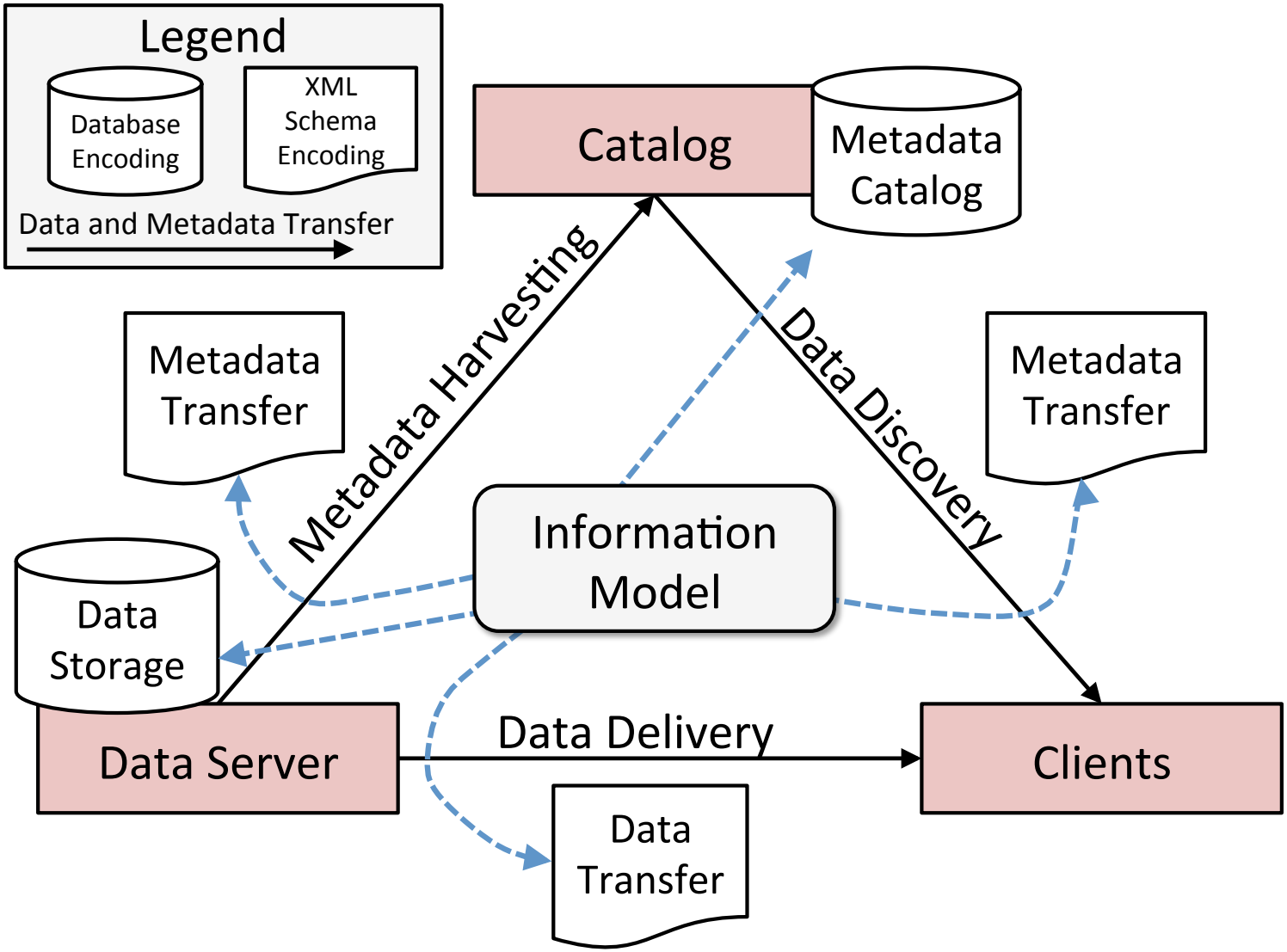
Foreign Key

DateTime	StationID (FK)	Sensor	StreamStage	Units (CV)
1/23/12 12:00	1	Solinist	23	cm
1/23/12 12:15	1	Solinist	24	cm
1/23/12 12:30	1	Solinist	25	cm
1/23/12 12:45	1	Solinist	26	cm
1/23/12 13:00	1	Solinist	27	cm
1/23/12 13:15	1	Solinist	28	cm
1/23/12 13:30	1	Solinist	29	cm
1/23/12 13:45	1	Solinist	30	cm
1/23/12 14:00	1	Solinist	31	cm

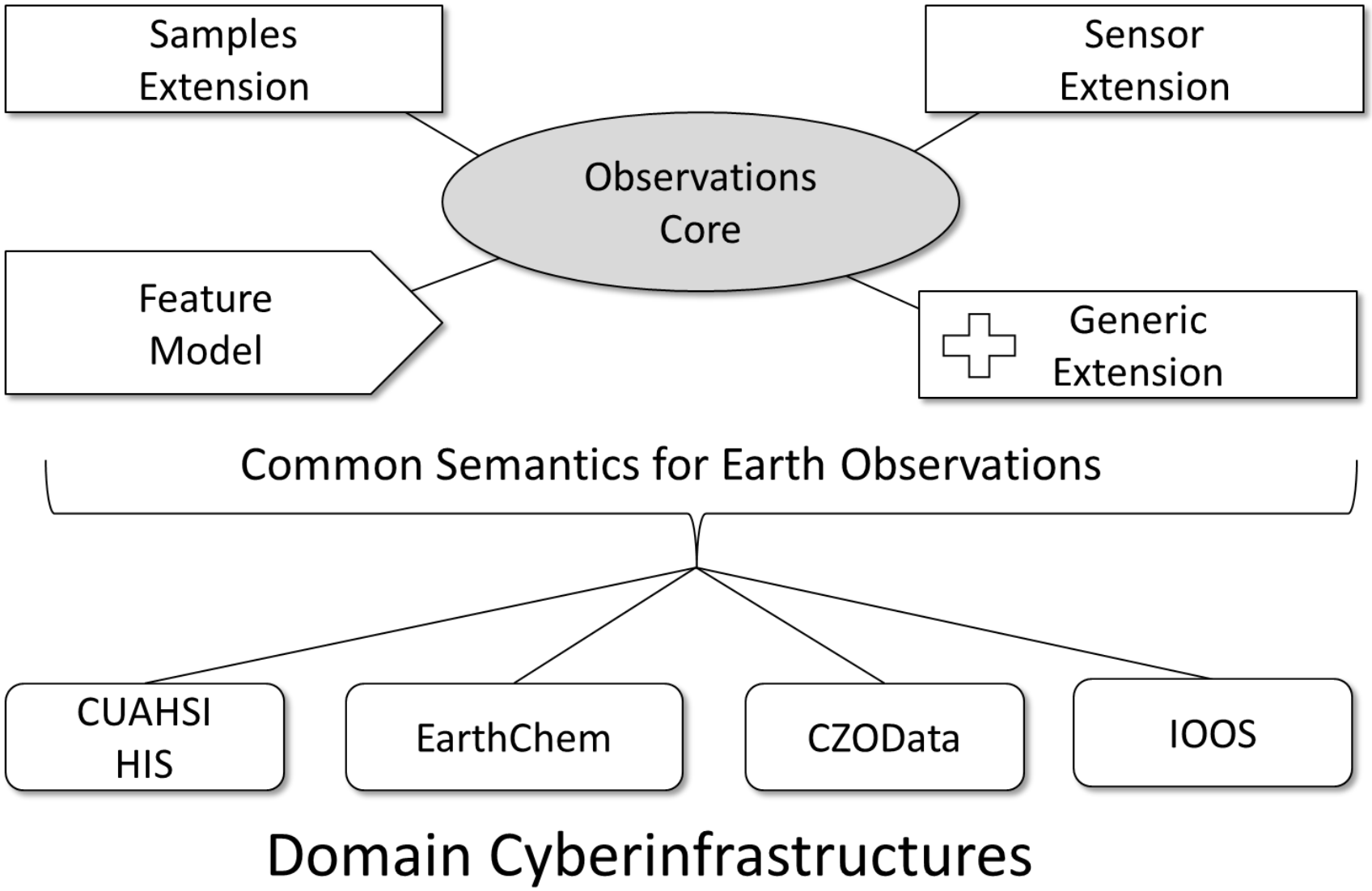
Primary Key

StationID (PK)	StationName	Latitude	Longitude
1	WCC-PumpHouse	39.86066	-75.783855
2	WCCLab	39.8594	-75.78381
3			
4			

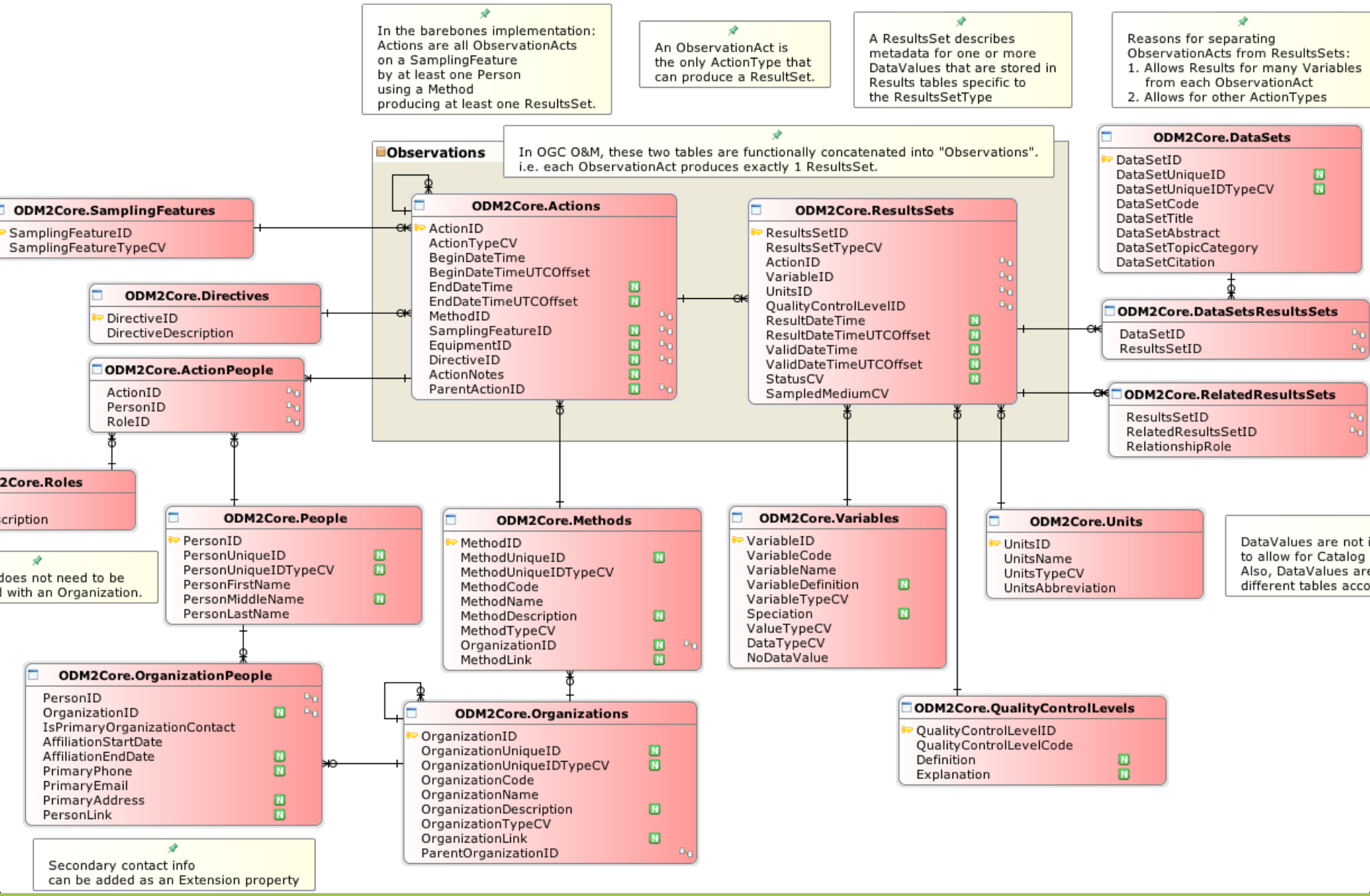
ODM2: Common to All Components



ODM2: Common to Most Data Types



ODM2: Core Schema



CZO Data Challenges

1. Community Engagement
2. Web: Get small “long tail” data out of the dark!
3. Metadata: Integrate data from all CZ disciplines
4. Workflows: Enhance data management practice

CZOData solutions:

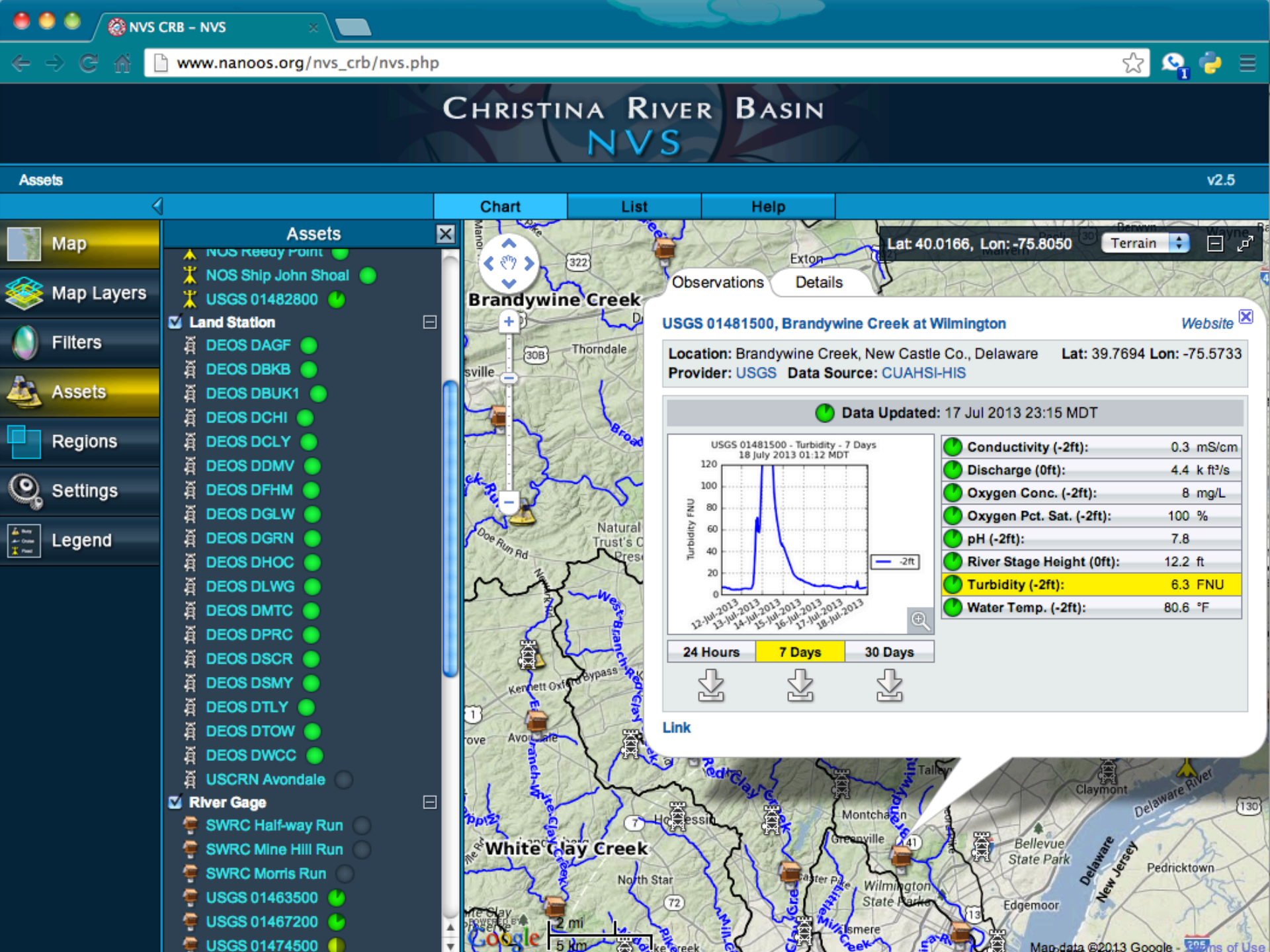
- CZO Display File version 2
- ODM2 extensions for data management developing with BiG CZ SSI Toolbox
  - Sensor/Instrument tracking
  - Sample tracking

CZO Data Challenges

1. Community Engagement
2. Web: Get small “long tail” data out of the dark!
3. Metadata: Integrate data from all CZ disciplines
4. Workflows: Enhance data management practice
5. Web-based data discovery & visualization client

CZOData solutions:

- Adapting NANOOS Visualization System (NVS) for CZOs
- See http://nvs.nanoos.org/Explorer


NVS CRB – NVS
www.nanoos.org/nvs_crb/nvs.php
CHRISTINA RIVER BASIN
NVS
Assets
v2.5
Chart
List
Help
Map
Map Layers
Filters
Assets
Regions
Settings
Legend
Assets
NOS Reedy Point
NOS Ship John Shoal
USGS 01482800
Land Station
DEOS DAGF
DEOS DBKB
DEOS DBUK1
DEOS DCHI
DEOS DCLY
DEOS DDMV
DEOS DFHM
DEOS DGLW
DEOS DGRN
DEOS DHOC
DEOS DLWG
DEOS DMTC
DEOS DPRC
DEOS DSCR
DEOS DSMY
DEOS DTLY
DEOS DTOW
DEOS DWCC
USCRN Avondale
River Gage
SWRC Half-way Run
SWRC Mine Hill Run
SWRC Morris Run
USGS 01463500
USGS 01467200
USGS 01474500
Lat: 40.0166, Lon: -75.8050
Terrain
Brandywine Creek
White Clay Creek
Observations
Details
USGS 01481500, Brandywine Creek at Wilmington
Website
Location: Brandywine Creek, New Castle Co., Delaware
Lat: 39.7694 Lon: -75.5733
Provider: USGS Data Source: CUAHSI-HIS
Data Updated: 17 Jul 2013 23:15 MDT
USGS 01481500 - Turbidity - 7 Days
18 July 2013 01:12 MDT
Turbidity FNU
-2ft
Conductivity (-2ft): 0.3 mS/cm
Discharge (0ft): 4.4 k ft³/s
Oxygen Conc. (-2ft): 8 mg/L
Oxygen Pct. Sat. (-2ft): 100 %
pH (-2ft): 7.8
River Stage Height (0ft): 12.2 ft
Turbidity (-2ft): 6.3 FNU
Water Temp. (-2ft): 80.6 °F
24 Hours
7 Days
30 Days
Link
Map data ©2013 Google

NSF Scientific Software Integration

BiG CZ SSI project (Dec. 2013 to Nov. 2015):

The community-driven BiG CZ software system for integration and analysis of bio- and geoscience data in the critical zone

- Community Engagement in Software Design
  through co-design, training & testing workshops.
- BiG CZ Portal web application
  for high-performance map-based discovery, visualization, access & publication of data on critical zone structure & function
- BiG CZ Toolbox
  to enable cyber-savvy CZ scientists & data managers to manage and publish the data they produce through a single scientist-focused toolkit
- BiG CZ Central software stack
  to bridge data systems developed for multiple critical zone domains

BiG CZ Portal

- A web-based integration and visualization environment for joint analysis of cross-scale bio and geoscience processes in the critical zone, spanning experimental and observational designs
  - intuitive, high-performance map-based discovery, visualization, access and publication of heterogeneous datasets, including:
    - Points with sensor & sample observations;
    - 2D satellite and GIS imagery from many different agencies and sources;
    - 2D and 3D interactive visualization of datasets.
  - Data publication/registration forms, to easily “shed light on dark data”

BiG CZ Toolbox

- A single, easy-to-install and easy-to-teach cross-platform package with components to facilitate local data management workflows and analysis, and publication to data repositories.
  - A ready-to-use relational database for many data types
  - Web service interfaces for both accessing and sharing near real-time or historical data
  - Streaming sensor middleware and other software for managing, visualizing and QAing data from sensor networks
  - A publication coordination tool to maintain common identifiers, spatial-temporal registration and controlled vocabularies when publishing to repositories.

BiG CZ Central

- Cloud software stack to bridge data systems developed for multiple critical zone domains into a single metadata catalog
  - Central catalog for single search of multiple repositories, with focus on bridging bio & geo.
  - Unified web services API to access multiple repositories using a variety of standards/protocols

Thank You